

# Hierarchical Codebook Design and Analytical Beamforming Solution for IRS Assisted Communication

Xiyuan Liu, *Graduate Student Member, IEEE*, Qingqing Wu, *Senior Member, IEEE*, Die Hu, *Member, IEEE*, Rui Wang, *Senior Member, IEEE*, Jun Wu, *Senior Member, IEEE*,

## Abstract

In intelligent reflecting surface (IRS) assisted communication, beam search is usually time-consuming as the multiple-input multiple-output (MIMO) of IRS is usually very large. Hierarchical codebooks is a widely accepted method for reducing the complexity of searching time. The performance of this method strongly depends on the design scheme of beamforming of different beamwidths. In this paper, a non-constant phase difference (NCPD) beamforming algorithm is proposed. To implement the NCPD algorithm, we first model the phase shift of IRS as a continuous function, and then determine the parameters of the continuous function through the analysis of its array factor. Then, we propose a hierarchical codebook and two beam training schemes, namely the joint searching (JS) scheme and direction-wise searching (DWS) scheme by using the NCPD algorithm which can flexibly change the width, direction and shape of the beam formed by the IRS array. Simulation results show that the NCPD algorithm is more accurate with smaller side lobes, and also more stable on IRS of different sizes compared to other wide beam algorithms. The misalignment rate of the beam formed by the NCPD method is significantly reduced. The time complexity of the NCPD algorithm is constant, thus making it more suitable for solving the beamforming design problem with practically large IRS.

X. Liu is with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: 1910670@tongji.edu.cn). Q. Wu is with the Department of Electronic Engineering, Shanghai Jiao Tong University, 200240, China (e-mail: qingqingwu@sjtu.edu.cn). D. Hu is with the Key Laboratory of EMW Information, Fudan University, Shanghai 200433, China (e-mail: hudie@fudan.edu.cn). R. Wang is with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China, and also with the Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 201804, China (e-mail: ruiwang@tongji.edu.cn). J. Wu is with the School of Computer Science, Fudan University, Shanghai 200433, China, and also with the Shanghai Qi Zhi Institute, Shanghai 200030, China (e-mail: wujun@fudan.edu.cn).

## Index Terms

Intelligent reflecting surface, continuous array factor, hierarchical codebook, beamforming.

## I. INTRODUCTION

Intelligent reflecting surface (IRS) is a key technology often mentioned in 6th generation (6G) white papers. An IRS is generally composed of a large number of passive elements each being able to reflect the incident signal with an adjustable phase shift. Since IRS is passive, its cost is low and large-scale deployment can be achieved within a limited cost. IRS can be deployed on a flat surface such as the outer surface of a building to enhance the signal quality in a blind area of a base station [?]. Since IRS has no RF chain, the array scale on IRS is usually very large to provide sufficient beamforming gain [?]. The larger the array scale, the higher is the resolution of the array and the narrower is the formed beam [?]. The narrower the beam, the greater is the gain of the beam, and the better is the effect for communication. But the narrower the beam, the greater is the number of beams, which makes it to take longer time to search for a suitable beam [?]. This type of narrow beams are called as the pencil beam [?]. Therefore, how to speed up the beam search process is an important issue in the IRS-assisted communication scenarios.

Although the beam search is still immature in the IRS-assisted communication system, it is widely used in the conventional base station (BS) and user equipment (UE) scenarios. According to whether channel state information (CSI) information is utilized, beam search work can be divided into two categories: CSI based and CSI free. There are four main CSI based algorithms. The first one is the context-based beam search method [?], [?], [?], which exploits the user's location or movement related information to assist the beam search. This method first needs the user's location information, and then uses the user's position related information to assist the communication. But sometimes the location information is not accurate enough, which may cause the selected beam to be misaligned. The second method is based on machine learning [?], [?], [?] where the required beams can be obtained by learning the channel state information (CSI). However, a machine learning method requires a large amount of data to train the model. The third method uses the cascaