Optimization and Quantization of Multibeam Beamforming Vector for Joint Communication and Radio Sensing

Yuyue Luo, *Student Member, IEEE*, J. Andrew Zhang, *Senior Member, IEEE*, Xiaojing Huang, *Senior Member, IEEE*, Wei Ni, *Senior Member, IEEE*, Jin Pan

***Abstract*—Joint communication and radio sensing (JCAS) in millimeter wave (mmWave) systems requires the use of steerable beam. For analog antenna arrays, a single beam is typically used, which limits the sensing area to within the direction of the communication. Multibeam technology can overcome this limitation by separately generating package-level direction-varying sensing subbeams and fixed communication subbeams and then combing them coherently. In this paper, we investigate the optimal combination of the two subbeams and the quantization of the beamforming (BF) vector that generates the combined beam. When either full channel matrix or only the angle of departure (AoD) of the dominating line-of-sight (LOS) path is known at the transmitter, we derive closed-form expressions for the optimal combining coefficients that maximize the received communication signal power. For quantization of the BF vector, we focus on the two-phase-shifter array where two phase shifters are used to represent each BF weight. We propose novel joint quantization methods by combining the codebooks of the two phase shifters. The mean squared quantization error is derived for various quantization methods. Extensive simulation results validate the accuracy of the analytical results and the effectiveness of the proposed multibeam optimization and joint quantization methods.**

***Index Terms*—Multibeam, beamforming, joint communication and radio sensing, quantization, phase shift.**

I. Introduction

Joint communication and radio sensing (JCAS, also known as Radar-Communications) techniques have received strong interests from both academia and industry. JCAS integrates radio communication and sensing into one system, sharing the same transmitted signals [1]–[6]. Rooting in traditional radar technology, radio sensing here is referred to as information retrieval for environment surrounding radio transceivers, based on estimating the position, speed and feature signal of objects, activities, and events. JCAS systems have appealing features such as low-cost, resource-saving, reduced size and weight, and mutual sharing of information for improved communication and sensing performance [7].

Millimeter wave (mmWave) is particularly promising for JCAS due to its very wide bandwidth and hence fine resolution capability. However, there are also particular challenges associated with its usage of beamforming (BF). For mmWave, BF with steerable beam is essential for overcoming large propagation attenuation, supporting mobility, and estimating sensing parameters such as angle-of-arrival (AoA) and angle-of-departure (AoD) of signals. To reduce hardware cost, BF in mmWave systems is typically realized by using analog antenna array or hybrid array [8]. A mmWave JCAS system using low-cost and low-resolution analog-to-digital converters (ADC) is also investigated in [9], where the Cramér Rao bound (CRB) for sensing and the achievable data rate for communication are characterized with respect to ADC's quantization step.

Although extensive studies have been conducted on BF for sensing and communication separately [10]–[12], the results cannot be directly applied to BF and beam-steering in mmWave JCAS with analog or hybrid array. The primary challenge is that communication and sensing have different requirements for BF. Radio sensing often requires time-varying directional scanning beams, while a stable and accurately-pointing beam is usually demanded by communication.

BF design for JCAS systems has been investigated in [3]–[5], [13]–[15]. For digital multiple-input-multiple-output (MIMO) systems, flexible system design and optimization can be realized due to the multiple degrees of freedom in the spatial domain. As a result, beams with multiple mainlobes can be generated to support communication and sensing in different directions. In [4], sparse antenna array and BF optimization are studied for JCAS MIMO systems. In [5], waveform optimization is studied for minimizing the difference between the generated signal and the desired sensing waveform under the constraints of signal-to-interference-and-noise ratio (SINR) for multiuser MIMO communications. In [15], globally optimal waveforms are derived for multiple desired radar beam patterns, using the metric of minimizing multiuser interference for communications. Unfortunately, these problem formulations are based on digital MIMO systems and are not suitable for a cost-effective, compact, and computationally

This work was supported by the Foundation for Innovative Research Group of the National Natural Science Foundation of China under Grant Nos. 61721001.

Yuyue Luo is with School of Electronic Science and Engineering, University of Electronic Science and Technology of China, China. She is also with School of Electrical and Data Engineering, University of Technology Sydney, Australia. Email: Yuyue.Luo@student.uts.edu.au.

J. Andrew Zhang and Xiaojing Huang are with School of Electrical and Data Engineering, University of Technology Sydney, Australia. Email:{Andrew.Zhang;Xiaojing.Huang}@uts.edu.au.

Wei Ni is with Data61, CSIRO, Sydney, Australia, NSW 2122. E-mail: wei.ni@data61.csiro.au.

Jin Pan is with School of Electronic Science and Engineering, University of Electronic Science and Technology of China, China. Email: panjin@uestc.edu.cn.

Part of the work presented in this paper was accepted for publication in IEEE ICC2019. More than 50% of the work here is new and different to that paper.

efficient analog antenna array where there are much lower degrees of freedom in optimization due to a single RF chain of the array. For example, these MIMO designs can optimize a digital BF precoding matrix, but with an analog array, we can only optimize an analog BF weighting vector. For JCAS with analog BF, most studies such as [3], [13], [14] only consider a single beam for communication and sensing, hence sensing is restricted to the communication direction. In a relevant yet different context, BF design for other dual function systems such as joint wireless information and power transfer has also been investigated in [16], [17]. Such designs also consider joint optimization of two cost functions, but they use very different objective functions and are also based on digital MIMO systems.

In [6], [18], multibeam technology is introduced for mmWave JCAS, with the use of analog antenna arrays. In that work, multibeam is defined as a BF waveform with two or more mainlobes (also called subbeams) generated by a single analog array at a time. It provides a fixed communication subbeam along with direction-varying scanning subbeams across different packets. Several methods are proposed for generating the multibeam in [6]. Our multibeam scheme is shown to be superior in balancing complexity and performance by separately generating two basic beams for communication and sensing and then combining them according to the desired directions. In this paper, we improve the multibeam method in [6] by addressing two important problems: the optimal combination of separately generated sensing and communication beams, and the quantization of the combined BF vector for practical systems with only discrete phase shifting values.

The first problem arises from the fact that two separately generated beams can counteract each other if they are simply added up, particularly when their mainlobes point to adjacent directions. An adaptive coefficient is applied in [6] to combine the two subbeams coherently to increase the BF gain. The computation of this combining coefficient in [6] is simple but not optimized. We note that some existing studies on constructive interference for multiuser MIMO communications [19] and radar-communication coexistence systems [20] can potentially apply symbol-level precoding to align multiuser's signals so that the signals can be combined at each user's receiver constructively. Although their goal is similar to ours, our problem here is different as only a single symbol is transmitted from all antennas and there is no interference between the sensing and communication subbeams. This leads to different optimization strategies, as will be described in Section III.

The second problem is due to the fact that practical analog arrays only have phase shifters with discrete phase shifting values, while in [6] only continuous and non-quantized BF vectors are considered. Quantization of BF vectors can cause large mismatches on BF waveform and degradation on BF gain [21], [22], particularly when BF weights can only be quantized as discrete phase shifting values. Recently, a two-phase-shifter (2-PS) scheme is proposed for BF weight quantization in [23]. Using two phases can potentially represent any element in a normalized BF vector with a negligible quantization error, when the number of quantization bits is sufficiently large. Basic performance analysis for the 2-PS arrays is provided in [23], but detailed quantization methods are not presented.

In this paper, for the multibeam method in [6], we first study the optimization of coefficients for combining communication and sensing subbeams, and then investigate several methods for quantizing the combined BF vector and characterize the quantization error. Our main contributions in this paper are summarized below.

- We derive the optimal coefficients for combining communication and sensing subbeams. These coefficients can maximize the received signal power in two cases: (1) when the full channel matrix $\mathbf{H}$ is known, and (2) when the (estimated) AoD of the dominating path is known. The optimality is analytically proven and validated by simulation results.
- We introduce several quantization methods for quantizing the BF vector, particularly for the 2-PS arrays. We propose a novel joint quantization method that combines the codebooks of the two phase shifters. With the new codebooks, particular the one established by introducing a fixed phase shifting value to one of the phase shifters, we show that even scalar quantization can achieve performance close to the non-quantized case when there are more than 3 quantization bits in each phase shifter. We also develop improved golden section search-quantization (IGSS-Q) method that enables better scalar quantization by considering the property of vector quantization.
- We analytically evaluate and compare the mean squared quantization error (MSQE) for several 1-PS and 2-PS quantization methods. These analytical results are shown to match the simulated results well.

Extensive simulation results are provided and compared to other schemes to validate the optimality of the derived combining coefficients in terms of signal power, the effectiveness of the quantization methods, and the match between analytical and simulated quantization errors.

The rest of this paper is organized as follows. We introduce the system model, formulate the problem and address our principle of multibeam optimization in Section II. The optimal combining coefficients are investigated in Section III. We then study the quantization methods in Section IV, and analytically evaluate the quantization error for these methods in Section V. In Section VI, extensive simulation results are provided, and finally, concluding remarks are provided in Section VII.

Notations: $(\cdot)^H$, $(\cdot)^*$, $(\cdot)^T$, $(\cdot)^{-1}$, and $(\cdot)^\dagger$ denote the Hermitian transpose, conjugate, transpose, inverse, and pseudo-inverse, respectively. $|\cdot|$, $\|\cdot\|$, and $\|\cdot\|_2$ denote the element-wise absolute value, the norm, and the Euclidean norm, respectively. $E(\cdot)$ denotes the expected value. $\arg(\cdot)$ denotes the argument of a complex number.

II. System Model and Problem Formulation

In this section, we briefly review the system model, present the concept of multibeam technologies based on one analog array, and discuss the principle of our multibeam optimization.

A. System Model

Our work on multibeam generation inherits the JCAS system proposed in [6]. In the system, two nodes perform two-way point-to-point communications in the time division duplex (TDD) mode and simultaneously sensing the environment to determine locations and speed of nearby objects. Each node uses two spatially separated analog antenna arrays to enable radio sensing while transmitting the signal. In each array, we assume that both the amplitude and phase of the signal at each antenna can be adjusted, which can be realized by active phased arrays [24], [25] or the two-phase-shifter passive array that is the focus of this paper. Only a single RF and baseband module is used and connected to the two arrays. Below we only briefly describe the essential system setup to make this paper self-contained. Readers are referred to [6] for more details of the system and the multibeam JCAS technology.

We consider M-element antenna arrays where antenna elements are equally spaced at an interval of half wavelength. We assume planar wave-front and consider a narrow-band beamforming model. The array response vector is given by

$$\mathbf{a}(\theta) = [1, e^{j\pi \sin(\theta)}, \cdots, e^{j\pi(M-1)\sin(\theta)}]^T. \tag{1}$$

A quasi-static channel model can be used for both communication and sensing, although the values of their parameters are different. Consider L multipath signals with AoDs $\theta_{t,\ell}$ and AoAs $\theta_{r,\ell}$. For simplicity, we assume that transmitter and receiver arrays have the same number of antennas M, and the results in this paper can be straightforwardly extended to arrays with different numbers of antennas. The quasi-static physical channels between the transmitting and receiving antennas can then be represented as

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

where for the ℓ-th path, b_ℓ is its amplitude of complex value, τ_ℓ is the propagation delay, and $f_{D,\ell}$ is the associated Doppler frequency. Note that as discussed in [6], these channel parameters are generally different for communications and sensing for active sensing, and are the same for passive sensing. We do not differentiate between active and passive sensing in this paper, and the results here are generally applicable to both. We consider typical multipath mmWave channels where there exists a line-of-sight (LOS) path and $(L-1)$ non-line-of-sight (NLOS) paths. The LOS path is assumed to be dominating in terms of signal power.

Let the transmitted baseband signal be $s(t)$, and the transmitter and receiver BF vectors be $\mathbf{w}_t$ and $\mathbf{w}_r$, respectively. The received signal for either sensing or communication can be written as:

$$\begin{aligned} y(t) &= \mathbf{w}_r^T \mathbf{H} \mathbf{w}_t \, s(t-\tau_\ell) + \mathbf{w}_r^T \boldsymbol{z}(t) \qquad (3) \\ &= \sum_{\ell=1}^{L} b_\ell e^{j2\pi f_{D,\ell} t} (\mathbf{w}_r^T \mathbf{a}(\theta_{r,\ell}))(\mathbf{a}^T(\theta_{t,\ell})\mathbf{w}_t) s(t-\tau_\ell) + \mathbf{w}_r^T \boldsymbol{z}(t), \end{aligned}$$

where $\boldsymbol{z}(t)$ is the independently and identically distributed additive white Gaussian noise (AWGN) vector at the receiving antennas. Consequently, the received signal-to-noise ratio (SNR) can be written as

$$\gamma = \frac{||\mathbf{w}_r^T \mathbf{H} \mathbf{w}_t||^2}{||\mathbf{w}_r||^2} \cdot \frac{\sigma_s^2}{\sigma_n^2}, \tag{4}$$

where σ_s^2 is the mean power of $s(t)$ and σ_n^2 is the variance of AWGN.

B. Multibeam for JCAS

We want to generate a BF waveform with one subbeam (mainlobe) for communication and another one or more subbeams for sensing which may need to scan areas in different directions from communication. For this purpose we proposed two multibeam generation methods in [6].

Both methods in [6] use an iterative least square (ILS) method to generate the BF vectors according to the desired beam patterns, which are usually specified as the magnitude of the BF waveform. The first method generates two BF vectors for communication and sensing respectively based on their desired beam pattern. Then it combines the two BF vectors using a phase shifting term $e^{j\varphi}$ and power distribution factor ρ, as shown below

$$\mathbf{w}_t = \sqrt{\rho}\mathbf{w}_{t,c} + \sqrt{1-\rho} e^{j\varphi} \mathbf{w}_{t,s}, \tag{5}$$

where $\mathbf{w}_{t,c}$ and $\mathbf{w}_{t,s}$ are the respective BF vectors for communication and sensing, the power distribution factor ρ $(0 < \rho < 1)$ controls the power allocation between two BF vectors. The value of ρ can be flexibly set. For a given shape of the BF waveform, it is shown in [6] that BF pointing to a different direction can be easily generated by multiplying a phase shifting sequence to the basic BF vectors. The second method generates a single BF vector directly based on the desired joint BF waveform for communication and sensing.

The second method has the advantage in generating a BF waveform with the shape closer to the desired one. However, the first method is more appealing owing to the following advantages. 1) It provides great flexibility for varying BF directions and power distribution between communication and sensing; 2) It potentially enables the constructive combination of communication and sensing subbeams at the communication receiver to improve the received signal power, especially when the two subbeams are overlapped. One example of the multibeam is shown in Fig. 1.

C. Principle of Our Multibeam Optimization

In this paper, we further study the first multibeam generation method, by proposing in-principal optimal solutions to the phase shifting term and investigating its performance under practical situation with quantized magnitude and phase for the BF vectors.

The optimization of multibeam generation in (5) involves both ρ and φ. Here, we consider a sub-optimal two-stage process for determining the values of ρ and φ. In the first stage, ρ can be decided based on the required communication performance and the sensing ranges; and in the second stage, φ is uniquely optimized for any given ρ. Although suboptimal, this two-stage process is well suitable for practical mmWave systems, where the channel fading varies fast due to the small

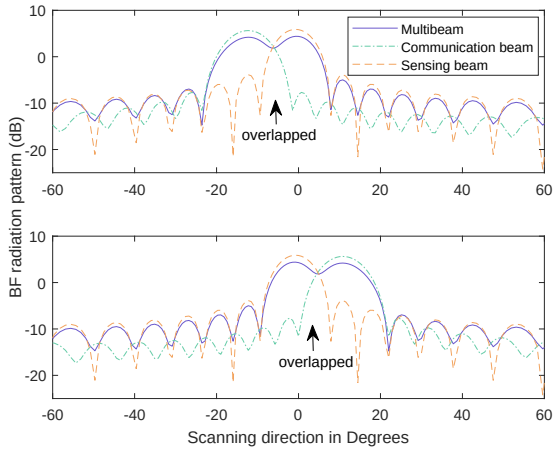


Fig. 1. Example of two separately generated subbeams and the combined multibeam using Method 1 in [6]. Communication subbeam points at 0 degree, and scanning subbeam points at -12.3 (top subfigure) and 10.8 (bottom subfigure) degrees.

wavelength of mmWave signals, while the path loss can remain stable over a relatively long period. Over this period, we only need to update φ in response to fast-changing channel fading.

The value of ρ depends on specific communication and sensing requirements. It can be adjusted to trade off between the performances of communication and sensing. When sensing is given priority, the required power for sensing can be first decided based on the desired sensing distance, and then the proper modulation and coding scheme can be decided for communication. When communication is given priority, the power can be allocated to meet the communication capacity, while being reserved to meet the requirement for a minimal sensing distance. In either case, the allocation is straightforward. Hence we ignore detailed design of ρ here and focus on optimizing the value of φ for a given ρ. As will be shown later, the optimized φ can significantly increase the received signal power for communications.

When optimizing φ, we focus on its impact on communication signals. This is because its impact on sensing waveform is generally much weaker. Firstly, the proposed methods generally cause insignificant variation of the BF waveform, as we will see later (in Figs. 6 and 7 in Section VI). Secondly, the combined beam pattern only deviates notably from the desired one when the sensing and communication subbeams are very close in directions. In our multibeam scheme, the reflected signals of the communication subbeam is also used for sensing. In this case, the combined beam still provides good coverage for the targeted sensing area as the total energy of the beam remains unchanged and concentrated in these directions.

Compared to existing globally optimal solutions such as those reported in [5], [15], low complexity and fast adaptation to time-varying channels are the key advantages of our multibeam scheme. For a given shape of the BF waveform, it was shown in [6] that BF pointing to a different direction can be easily generated by multiplying a phase shifting sequence to the basic BF vectors with the computational complexity of $O(M)$. Since the BF vector for generating the basic BF waveform can typically be pre-computed and stored in the system, the complexity for computing $\mathbf{w}_{t,s}$ and $\mathbf{w}_{t,c}$ is negligible. The complexity of finding the optimal φ is $O(M)$ (or $O(M^2)$), when the dominating AoD (or the full channel matrix) is known at the transmitter, as will be shown in Section IV. Therefore, the complexity of our multibeam schemes is much lower than the global optimization schemes for MIMO JCAS systems, e.g., in [5], [15].

To the best of our knowledge, no globally optimal solution for JCAS systems with an analog array has been reported. To provide a benchmark for comparison, we propose the following heuristic BF design

$$\begin{aligned} \mathbf{w}_{\text{BM}} &= \lambda \mathbf{w}^{\star}_{\text{WF}} + (1-\lambda)\mathbf{w}^{\star}_{\text{Gain}}, \\ \mathbf{w}_{\text{BM}} &= \mathbf{w}_{\text{BM}}/\|\mathbf{w}_{\text{BM}}\|, \end{aligned} \tag{6}$$

where $\mathbf{w}^{\star}_{\text{WF}}$ is the BF vector minimizing the mean squared error (MSE) between the generated BF waveform and the desired one, and $\mathbf{w}^{\star}_{\text{Gain}}$ is the BF vector maximizing the received signal power for communication, and $0 \le \lambda \le 1$ is the weighting factor. By adjusting λ, the combined BF vector $\mathbf{w}_{\text{BM}}$ can prioritize and balance between the conformity of the generated BF waveform for sensing and the received signal power for communication. In Section VI.A, we will numerically generate the benchmark BF vectors $\mathbf{w}_{\text{BM}}$ under different values of λ, and compare their performance with our proposed schemes.

III. Optimal Phase Alignment for JCAS

In this section, we first demonstrate the impact of the phase shifting term $e^{j\varphi}$ on the received signal power, and then propose approaches for determining optimal φ in two scenarios, where (1) the full channel matrix and (2) the AoD of the dominating path is known at the transmitter.

A. Impact of Combining Coefficient

The combing coefficient, i.e., the phase shifting term $e^{j\varphi}$ in (5), can have a significant impact on the received signal power for communication. An example can be seen from Fig. 4 that will be presented in Section VI.A. In [6], we developed a method for determining $e^{j\varphi}$, which is effective but not optimized. Let $\mathbf{q}$ denote the right eigenvector corresponding to the maximum eigenvalue of $\mathbf{H}$ when $\mathbf{H}$ is known. When only the dominating path direction is known, $\mathbf{q}$ denotes the conjugate of the array steering vector at that direction. The method simply aligns the phases of the two outputs $\mathbf{q}^H\mathbf{w}_{t,c}$ and $\mathbf{q}^H\mathbf{w}_{t,s}$ via letting

$$e^{j\varphi} = \frac{\mathbf{q}^H\mathbf{w}_{t,c}(\mathbf{q}^H\mathbf{w}_{t,s})^H}{|\mathbf{q}^H\mathbf{w}_{t,c}(\mathbf{q}^H\mathbf{w}_{t,s})^H|}. \tag{7}$$

Although this guarantees that the two subbeams can be constructively added up at the communication receiver, it is not optimal. This is because $\mathbf{w}$ needs to be normalized to $||\mathbf{w}||^2$ which may not be the smallest for the choice of $e^{j\varphi}$ in (7). Thus the overall optimality is not guaranteed by simply aligning the phase like in (7).

B. Optimal Solution when $\mathbf{H}$ is Known at Transmitter

When $\mathbf{H}$ is known at the transmitter, eigenbeam is the ideal beamforming, and the transmitter and receiver BF vectors, $\mathbf{w}_t$ and $\mathbf{w}_r$, shall be the left and right eigenvectors of $\mathbf{H}$. However, $\mathbf{w}_t$ needs to vary over packets and hence cannot always be the eigenvector. Hence we do not particularly consider the optimization of $\mathbf{w}_{t,c}$ and $\mathbf{w}_{t,s}$, but study how to optimize the phase parameter φ for any given $\mathbf{w}_{t,c}$ and $\mathbf{w}_{t,s}$. The optimal φ, φ_{opt}, is obtained when the receiver SNR is maximized, which can be formulated as

$$\varphi_{\text{opt}} = \arg\max_{\varphi}\{\gamma; ||\mathbf{w}_t||_2 = 1\} = \arg\max_{\varphi} \frac{||\mathbf{w}_r^T \mathbf{H} \mathbf{w}_t||^2}{||\mathbf{w}_r||^2 ||\mathbf{w}_t||^2}, \tag{8}$$

where the transmit BF vector $\mathbf{w}_t$ is normalized to ensure equal transmission power for different $\mathbf{w}_t$ values. For $\mathbf{w}_r$, we assume that maximal ratio combining (MRC) [26] is applied in the analog domain and $\mathbf{w}_r = (\mathbf{H}\mathbf{w}_t)^*$. We can then rewrite (8) as

$$\begin{aligned} &\varphi_{\text{opt}} = \arg\max_{\varphi} \frac{\mathbf{w}_t^H \mathbf{H}^H \mathbf{H} \mathbf{w}_t}{||\mathbf{w}_t||^2}, \\ &\text{with } \mathbf{w}_t = \sqrt{\rho}\mathbf{w}_{t,c} + \sqrt{1-\rho}e^{j\varphi}\mathbf{w}_{t,s}. \end{aligned} \tag{9}$$

Since an MRC receiver maximizes the received power, we can see the equivalence between maximizing the received SNR and power here.

Let $g_1(\varphi) = \mathbf{w}_t^H \mathbf{H}^H \mathbf{H} \mathbf{w}_t$ and $g_2(\varphi) = |\mathbf{w}_t|^2$. Equation (9) can be rewritten as

$$\begin{aligned} &\varphi_{\text{opt}} = \arg\max_{\varphi} \left(f(\varphi) \triangleq \frac{g_1(\varphi)}{g_2(\varphi)} \right), \\ &\text{with} \\ &g_1(\varphi) = \rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2 \\ &\qquad + Pe^{j\varphi}\mathbf{w}_{t,c}^H \mathbf{H}^H \mathbf{H} \mathbf{w}_{t,s} + Pe^{-j\varphi}\mathbf{w}_{t,s}^H \mathbf{H}^H \mathbf{H} \mathbf{w}_{t,c}, \\ &g_2(\varphi) = \rho\|\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{w}_{t,s}\|^2 + Pe^{j\varphi}\mathbf{w}_{t,c}^H \mathbf{w}_{t,s} \\ &\qquad + Pe^{-j\varphi}\mathbf{w}_{t,s}^H \mathbf{w}_{t,c}, \end{aligned} \tag{10}$$

where $P \triangleq \sqrt{\rho(1-\rho)}$.

By studying the monotonicity of $f(\varphi)$, we can find the optimal phase combiner as

$$\varphi_{\text{opt}} = \begin{cases} \pi + \mu_0 - \gamma + 2l\pi, & \text{when } X_1 \geq 0, \\ \mu_0 - \gamma + 2l\pi, & \text{when } X_1 < 0, \end{cases} \quad l = 0, \pm 1, \pm 2 \cdots \tag{11}$$

where

$$\begin{aligned} &\gamma \triangleq \arctan(X_2/X_1),\ \mu_0 \triangleq \arcsin\left(\frac{L}{\sqrt{X_1^2 + X_2^2}}\right), \\ &X_1 \triangleq 2P|a_1|\cos(\alpha_1) + \\ &\qquad 2P|a_2|[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]\cos(\alpha_2), \\ &X_2 \triangleq -2P|a_1|\sin(\alpha_1) + \\ &\qquad 2P|a_2|[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]\sin(\alpha_2), \\ &L \triangleq -4P^2|a_1||a_2|\sin(\alpha_1 - \alpha_2), \\ &a_1 = |\mathbf{w}_{t,c}^H \mathbf{H}^H \mathbf{H} \mathbf{w}_{t,s}|,\ a_2 = |\mathbf{w}_{t,c}^H \mathbf{w}_{t,s}|, \\ &\alpha_1 = \arg(\mathbf{w}_{t,c}^H \mathbf{H}^H \mathbf{H} \mathbf{w}_{t,s}),\ \alpha_2 = \arg(\mathbf{w}_{t,c}^H \mathbf{w}_{t,s}). \end{aligned}$$

The detailed derivation is provided in Appendix A, and the existence of μ_0 via $|L| \leq \sqrt{X_1^2 + X_2^2}$ is proven in Appendix B. The complexity of calculating φ_{opt} here is $O(M^2)$.

C. Optimal Solution when only Dominating AoD is Known at Transmitter

It is generally challenging to get the full knowledge on the channel matrix $\mathbf{H}$, while it is more practical to estimate the dominating AoD. Here, we derive the optimal phase $\tilde{\varphi}_{opt}$ that maximizes the power at the dominating AoD θ_t. The problem can be formulated as

$$\begin{aligned} &\tilde{\varphi}_{\text{opt}} = \arg\max_{\varphi} \frac{\|\mathbf{a}^T(\theta_t)\tilde{\mathbf{w}}_t\|^2}{||\tilde{\mathbf{w}}_t||^2} \\ &\text{with } \tilde{\mathbf{w}}_t = \sqrt{\rho}\mathbf{w}_{t,c} + \sqrt{1-\rho}e^{j\varphi}\mathbf{w}_{t,s}, \end{aligned} \tag{12}$$

where $\mathbf{a}(\theta_t)$ is the steering vector at the dominating AoD θ_t. Let $\tilde{g}_1(\varphi) = \|\mathbf{a}^T(\theta_t)\tilde{\mathbf{w}}_t\|^2$ and $\tilde{g}_2(\varphi) = ||\tilde{\mathbf{w}}_t||^2$. Then (12) can be rewritten as

$$\begin{aligned} &\tilde{\varphi}_{\text{opt}} = \arg\max_{\varphi} \frac{\tilde{g}_1(\varphi)}{\tilde{g}_2(\varphi)}, \\ &\text{with} \\ &\tilde{g}_1(\varphi) = \rho\|\mathbf{w}_{t,c}^H \mathbf{a}^*\|^2 + (1-\rho)\|\mathbf{w}_{t,s}^H \mathbf{a}^*\|^2 \\ &\qquad + Pe^{j\varphi}\mathbf{w}_{t,c}^H \mathbf{a}^* \mathbf{a}^T \mathbf{w}_{t,s} + Pe^{-j\varphi}\mathbf{w}_{t,s}^H \mathbf{a}^* \mathbf{a}^T \mathbf{w}_{t,c}, \\ &\tilde{g}_2(\varphi) = \rho\|\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{w}_{t,s}\|^2 \\ &\qquad + Pe^{j\varphi}\mathbf{w}_{t,c}^H \mathbf{w}_{t,s} + Pe^{-j\varphi}\mathbf{w}_{t,s}^H \mathbf{w}_{t,c}. \end{aligned}$$

Similar to the process in Section III-B, we can obtain $\tilde{\varphi}_{\text{opt}}$ as

$$\tilde{\varphi}_{\text{opt}} = \begin{cases} \pi + \tilde{\mu}_0 - \tilde{\gamma} + 2l\pi, \text{ when } \tilde{X}_1 > 0, \\ \tilde{\mu}_0 - \tilde{\gamma} + 2l\pi, \text{ when } \tilde{X}_1 < 0, \end{cases} \quad l = 0, \pm 1, \pm 2 \cdots \tag{13}$$

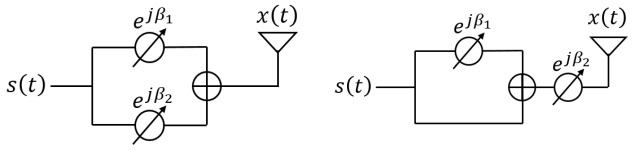
where

$$\begin{aligned} &\tilde{\gamma} \triangleq \arctan(\tilde{X}_2/\tilde{X}_1),\ \tilde{\mu}_0 \triangleq \arcsin\left(\tilde{L}/\sqrt{\tilde{X}_1^2 + \tilde{X}_2^2}\right), \\ &\tilde{X}_1 \triangleq -2P\tilde{a}_2\tilde{a}_3\cos(\tilde{\alpha}_2 + \tilde{\alpha}_3) \\ &\qquad + 2P\tilde{a}_1(\rho\tilde{a}_2^2 + (1-\rho)\tilde{a}_3^2)\cos(\tilde{\alpha}_1) \\ &\tilde{X}_2 \triangleq -2P\tilde{a}_2\tilde{a}_3\sin(\tilde{\alpha}_2 + \tilde{\alpha}_3) \\ &\qquad + 2P\tilde{a}_1(\rho\tilde{a}_2^2 + (1-\rho)\tilde{a}_3^2)\sin(\tilde{\alpha}_1) \\ &\tilde{L} \triangleq -4P^2\tilde{a}_1\tilde{a}_2\tilde{a}_3\sin(\tilde{\alpha}_2 + \tilde{\alpha}_3 - \tilde{\alpha}_1), \\ &\tilde{a}_1 = |\mathbf{w}_{t,c}^H \mathbf{w}_{t,s}|,\ \tilde{a}_2 = |\mathbf{w}_{t,c}^H \mathbf{a}^*|,\ \tilde{a}_3 = |\mathbf{a}^T \mathbf{w}_{t,s}|, \\ &\tilde{\alpha}_1 = \arg(\mathbf{w}_{t,c}^H \mathbf{w}_{t,s}),\ \tilde{\alpha}_2 = \arg(\mathbf{w}_{t,c}^H \mathbf{a}^*),\ \tilde{\alpha}_3 = \arg(\mathbf{a}^T \mathbf{w}_{t,s}). \end{aligned} \tag{14}$$

The detailed derivation is provided in Appendix C. The complexity of calculating $\tilde{\varphi}_{\text{opt}}$ is $O(M)$.

IV. Quantization of BF Vector with Phase Shifters

In practical analog arrays, beamforming weights can typically be represented only as quantized and discrete values instead of continuous ones. In this section, we study the impact of BF vector quantization on multibeam generation, by using phase shifters in the array only. After obtaining $\mathbf{w}_t$ via (5) and



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

Fig. 2. Optional parallel and serial structures with two phase shifters.

(11) (or (13)), the quantization is applied to the final BF vector $\mathbf{w}_t = \{w_i\} = \{|w_i|e^{j\psi_i}\}, i = 1, \cdots, M$, where $\psi_i = \angle(w_i)$. Let b be the number of quantization bits in each phase shifter. We assume that the discrete phase values are equally spaced over $[0, 2\pi]$ with a quantization step of $\Delta = 2\pi 2^{-b}$.

We study two cases when a single and double phase shifters, abbreviated as 1-PS and 2-PS, are used respectively. The two optional structures for 2-PS are shown in Fig. 2. As shown in [23], the 2-PS array can provide significantly reduced quantization errors compared to 1-PS, at increased hardware cost.

In this paper, we will mainly study element-wise scalar quantization. Although vector quantization can achieve better performance than element-wise quantization, its complexity is higher. As we will see in Section VI, element-wise quantization, particularly for the 2-PS array, can achieve performance approaching to a non-quantized one when the number of quantization bit b is moderately large, e.g., $b \geqslant 3$. When b is small, formulating BF vector quantization as non-coherent detection problems can be an effective approach for reducing quantization distortion at an affordable complexity, e.g., trellis based searching algorithms [27], [28] or maximum likelihood (ML) detection algorithms [29], [30].

For the 1-PS array, the element-wise quantization can be represented as

$$\hat{\beta}^{(i)} = \arg\min_{\hat{\beta}\in\mathcal{B}} |\mathrm{mod}_{2\pi}(\psi_i - \hat{\beta})| \tag{15}$$

where $\hat{\beta} \in \mathcal{B} = \{0, \Delta_\beta, 2\Delta_\beta, \ldots, (2^b - 1)\Delta\}$ with quantization step Δ_β, and $\mathrm{mod}_{2\pi}(x)$ stands for x modulo 2π.

Since the 1-PS array only represents a value with unit magnitude, the resulting amplitude mismatches can cause notable sidelobe even using phase shifters with infinite number of quantization bits [21], [22]. This can cause a waste of energy and difficulty for the angle of arrival estimation in multibeam sensing. This problem may be solved by adding power amplifiers/attenuators for each phase shifter. It can also be solved by using the 2-PS array, which is the main approach being investigated in this paper.

Next, we study several methods for determining the phase shifting values in the two 2-PS structures, including a novel joint quantization method that is particularly promising, as will be shown analytically and numerically later.

A. Separate Quantization of Individual Phase Shifts

For the parallel and serial structures in Fig. 2, the phase shifting values satisfy

$$w_i = |w_i|e^{j\psi_i} = e^{j\beta_1^{(i)}} + e^{j\beta_2^{(i)}}, \tag{16a}$$

and

$$w_i = |w_i|e^{j\psi_i} = e^{j\beta_1^{(i)}}(1 + e^{j\beta_2^{(i)}}), \tag{16b}$$

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

$$\beta_1^{(i)} = \psi_i + \arccos(|w_i|/2),\ \beta_2^{(i)} = \psi_i - \arccos(|w_i|/2), \tag{17a}$$

$$\beta_1^{(i)} = \psi_i - \arccos\left(\frac{|w_i|}{2}\right),\ \beta_2^{(i)} = 2\arccos\left(\frac{|w_i|}{2}\right), \tag{17b}$$

respectively.

A straightforward and simple way to decide the quantized phase shifts is then through quantizing each of them separately. This is given by

$$\begin{aligned}\hat{\beta}_1^{(i)} &= \arg\min_{\hat{\beta}_1\in\mathcal{B}_1} |\mathrm{mod}_{2\pi}(\beta_1^{(i)} - \hat{\beta}_1)|,\\ \hat{\beta}_2^{(i)} &= \arg\min_{\hat{\beta}_2\in\mathcal{B}_2} |\mathrm{mod}_{2\pi}(\beta_2^{(i)} - \hat{\beta}_2)|,\end{aligned} \tag{18}$$

where $\hat{\beta}_1 \in \mathcal{B}_1 = \{0, \Delta_{\beta_1}, 2\Delta_{\beta_1}, \ldots, (2^{b_1}-1)\Delta_{\beta_1}\}$ and $\hat{\beta}_2 \in \mathcal{B}_2 = \{0, \Delta_{\beta_2}, 2\Delta_{\beta_2}, \ldots, (2^{b_2} - 1)\Delta_{\beta_2}\}$ are the sets of the quantized phase values. Δ_{β_1} and Δ_{β_2} are the quantization steps depending on the number of quantization bits b_1 and b_2 respectively. As pointed out in [23], when the quantization step is sufficiently small, this method can leads to negligible quantization error.

B. Joint Quantization Using Combined Quantization Codebooks

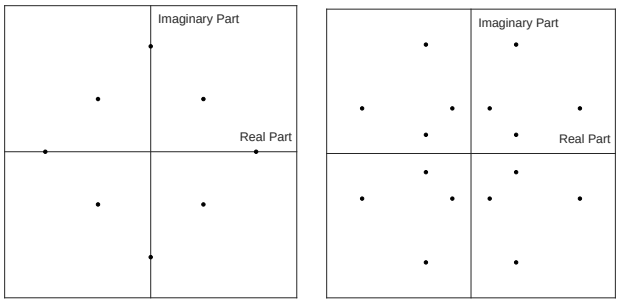
In this scheme, we generate a combined codebook from the two separate codebooks $\mathcal{B}_1$ and $\mathcal{B}_2$ for the two phase shifters. That is, given the code $\hat{\beta}_1 \in \mathcal{B}_1$ and $\hat{\beta}_2 \in \mathcal{B}_2$, we combine and generate a new combined codebook $\mathcal{C}$ with code $\hat{c}$, and $\hat{c} = e^{j\hat{\beta}_1} + e^{j\hat{\beta}_2}$. Note that the codes in $\mathcal{C}$ do not have unit magnitude any more. This scheme will output the same quantization codebook for the two structures in Fig. 2.

Consider a pair of generalized quantization codebooks

$$\begin{aligned}\mathcal{B}_1 &= \{0, \Delta_{\beta_1}, 2\Delta_{\beta_1}, \ldots, (2^{b_1} - 1)\Delta_{\beta_1}\},\\ \mathcal{B}_2 &= \{\phi, \phi + \Delta_{\beta_2}, \ldots, \phi + (2^{b_2} - 1)\Delta_{\beta_2}\},\end{aligned} \tag{19}$$

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., a fixed length of delay line in the circuit. This phase shift can effectively increase the number of codes in the combined codebook as we shall see shortly.

We consider a typical case when the quantization step and bits are the same for the two phase shifters. In this case, it is easy to see that if $\phi = 0$, there will be 2^b repetitive values out of the total 2^{2b} codes in $\mathcal{C}$. The reduced number of distinct codes will lead to increased quantization errors. We can let ϕ be a non-zero value to increase the number of non-identical



(a) Constellation of codebook $\mathcal{C}_1$ with $\phi = 0$. (b) Constellation of codebook $\mathcal{C}_2$ with $\phi = \Delta_{\beta_2}/2$.

Fig. 3. Two types of Constellations when $b = 2$.

codes and reduce the average quantization error. Even when the quantization steps for the two phase shifters are the same, a different ϕ can also lead to different quantization errors.

Hence in principle we can optimize the value of ϕ to get the minimum quantization error. However, the optimization is tricky as both the total number of different codes and their values have notable impact on the final quantization performance. So we leave general optimization of ϕ as an open problem.

In this paper, we will compare two cases with $\phi = 0$ and $\phi = \Delta_{\beta_2}/2$. In the second case, we get $n_{c_2} = 2^{2b}$ different codes. Fig. 3 displays the constellation plot for the combined codebook $\mathcal{C}_1$ and $\mathcal{C}_2$ corresponding to these two cases. These constellation points are normalized so that $E[|\hat{c}_{k,i}|^2] = 1/M$, consistent with the norm of the BF vector, where $\hat{c}_{k,i}$ is the i-th element of $\mathcal{C}_k$. The normalization factor is given by

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

and

$$h_2 = \sqrt{\frac{M}{2^{b}} \sum_{i=1}^{2^{b}} \hat{c}_{k,i}^2} = \sqrt{2M}, \tag{20}$$

for $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively.

For the codebook $\mathcal{C}_k$, we can then apply the element-wise quantization for each BF weight w_i and obtain

$$\hat{w}_i = \arg\min_{\hat{c} \in \mathcal{C}_k} |w_i - \hat{c}_{k,i}|^2. \tag{21}$$

As we are going to show in Section VI, the element-wise quantization algorithm based on our proposed codebooks can already achieve sufficiently good performance with only a small number of quantization bits.

C. Quantization with Optimized Scaling Factor

In the last subsection, our codebook $\mathcal{C}_k$ was normalized to h_k, but h_k are statistical values which cannot guarantee the instantaneous optimality for quantizing a particular BF vector $\mathbf{w}_t$. With the goal of finding a better solution, we propose an algorithm, what we call as IGSS-Q, based on the improved golden section search (IGSS) algorithm [31]. The IGSS algorithm is an efficient one-dimension linear searching method that relaxes the unimodal requirement for the classic golden-section search method.

Algorithm 1 IGSS-Q Algorithm

Input: a_1, a_2, $L_{\max}$, ϵ_0, $\rho = \frac{\sqrt{5}-1}{2}$.
1) $l = 0$, $a_1^{(0)} = a_1$, $a_2^{(0)} = a_2$, $d^{(0)} = a_2^{(0)} - a_1^{(0)}$, $x_1^{(0)} = a_1^{(0)} + (1-\rho)d^{(0)}$, $x_2^{(0)} = a_1^{(0)} + \rho d^{(0)}$; go to 2).
2) $d^{(l)} = a_2^{(l)} - a_1^{(l)}$; If $l \le L_{\max}$ & $|d^{(l)}| > \epsilon_0$, go to 3); otherwise, go to 5).
3) Calculate $e(a_1^{(l)})$, $e(x_1^{(l)})$, $e(x_2^{(l)})$ and $e(a_2^{(l)})$ through (24); 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 [31], (23) and (24). 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 (23) and $\nu_{\min}$.
Output: $\nu_{\min}$, $\hat{\mathbf{q}} = [\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_M]^T$

Our IGSS-Q method solves the following problem

$$\nu_{\text{opt}} = \arg\min_{\nu} \|\nu \mathbf{w}_t - \hat{\mathbf{q}}(\nu)\|_2^2 \tag{22}$$

iteratively. In each iteration, $\hat{\mathbf{q}}(\nu)$ is obtained by scalar quantization with $\mathcal{C}_k$ and the i-th element of $\hat{\mathbf{q}}(\nu)$ is given by

$$\hat{q}_i = \arg\min_{\hat{c} \in \mathcal{C}_k} |\nu w_i - \hat{c}_{k,i}|^2, \tag{23}$$

where $i \in \{1, \cdots, M\}$. For a fixed ν and the $\hat{q}_i$ values obtained in (23), the quantization error $e(\nu)$ can be expressed as

$$e(\nu) = \sum_{i=1}^{M} |\nu w_i - \hat{q}_i|^2. \tag{24}$$

The IGSS-Q method starts with setting the initial searching interval $\nu \in [a_1, a_2]$ and then defines 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 (23) and (24) 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 down 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 times $L_{\max}$ is reached. The detailed process of the IGSS-Q method is provided in Algorithm 1.

The computational complexity of IGSS-Q is low and can be approximately represented as $O(ML_{\max})$, where $L_{\max}$ is the number of iterations. With the output $\hat{\mathbf{q}}$ from the algorithm, we can get the quantized BF vector as

$$\hat{\mathbf{w}}_t = \frac{\hat{\mathbf{q}}}{\|\hat{\mathbf{q}}\|}, \ \hat{\mathbf{q}} = [\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_M]^T. \tag{25}$$

V. Quantization Error Analysis

In this section, we analyze element-wise quantization error for the 1-PS and 2-PS quantization schemes presented in

Section IV. We use the mean squared quantization error (MSQE) as the performance metric, which is defined as

$$\varepsilon = E[\frac{1}{M}\sum_{i=1}^{M}(|w_i - \hat{w}_i|^2)], \tag{26}$$

where $\hat{w}_i$ is the quantized BF weight. For both 1-PS quantization and 2-PS separate quantization, we provide more accurate analytical results compared with [23]. For 2-PS joint quantization, our analytical results are new, as well as the quantization scheme itself.

A. 1-PS Array

When only one phase shift is used to represent a BF weight, the ith element of the quantized BF vector can be represented as $\hat{w}_i = \frac{1}{\sqrt{M}}e^{j(\psi_i+\delta_\psi)}$, where δ_ψ denotes the phase quantization error. Assume that δ_ψ is uncorrelated with ψ and uniformly distributed over $[-\Delta_\psi/2, \Delta_\psi/2)$. The MSQE in this case can be expanded to

$$\begin{aligned}\varepsilon_0 = E(|\hat{w}_i - w_i|^2) &= \int_{-\frac{\Delta_\psi}{2}}^{\frac{\Delta_\psi}{2}} |\hat{w}_i - w_i|^2 \frac{1}{\Delta_\psi} d\delta_\psi \\ &= \frac{1}{M} + E(|w_i|^2) - \frac{2E(|w_i|)}{\sqrt{M}}(\frac{2}{\Delta_\psi}\sin\frac{\Delta_\psi}{2}).\end{aligned} \tag{27}$$

When b is large, and in the extreme case $\Delta_\psi \to 0$, we have

$$\begin{aligned}\lim_{\Delta_\psi\to 0}\varepsilon_0 &= \frac{1}{M} + E(|w_i|^2) - \frac{2E(|w_i|)}{\sqrt{M}} \\ &= E[(|w_i| - \frac{1}{\sqrt{M}})^2] + \text{Var}(|w_i|) \\ &\geq \text{Var}(|w_i|),\end{aligned} \tag{28}$$

where $\text{Var}(|w_i|)$ is the variance of $|w_i|$. This clearly shows that the quantization error does not vanish by only using phase shifting values to represent the BF weights with varying magnitudes. There will be an error floor despite the value of the quantization step.

B. 2-PS with Parallel Structure Using Separate Quantization

When phase shifts are quantized, $\hat{w}_i = e^{j\beta_1^{(i)}+\delta_{\beta_1}} + e^{j\beta_2^{(i)}+\delta_{\beta_2}}$, where δ_{β_1} and δ_{β_2} are the phase quantization errors (referring to (16a)). The MSQE ε_1 is then given by

$$\begin{aligned}\varepsilon_1 =& 4 + 2E[\cos(\beta_1 - \beta_2)] - 2E[\cos(\beta_1 - \beta_2)]\cdot \\ &\{E(\cos\delta_{\beta_1}) - E(\cos\delta_{\beta_2}) - E[\cos(\delta_{\beta_1} - \delta_{\beta_2})]\} \\ &- 2E(\cos\delta_{\beta_1} + \cos\delta_{\beta_2}).\end{aligned} \tag{29}$$

The quantization errors δ_{β_1} and δ_{β_2} are assumed to be uncorrelated and uniformly distributed over $[-\Delta_{\beta_1}/2, \Delta_{\beta_1}/2)$ and $[-\Delta_{\beta_2}/2, \Delta_{\beta_2}/2)$, respectively. It can be calculated that

$$\begin{aligned}&E(\cos\delta_{\beta_1}) = \frac{2}{\Delta_{\beta_1}}\sin(\frac{\Delta_{\beta_1}}{2}), \\ &E(\cos\delta_{\beta_2}) = \frac{2}{\Delta_{\beta_2}}\sin(\frac{\Delta_{\beta_2}}{2}), \\ &E[\cos(\delta_{\beta_1} - \delta_{\beta_2})] = E[\cos(\delta_{\beta_1} + \delta_{\beta_2})] \\ &\qquad = \frac{4}{\Delta_{\beta_1}\Delta_{\beta_2}}\sin(\frac{\Delta_{\beta_1}}{2})\sin(\frac{\Delta_{\beta_2}}{2}), \\ &E(\sin\delta_{\beta_1}) = E(\sin\delta_{\beta_2}) = 0, \\ &E[\sin(\delta_{\beta_1} - \delta_{\beta_2})] = E[\sin(\delta_{\beta_1} + \delta_{\beta_2})] = 0.\end{aligned}$$

Since $E[\cos(\beta_1 - \beta_2)] = E(\frac{|w_i|^2 - 1}{2})$, (29) can be expressed as

$$\begin{aligned}\varepsilon_1 =& 3 + E(|w_i|^2) + E(|w_i|^2)[\frac{4}{\Delta_{\beta_1}\Delta_{\beta_2}}\sin(\frac{\Delta_{\beta_1}}{2})\sin(\frac{\Delta_{\beta_2}}{2}) \\ &- \frac{2}{\Delta_{\beta_1}}\sin(\frac{\Delta_{\beta_1}}{2}) - \frac{2}{\Delta_{\beta_2}}\sin(\frac{\Delta_{\beta_2}}{2})] - \frac{2}{\Delta_{\beta_1}}\sin(\frac{\Delta_{\beta_1}}{2}) \\ &- \frac{4}{\Delta_{\beta_1}\Delta_{\beta_2}}\sin(\frac{\Delta_{\beta_1}}{2})\sin(\frac{\Delta_{\beta_2}}{2}) - \frac{2}{\Delta_{\beta_2}}\sin(\frac{\Delta_{\beta_2}}{2}).\end{aligned} \tag{30}$$

When x is small, the function $\sin x$ can be approximated as $\sin x \approx x - \frac{x^3}{6}$ using the Taylor series expansion for $\sin x$. To this end, (30) can be approximated as

$$\varepsilon_1 \approx \frac{\Delta_{\beta_1}^2}{12} + \frac{\Delta_{\beta_2}^2}{12} - (E(|w_i|^2) - 1)\frac{\Delta_{\beta_1}^2\Delta_{\beta_2}^2}{576}. \tag{31}$$

Note that the first two terms in (31) are just the results in [23]. Our result here provides a closer approximation to the MSQE ε_1.

In the extreme case of $\Delta_{\beta_1} \to 0$, $\Delta_{\beta_2} \to 0$, we can get

$$\lim_{\Delta_{\beta_1},\ \Delta_{\beta_2}\to 0}\varepsilon_1 = 0. \tag{32}$$

This shows that with the 2-PS parallel structure, the MSQE can reach zero when the quantization bit b is sufficiently large, even when the actual BF weights have varying amplitudes.

C. 2-PS with Serial Structure Using Separate Search

For the 2-PS array with the serial structure, $\hat{w}_i = e^{j(\beta_1^{(i)}+\delta_{\beta_1})}[1 + e^{j(\beta_2^{(i)}+\delta_{\beta_2})}]$. Thus, the MSQE ε_2 can be written as

$$\begin{aligned}\varepsilon_2 =& 4 + 2E(\cos\beta_2)[1 - \frac{2}{\Delta_{\beta_1}}\sin(\frac{\Delta_{\beta_1}}{2}) + \frac{2}{\Delta_{\beta_2}}\sin(\frac{\Delta_{\beta_2}}{2})] \\ &- 2E(\cos\beta_2 + 1)\frac{4}{\Delta_{\beta_1}\Delta_{\beta_2}}\sin(\frac{\Delta_{\beta_1}}{2})\sin(\frac{\Delta_{\beta_2}}{2}) \\ &- \frac{4}{\Delta_{\beta_1}}\sin(\frac{\Delta_{\beta_1}}{2}).\end{aligned}$$

Since $E(\cos(\beta_2)) = E(\frac{|w_i|^2}{2} - 1)$, ε_2 can be rewritten as

$$\begin{aligned}\varepsilon_2 =& 2 + E(|w_i|^2)[1 - \frac{2}{\Delta_{\beta_1}}\sin(\frac{\Delta_{\beta_1}}{2}) + \frac{2}{\Delta_{\beta_2}}\sin(\frac{\Delta_{\beta_2}}{2}) \\ &- \frac{4}{\Delta_{\beta_1}\Delta_{\beta_2}}\sin(\frac{\Delta_{\beta_1}}{2})\sin(\frac{\Delta_{\beta_2}}{2})] - \frac{4}{\Delta_{\beta_2}}\sin(\frac{\Delta_{\beta_2}}{2}).\end{aligned}$$

Using $\sin x \approx x - \frac{x^3}{6}$, we can approximate ε_2 as

$$\varepsilon_2 \approx \frac{1+E(|w_i|^2)}{12}\Delta_{\beta_1}^2 - \frac{E(|w_i|^2)}{576}\Delta_{\beta_1}^2\Delta_{\beta_2}^2. \quad (33)$$

Similarly, we can show that

$$\lim_{\Delta_{\beta_1},\, \Delta_{\beta_2}\to 0} \varepsilon_2 = 0. \quad (34)$$

D. 2-PS Using Joint Quantization

We consider the case when the two phase shifters have the same number of quantization bits.

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 $a \le |\hat{c}|_{\max}$ since the probability of $a > |\hat{c}|_{\max}$ is low. The MSQE metric for joint quantization becomes

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

When element-wise scalar quantization is used, $\nu_{\min} = 1$.

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 $\delta_{c,\max} = \frac{\sqrt{2-2\cos\Delta}}{\nu_{1,\min}}$. The corresponding MSQE is

$$\varepsilon_{c1} = [E(|\delta_c|)]^2 + \text{Var}(|\delta_c|) = \frac{2-2\cos\Delta}{3E[\nu_{1,\min}^2]}. \quad (35)$$

When the quantization step is small, $1-\cos\Delta \sim \frac{\Delta^2}{2}$, and (35) can be approximated as

$$\varepsilon_{c1} \approx \frac{\Delta^2}{3E[\nu_{1,\min}^2]} \approx \frac{\Delta^2}{3h_1^2} = \frac{\Delta^2}{3(2+2^{2-b})M}. \quad (36)$$

Note that (36) is also the MSQE for element-wise scalar quantization using codebook $\mathcal{C}_1$. This implies that scalar quantization using the proposed codebook can achieve comparable performance to IGSS-Q, when quantization bits are sufficiently large.

2) MSQE for Codebook $\mathcal{C}_2$: For codebook $\mathcal{C}_2$ with $\phi = \Delta/2$, we show in Appendix D that $\delta_{c,\max} = \frac{\sqrt{2-2\cos\frac{\Delta}{2}}}{\nu_{2,\min}}$. The MSQE in this case can be obtained as

$$\varepsilon_{c2} = \frac{2-2\cos(\frac{\Delta}{2})}{3E[\nu_{2,\min}^2]}. \quad (37)$$

When Δ is small, $1-\cos\frac{\Delta}{2} \sim \frac{\Delta^2}{8}$, and (37) can be approximated as

$$\varepsilon_{c2} \approx \frac{\Delta^2}{3h_2^2} = \frac{\Delta^2}{24M}. \quad (38)$$

Similarly, (38) is also the MSQE for element-wise scalar quantization using codebook $\mathcal{C}_2$.

When $\Delta \to 0$, we have

$$\lim_{\Delta\to 0}\varepsilon_{c1} = 0, \lim_{\Delta\to 0}\varepsilon_{c2} = 0. \quad (39)$$

E. Comparison of MSQE for Different Quantization Methods

Referring to the approximated values of MSQE in (31), (33), (36), and (38), we compare the performance for these quantization methods.

Since $||\mathbf{w}_t|| = 1$, we see $0 < |w_i| < 1$. According to (31) and (33), it can be found that $\varepsilon_2 < \varepsilon_1$, which indicates that for 2-PS, separate quantization using the parallel structure can achieve slightly better performance than using the serial structure, when the quantization step is reasonably small.

From (20), we can find that for $b \ge 2$, $\sqrt{2M} < h_1 \le \sqrt{3M}$. Therefore, ε_{c1} satisfies

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

For large arrays with more than, e.g., $M = 8$ antennas, it can be readily verified that

$$\varepsilon_{c2} < \varepsilon_{c1} < \varepsilon_2 < \varepsilon_1. \quad (40)$$

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

In Table I, we summarize the comparison results for these quantization methods.

VI. Simulation Results

In this section, simulation results are presented to verify the proposed combination and quantization methods. For all simulations, a uniform linear array with $M = 16$ omnidirectional antennas (spaced at half wavelength) is used. We assume that the basic reference beam for communication and sensing are pointed at zero degree. The 3dB beamwidth for a linear array with K_s antennas is approximately $2\arcsin(\frac{1.2}{K_s})$ in radius. We generate the basic beams with $K_s = 16$ and $K_s = 12$ for the communication and sensing subbeams, respectively. The reason for using a wider subbeam for scanning is to cover the sensing directions from -60 to 60 degrees with fewer times of scanning. The desired actual pointing directions of the 8 scanning subbeams is at -54.3, -37.8, -24.4, -12.3, 10.8, 22.8, 35.9 and 51.9 degrees. Note the nonuniform actual scanning directions are because of the requirement of applying the simple displaced BF waveform generation method as described in [6]. The power distribution factor ρ is set as 0.5 unless noted otherwise.

In the simulation, $\mathbf{w}_{t,c}$ is set pointing to the dominating AoD, and $\mathbf{w}_{t,s}$ is generated by multiplying a phase-shifting sequence to the basic sensing subbeam to change the pointing directions, as described in [6]. For all results on the received signal power, an MRC receiver is assumed to be used, and they are normalized to the power value when the whole transmitter array generates a single beam pointing to the dominating AoD.

Assume there is an LOS path between the transmitting and receiving nodes for communication. All the other multipath components are uniformly distributed within an angular range of 14 degrees centered at the LOS direction. The mean power ratio between the LOS and the NLOS signals is 10dB.

TABLE I
COMPARISON OF PHASE SHIFTER STRUCTURES AND DIFFERENT SEARCHING METHODS.

	1-PS Array	2-PS Arrays		
		Separate Search	Joint Search	
MSQE	ε_0	$\varepsilon_1 > \varepsilon_2$	$\varepsilon_{c1} > \varepsilon_{c2}$, lowest overall	
$\lim_{b\to\infty} \varepsilon_i, i = 0, 1, 2$	$E[(\lvert w_i\rvert - \frac{1}{M})^2] + D(\lvert w_i\rvert) > 0$	0		
Number of Phase Shifters	M	2M		
Hardware Complexity	normal	relatively complex		
Drawbacks	large error floor exists	large error	more constellation points to be compared and stored	
Computation Complexity	$O(2^b M)$	$O(2^{b+1} M)$	$O(2^{n_c} M)$ (Scalar)	$O(2^{n_c} M^2)$ (IGSS-Q)

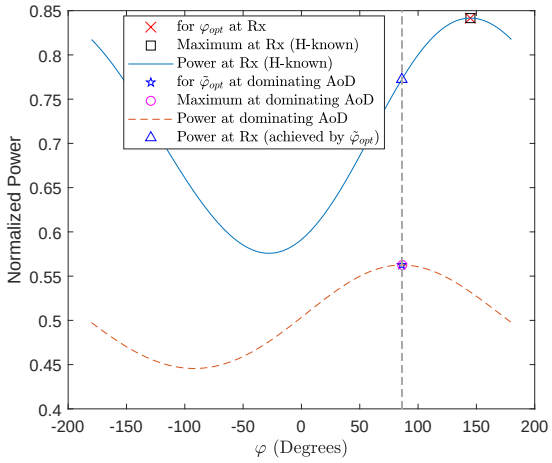


Fig. 4. Normalized signal power at the receiver (Rx) and at the dominating AoD versus combining phase φ for a fixed sensing subbeam pointing at $10.8°$, for a random channel realization.

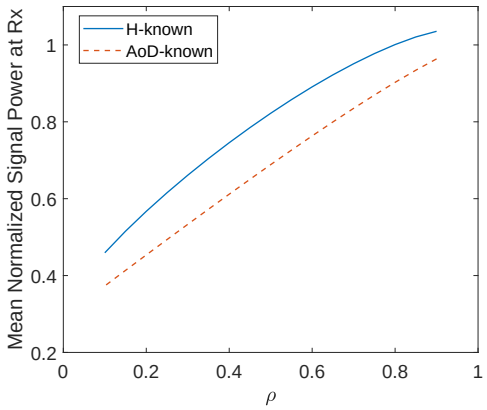


Fig. 5. Normalized mean received signal power versus power distribution factor ρ for optimized φ when the sensing subbeam points to $10.8°$.

A. Non-quantized Results

We first demonstrate the effect of improved received signal power via optimizing the combining coefficient value as proposed in Section III. Non-quantized BF vector is used. Denote the cases when the full channel or only the dominating AoD is known at transmitter as "H-known" and "AoD-known", respectively.

In Fig. 4, we present the signal power at the receiver and at the dominating AoD $\|\mathbf{a}^T(\theta_t)\tilde{\mathbf{w}}_t\|^2$ with varying phase values φ, when the scanning beam points to $10.8°$. The optimal values obtained by our solutions for "H-known" and "AoD-known", together with the actual one via exhaustive search, are also shown for comparison. We can see that there is up to about 30% variation of the power at receiver and 20% variation at the dominating AoD between the optimal and non-optimal phase values, and the derived optimal phase values match the actual ones very well.

Fig. 5 demonstrates how the normalized mean received signal power varies with the value of the power distribution factor ρ when the optimized combining phase values are used. The figure shows that the proposed optimization methods efficiently increase the received signal power, almost linearly with the increasing of ρ.

In Fig. 6, we use a dual y-axis plot to illustrate how the mean normalized received power (left y-axis, curves are denoted as "Power") and the MSEs of BF waveform (right y-axis, curves are denoted as "MSE") for the benchmark BF design in (6). The vector $\mathbf{w}^\star_{\text{WF}}$ in (6) is generated using Method 2 in [6], and $\mathbf{w}^\star_{\text{Gain}}$ generates the beam pointing to the dominating communication AoD. Although $\mathbf{w}^\star_{\text{Gain}}$ is unrelated to the power distribution factor ρ, $\mathbf{w}^\star_{\text{WF}}$ is a function of ρ and hence both the received power and MSE are affected by ρ. Each BF waveform is normalized to its peak value before the MSE is evaluated. This figure demonstrates that, for most values of λ, our proposed scheme outperforms the benchmark in either signal power or waveform conformity. There does exist a small range of λ values where the benchmark can achieve better performance in terms of both metrics. Note that, however, finding such better solutions is computational intensive, involving exhaustive search of λ, and ILS with iterative matrix inversion operations for generating each new $\mathbf{w}^\star_{\text{WF}}$.

Fig. 7 shows how the MSE of BF waveform and normalized received signal power change with the scanning directions for several methods, where we specifically study the case when scanning directions are close to communication. Each result is averaged over 5000 channel realizations. These methods include deriving the combining phase from (7) (i.e., Method 1 in [6] and denoted as "M1 in [6]"), the two proposed optimization methods with "H-known" and "AoD-known", the joint design method (Method 2) in [6] (denoted as "M2 in [6]"), and the benchmark design. The λ value for the benchmark design is decided by first finding a set of λ values that lead to larger power than the one obtained by "AoD-known", and then selecting the one from the set leading to the minimum MSE.

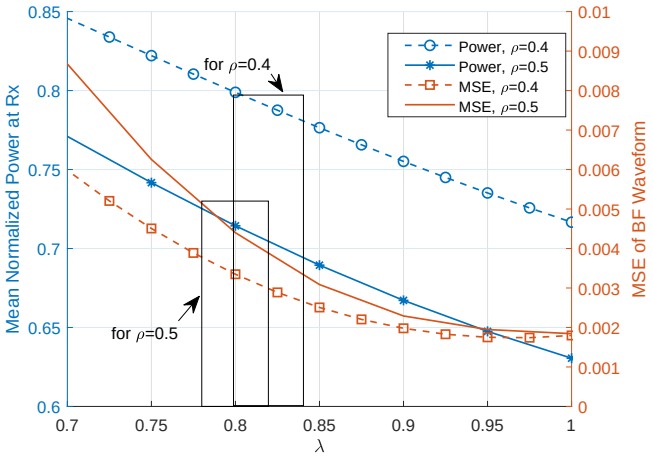


Fig. 6. Comparison with the benchmarking balanced BF design in (6) when the sensing subbeam points to $10.8°$. The curves with legends of "Power" and "MSE" represent the mean normalized received power (left y-axis) and the MSE of waveform (right y-axis), respectively. The dash and solid curves are for $\rho = 0.4$ and $\rho = 0.5$, respectively. For $\rho = 0.4$ and $\rho = 0.5$, the mean received power values for using our proposed combining phase with AoD-known are 0.7766 and 0.7023, respectively, and the MSEs of BF waveform are 0.5123 and 0.6629. The range of λ values within each black square box represent where the benchmarking design can achieve better performance in both signal power and waveform for one ρ.

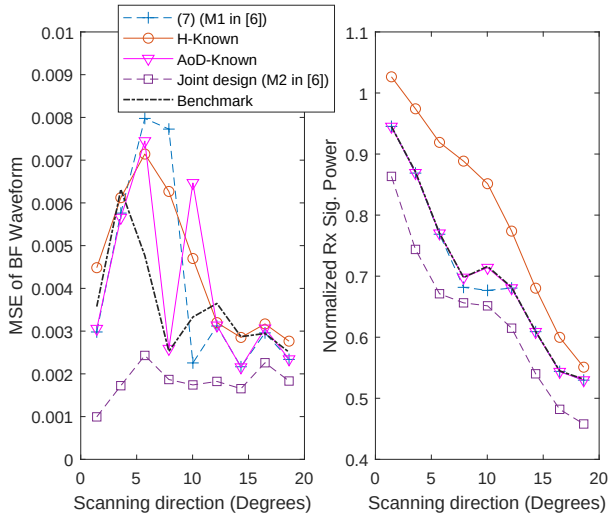
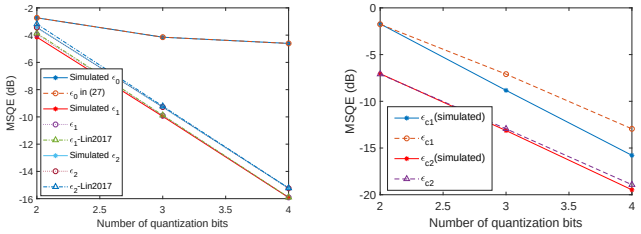


Fig. 7. Variation of MSE of BF waveform and normalized mean received signal power with the scanning directions of sensing subbeam.



(a) (b)

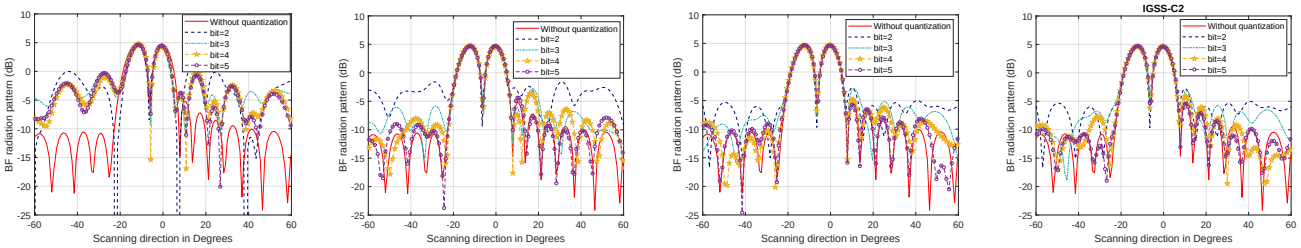
Fig. 8. MSQE versus number of quantization bits for various quantization schemes.

The communication subbeam always points to the dominating AoD at 0 degree. From the figure, no monotonicity can be observed between the MSE of BF waveform and the interval between the pointing directions of scanning and communication subbeam. The MSE gaps between these schemes, except for the joint design, are typically small. The received signal power generally decreases with the increase of the interval, as expected. The proposed methods improve the received signal power. The power gap between our proposed methods and the design in (7) is generally small. This is partially due to the large power ratio (10dB) between the dominating LOS path and NLOS paths used in the simulation. When the ratio reduces to 5dB, we observe approximately a 25% gap in the experiments. In the experiments, we also observe that the range of λ, where the benchmark design outperforms our methods, varies with the scanning direction and its span is mostly smaller than 0.1. Details are not presented here due to page limits.

B. Quantized Results

In Fig. 8, we show MSQE versus the number of quantization bits for various quantization methods. Apart from those studied and analysed in this paper, we also compute and plot the MSQE in [23], which is denoted as "ε_x-Lin2017" in the legend. The values of $w_i = |w_i|e^{j\psi_i}$ in $\mathbf{w}_t$ are generated randomly with $|w_i|$ following a uniform distribution over $[0, 2)$ and with ψ_i following a uniform distribution over $[0, 2\pi)$. The vector of $\mathbf{w}_t$ is then normalized so that its norm is 1. The MSQE is averaged over 10^5 realizations for the simulated values. From the figure, we can see that most of the analytical results match the simulated ones very well, except for ε_{c1}. The analytical results provided in this paper are shown to offer better accuracy than the one in [23], although the difference is small for this simulated example. When $b > 2$, the simulated ε_{c1} deviates from the analytical ε_{c1} derived in Section V-D. This is because for Codebook $\mathcal{C}_1$, many of the largest distances between any two nearest constellation points are smaller than $\delta_{c,\max}$. Therefore, the uniform distribution assumption in SectionV-D is not accurate enough. Overall, the joint quantization method using Codebook $\mathcal{C}_2$ achieves the lowest quantization error, as we have shown analytically in (40).

Fig. 9 shows how the BF radiation pattern varies with the number of quantization bits for our proposed joint quantization schemes, together with 1-PS vector quantization using the fast block noncoherent decoding (FBND) method [29] for comparison. From Fig. 9(a), we can see that for 1-PS, the sidelobe of the waveform for the quantized BF vector is quite large, even when the number of quantization bits is as large as 5. There is also a notable distortion in the mainlobe. So vector quantization cannot either improve the quantization error floor for the case with only quantized phase values. Comparatively, joint quantization for 2-PS achieves good match in the mainlobe with the non-quantized one and much lower sidelobe, as can be seen from Fig. 9(b), 9(c) and 9(d). For example, when the number of quantization bits is larger than 3, the power level of the sidelobe of the quantized results with codebook $\mathcal{C}_2$ is



(a) FBND quantization for 1-PS. (b) Joint quantization for 2-PS with Codebook $\mathcal{C}_1$. (c) Joint quantization for 2-PS with Codebook $\mathcal{C}_2$. (d) IGSS-Q for 2-PS with Codebook $\mathcal{C}_2$.

Fig. 9. BF radiation pattern for different number of quantization bits. The scanning beam points to $-12.3°$.

very close to the non-quantized one. It can also be observed that IGSS-Q can slightly reduce the sidelobe when $b \geq 4$, compared with element-wise scalar quantization.

We now further look at the impact of quantization on the received signal power and signal demodulation performance.

In Fig. 10, we show the received signal power for the 1-PS FBND quantization and the joint IGSS-Q method with Codebooks $\mathcal{C}_1$ and $\mathcal{C}_2$ (denoted as "Codebook 1" and "Codebook 2"), when the total number of quantization bits in each method is the same. Even in this case, the received signal power achieved by the two joint quantization methods is still larger at the angles close to the communication direction.

Fig. 11 shows the estimated average bit error rate (BER) for different quantization methods with $b = 4$ in the case where $\mathbf{H}$ is known. 16 quadrature amplitude modulation (QAM) is used. We can see that BER performance benefits significantly from the array gain, and our proposed joint quantization methods achieve BER approaching the non-quantized case. Between the proposed methods, the vector-wise IGSS-Q methods can slightly outperform the element-wise scalar quantization methods. An SNR loss of about 2dB can be observed at BER $= 10^{-3}$ compared to the case where all energy is used for communication. This implies an approximately 1dB gain with the use of our proposed phase combining coefficients and quantization methods. Otherwise, a loss of 3dB or more may occur due to the equal split of power between sensing and communication subbeams.

To summarize, in terms of both BF waveform and the received signal power and demodulation performance, we can conclude that the proposed joint quantization method with codebook $\mathcal{C}_2$ can achieve performance approaching to the non-quantized one.

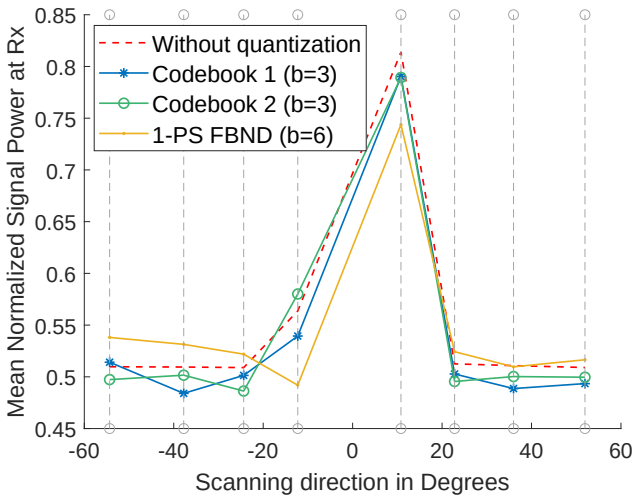


Fig. 10. Comparison of normalized received signal power for 2-PS using IGSS-Q method and 1-PS using FBND method with the same number of 6 total quantization bits.

VII. Conclusions

We have now studied the optimization of the coefficient for combining communication and sensing BF vectors and then the quantization of the combined BF vector, for the multibeam JCAS scheme proposed in [6]. We considered the cases when either the full channel knowledge or the dominating AoD is known, and provided closed-form expressions for the optimal combining coefficients that maximizes the received signal power in each case. Considering the practical constraint that BF weights in analog arrays are of discrete values, we investigated various element-wise quantization methods, particularly for the structures where two phase shifters are used to represent one BF weight. We proposed a novel joint quantization method using a combined codebook for the two phase shifters (2-PS). This method is shown to achieve BF waveform closely matching the one with non-quantized BF vector, as well as the received signal power, using a medium number of quantization bits. We also provided analytical expressions of the mean squared quantization error (MSQE) for these quantization methods. Simulation results match these analytical MSQE results well. Overall, the joint quantization method can approach the performance of non-quantized BF vector, and hence is very promising for the multibeam JCAS system.

The work in this paper can be enriched in various aspects. For example, the joint quantization method can be further improved by exploiting the noncoherent decoding methods [30], [32], and the underlying analog array can be replaced by more powerful hybrid arrays [8]. Although our proposed BF optimization methods generally cause insignificant BF waveform variation, a BF optimization scheme that directly takes the sensing requirement into formulation can lead to more accurate control of BF waveform. Our method can potentially be extended to the case where more than one sensing subbeam is generated to speed up scanning, at the cost of a reduced power allocated to each subbeam and

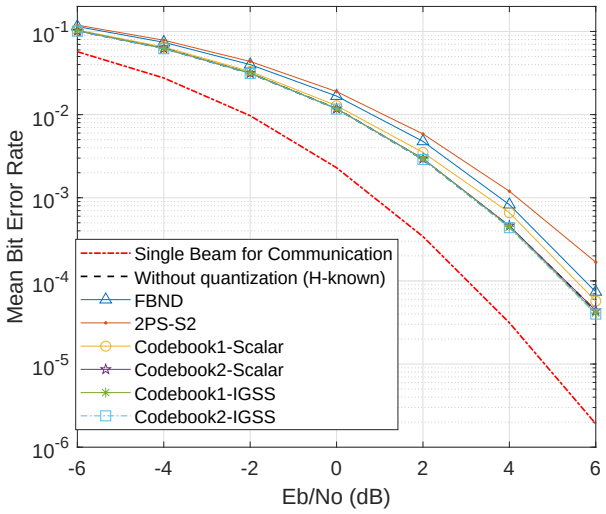


Fig. 11. BER for different methods with $b = 4$ (quantization bits) and the sensing subbeam pointing at $-12.3°$.

hence reduced sensing distance. Depending on the scattering environment, multiple combining coefficients may need to be optimized for communication, but the impact of combining coefficients on sensing is still limited to the sensing subbeam adjacent to the communication subbeam.

Appendix A
Derivation of φ_{opt}

We study the monotonicity of $f(\varphi)$ and try to look for its maximum via its derivatives. The first-order derivative of $f(\varphi)$ with respect to φ is

$$f'(\varphi) = \frac{g_1'(\varphi)g_2(\varphi) - g_2'(\varphi)g_1(\varphi)}{g_2(\varphi)^2}. \tag{41}$$

Obviously, $g_2^2(\varphi) > 0$. Let the numerator in (41) be $h(\varphi)$, and let $\mathbf{w}_{t,c}^H\mathbf{H}^H\mathbf{H}\mathbf{w}_{t,s} = a_1e^{j\alpha_1}$ and $\mathbf{w}_{t,c}^H\mathbf{w}_{t,s} = a_2e^{j\alpha_2}$, where $a_1 \geq 0$ and $a_2 \geq 0$. We have

$$\begin{aligned} h(\varphi) &= -2Pa_1\sin(\varphi+\alpha_1) - 4P^2a_1a_2\sin(\alpha_1-\alpha_2) + \\ &\quad 2Pa_2[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]\sin(\varphi+\alpha_2) \\ &= X_1\sin(\varphi) + X_2\cos(\varphi) + L, \end{aligned}$$

where $X1$, $X2$, L are given in (12). By considering the sign of X_1, $h(\varphi)$ can be represented as

$$h(\varphi) = \begin{cases} \sqrt{X_1^2+X_2^2}\sin(\varphi+\gamma) + L, & \text{when } X_1 \geq 0 \\ -\sqrt{X_1^2+X_2^2}\sin(\varphi+\gamma) + L, & \text{when } X_1 < 0, \end{cases}$$

where $\gamma = \arctan(X_2/X_1)$.

Since $h(\varphi)$ is a periodic function and the period is 2π, we study the monotonicity of $f(\varphi)$ in one period. During a period of length π, $f(\varphi)$ keeps increasing if $h(\varphi) > 0$, and keeps decreasing otherwise. So at the transition point where $h(\varphi) = 0$, we can obtain either the maximum or minimum of $f(\varphi)$. From $h(\varphi) = 0$, we can get

$$\varphi = \begin{cases} -\mu_0 - \gamma, & \text{when } X_1 \geq 0 \\ \mu_0 - \gamma, & \text{when } X_1 < 0, \end{cases} \tag{42}$$

where $\mu_0 \triangleq \arcsin\left(\frac{L}{\sqrt{X_1^2+X_2^2}}\right)$.

To make sure that μ_0 exists, $|L| \leq \sqrt{X_1^2+X_2^2}$ needs to be satisfied. In Appendix B, we prove that this is always guaranteed.

By studying the monotonic of $h(\varphi)$, the maximum of $f(\varphi)$ can then be found, as shown in (11).

Appendix B
Existence of μ_0

From (11), it can be observed that if μ_0 exists, the existence of φ_{opt} is guaranteed. Therefore, we need to prove

$$\frac{L^2}{X_1^2+X_2^2} \leqslant 1. \tag{43}$$

That is

$$\begin{aligned} &4P^2|a_1|^2|a_2|^2\sin^2(\alpha_1-\alpha_2) \leqslant \\ &|a_1|^2 + |a_2|^2[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]^2 \\ &- 2|a_1||a_2|[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]\cos(\alpha_1-\alpha_2). \end{aligned}$$

The right part of the inequality can be converted to

$$\begin{aligned} &|a_1|^2 + |a_2|^2[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]^2\cos^2(\alpha_1-\alpha_2) \\ &- 2|a_1||a_2|[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]\cos(\alpha_1-\alpha_2) \\ &- |a_2|^2[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]^2\cos^2(\alpha_1-\alpha_2) \\ &+ |a_2|^2[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]^2 = \\ &\underbrace{[|a_1| - |a_2|[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]cos(\alpha_1-\alpha_2)]^2}_{①} \\ &+ \underbrace{|a_2|^2[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]^2\sin^2(\alpha_1-\alpha_2)}_{②}. \end{aligned}$$

It can be easily verified that the term ① $\geqslant 0$. For the term ②,

$$\begin{aligned} &|a_2|^2[\rho\|\mathbf{H}\mathbf{w}_{t,c}\|^2 + (1-\rho)\|\mathbf{H}\mathbf{w}_{t,s}\|^2]^2\sin^2(\alpha_1-\alpha_2) \\ &\geqslant |a_2|^2[2\sqrt{\rho}\sqrt{1-\rho}\|\mathbf{H}\mathbf{w}_{t,c}\|^2\|\mathbf{H}\mathbf{w}_{t,s}\|^2]^2 sin^2(\alpha_1-\alpha_2) \\ &\geqslant 4P^2|a_1|^2|a_2|^2\sin^2(\alpha_1-\alpha_2). \end{aligned}$$

Therefore, (43) is proven.

Appendix C
Derivation of $\tilde{\varphi}_{\text{opt}}$

Similar to Appendix A, we evaluate the sign of $\tilde{h}(\varphi) = \tilde{g}_1'(\varphi)\tilde{g}_2(\varphi) - \tilde{g}_2'(\varphi)\tilde{g}_1(\varphi)$ for different values of φ. Let $\mathbf{w}_{t,c}^H\mathbf{w}_{t,s} = \tilde{a}_1e^{j\tilde{\alpha}_1}$, $\mathbf{w}_{t,c}^H\mathbf{a}^* = \tilde{a}_2e^{j\tilde{\alpha}_2}$, and $\mathbf{a}^T\mathbf{w}_{t,s} = \tilde{a}_3e^{j\tilde{\alpha}_3}$, where $\tilde{a}_1, \tilde{a}_2, \tilde{a}_3 \geq 0$, we can define

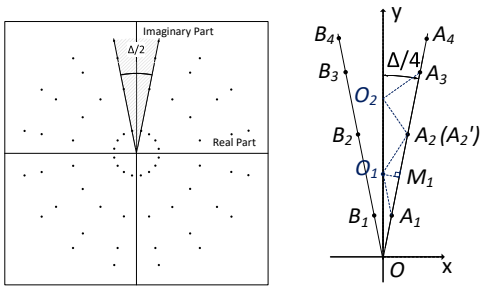
$$\tilde{h}(\varphi) = \tilde{X}_1\sin(\varphi) + \tilde{X}_2\cos(\varphi) + \tilde{L},$$

where $\tilde{X}_1$, $\tilde{X}_2$, $\tilde{L}$ are defined in (14).

Thus $\tilde{h}(\varphi)$ can be further written as

$$\tilde{h}(\varphi) = \begin{cases} \sqrt{\tilde{X}_1^2+\tilde{X}_2^2}\sin(\varphi+\gamma) + L, & \text{when } \tilde{X}_1 \geq 0 \\ -\sqrt{\tilde{X}_1^2+\tilde{X}_2^2}\sin(\varphi+\gamma) + L, & \text{when } \tilde{X}_1 < 0, \end{cases}$$

where $\tilde{\gamma} \triangleq \arctan(\tilde{X}_2/\tilde{X}_1)$. Applying the derivation similar to that in Section III-B and Appendix A, we can obtain $\tilde{\varphi}_{\text{opt}}$ in (13). The existence of $\tilde{\mu}_0$ can be proven using the similar process to Appendix B, and hence is omitted here.



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

Fig. 12. Constellation plot used for analyzing $d_{\max}$ for Codebook $\mathcal{C}_2$ ($b = 3$).

APPENDIX D
COMPUTATION OF $\delta_{c,\max}$ FOR CODEBOOK $\mathcal{C}_2$

Owing to the symmetry of the constellation, we can select a circular segment (the shaded region in Fig. 12(a)) to analyze $\delta_{c,\max}$. 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. 12(b). Obviously, every O_i locates on the y-axis. Note that for the simplicity, in the following analysis, we let the normalization factor $h_2 = 1$, which is independent of the position relationship between constellation points. According to (16), it can be easily proven that for the constellation points A_i,

$$|A_iO| = r_{i-1} = \sqrt{2 - 2\cos\left[(i-1)\Delta + \frac{\Delta}{2}\right]},\ i = 1, 2, \cdots \tag{44}$$

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,

$$\begin{aligned}|A_2'O| &= |OM_1| + |A_2M| = |OM_1| + (|OM_1| - |A_1O|)\\ &= 4r_0\cos^2(\frac{\Delta}{4}) - r_0 = \sqrt{2 - 2\cos(\frac{\Delta}{2})}(2\cos(\frac{\Delta}{2}) + 1)\\ &= \sqrt{2 - 8\cos^3(\frac{\Delta}{2}) + 6\cos(\frac{\Delta}{2})}.\end{aligned} \tag{45}$$

Since $\cos(\frac{3\Delta}{2}) = 4\cos^3(\frac{\Delta}{2}) - 3\cos(\frac{\Delta}{2})$, from (44) we get

$$|A_2O| = \sqrt{2 - 2\cos(\frac{3\Delta}{2})} = \sqrt{2 - 8\cos^3(\frac{\Delta}{2}) + 6\cos(\frac{\Delta}{2})}.$$

We can then find that A_2' overlaps with A_2. This implies that in the sector AOB, $d \le \delta_{c,\max} = r_0$.

According to the Cosine law, if $O_i(y_ie^{j2\pi})$ satisfies $|A_iO_i| = |A_{i+1}O_i|$, there 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 $y_i = \dfrac{r_{i+1} + r_i}{2\cos(\frac{\Delta}{4})}$. Therefore,

$$|A_iO_i|^2 = \frac{r_{i+1}^2 + r_i^2 - 2r_ir_{i+1}\cos(\frac{\Delta}{2})}{2\cos(\frac{\Delta}{2}) + 2}. \tag{46}$$

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

$$r_{i+1}^2 + r_i^2 - 2r_ir_{i+1}\cos(\frac{\Delta}{2}) = 2 - 2\cos\Delta. \tag{47}$$

According to the Cosine law again, the left part of (47) 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:

$$\begin{aligned}|A_iB_{i+1}| &= |(1 + e^{j(\frac{\Delta}{2} + i\Delta)}) - (1 + e^{j(\frac{\Delta}{2} + i\Delta + \Delta)})|\\ &= |e^{j\Delta} - 1| = 2 - 2\cos\Delta.\end{aligned} \tag{48}$$

Therefore, $|A_iO_i|^2 = r_0^2$ is proven. In summary, for all the points distributed inside the outermost layer of the constellation points, the maximal error distance $\delta_{c,\max} = r_0 = \left(\sqrt{2 - 2\cos\frac{\Delta}{2}}\right)/h_2$.

REFERENCES

[1] C. Sturm and W. Wiesbeck, "Waveform design and signal processing aspects for fusion of wireless communications and radar sensing," *Proc. IEEE*, vol. 99, no. 7, pp. 1236–1259, 2011.

[2] L. Han and K. Wu, "Joint wireless communication and radar sensing systems–state of the art and future prospects," *IET Microw. Antennas Propag.*, vol. 7, no. 11, pp. 876–885, 2013.

[3] P. Kumari, J. Choi, N. González-Prelcic, and R. W. Heath, "IEEE 802.11 ad-based radar: An approach to joint vehicular communication-radar system," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3012–3027, 2018.

[4] X. Wang, A. Hassanien, and M. Amin, "Dual-function MIMO Radar communications system design via sparse array optimization," *IEEE Trans. on Aerospace and Electronic Systems*, vol. PP, pp. 1–1, 08 2018.

[5] F. Liu, C. Masouros, A. Li, H. Sun, and L. Hanzo, "MU-MIMO communications with MIMO radar: From co-existence to joint transmission," vol. 17, no. 4, pp. 2755–2770, 2018.

[6] J. A. Zhang, X. Huang, Y. J. Guo, J. Yuan, and R. W. Heath, "Multibeam for joint communication and radar sensing using steerable analog antenna arrays," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 671–685, 2019.

[7] N. González-Prelcic, R. Méndez-Rial, and R. W. Heath, "Radar aided beam alignment in mmWave V2I communications supporting antenna diversity," in *Inform. Theory Appl. Workshop (ITA), 2016*. IEEE, 2016, pp. 1–7.

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

[9] P. Kumari, K. U. Mazher, A. Mezghani, and R. W. Heath, "Low resolution sampling for joint millimeter-wave mimo communication-radar," in *2018 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2018, pp. 193–197.

[10] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 5, no. 2, pp. 4–24, 1988.

[11] L. C. Godara, "Application of antenna arrays to mobile communications. II. Beam-forming and direction-of-arrival considerations," *Proc. IEEE*, vol. 85, no. 8, pp. 1195–1245, 1997.

[12] B. Wang, F. Gao, S. Jin, H. Lin, and G. Y. Li, "Spatial-and frequency-wideband effects in millimeter-wave massive MIMO systems," vol. 66, no. 13, pp. 3393–3406, 2018.

[13] L. Yan, X. Fang, H. Li, and C. Li, "An mmWave wireless communication and radar detection integrated network for railways," in *Veh. Technol. Conf. (VTC Spring), 2016 IEEE 83rd*. IEEE, 2016, pp. 1–5.

[14] V. Va and R. W. Heath, "Performance analysis of beam sweeping in millimeter wave assuming noise and imperfect antenna patterns," in *Veh. Technol. Conf. (VTC-Fall), 2016 IEEE 84th*. IEEE, 2016, pp. 1–5.

[15] F. Liu, L. Zhou, C. Masouros, A. Li, W. Luo, and A. Petropulu, "Toward dual-functional radar-communication systems: Optimal waveform design," vol. 66, no. 16, pp. 4264–4279, 2018.

[16] F. Zhou, Z. Chu, H. Sun, R. Q. Hu, and L. Hanzo, "Artificial noise aided secure cognitive beamforming for cooperative MISO-NOMA using SWIPT," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 4, pp. 918–931, 2018.

[17] F. Zhou, Z. Li, J. Cheng, Q. Li, and J. Si, "Robust AN-aided beamforming and power splitting design for secure MISO cognitive radio with SWIPT," vol. 16, no. 4, pp. 2450–2464, 2017.

[18] J. A. Zhang, A. Cantoni, X. Huang, Y. J. Guo, and R. W. Heath Jr, "Joint communications and sensing using two steerable analog atenna arrays," in *IEEE Veh. Technol. Conf., 2017*. IEEE, 2017, pp. 1–5.

[19] C. Masouros and G. Zheng, "Exploiting known interference as green signal power for downlink beamforming optimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 14, pp. 3628–3640, July 2015.

[20] F. Liu, C. Masouros, A. Li, T. Ratnarajah, and J. Zhou, "MIMO radar and cellular coexistence: A power-efficient approach enabled by interference exploitation," vol. 66, no. 14, pp. 3681–3695, 2018.

[21] 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 J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, 2016.

[22] 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 Trans. Antennas Propagat.*, vol. 51, no. 4, pp. 782–787, 2003.

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

[24] B. Sadhu, Y. Tousi, J. Hallin, S. Sahl, S. K. Reynolds, . Renstrm, K. Sjgren, O. Haapalahti, N. Mazor, B. Bokinge, G. Weibull, H. Bengtsson, A. Carlinger, E. Westesson, J. Thillberg, L. Rexberg, M. Yeck, X. Gu, M. Ferriss, D. Liu, D. Friedman, and A. Valdes-Garcia, "A 28-ghz 32-element trx phased-array ic with concurrent dual-polarized operation and orthogonal phase and gain control for 5g communications," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 12, pp. 3373–3391, Dec 2017.

[25] A. Naqvi and S. Lim, "Review of recent phased arrays for millimeter-wave wireless communication," *Sensors*, vol. 18, no. 10, p. 3194, 2018.

[26] T. K. Lo, "Maximum ratio transmission," in *Commun., 1999. ICC'99. 1999 IEEE Int. Conf. on*, vol. 2. IEEE, 1999, pp. 1310–1314.

[27] C. K. Au-Yeung, D. J. Love, and S. Sanayei, "Trellis coded line packing: Large dimensional beamforming vector quantization and feedback transmission," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1844–1853, 2011.

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

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

[30] 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 Trans. Commun.*, vol. 57, no. 4, 2009.

[31] E. Höpfinger, "On the solution of the unidimensional local minimization problem," *J. Optimiz. Theory Appl.*, vol. 18, no. 3, pp. 425–428, 1976.

[32] G. Colavolpe and R. Raheli, "Noncoherent sequence detection," *IEEE Trans. Commun.*, vol. 47, no. 9, pp. 1376–1385, 1999.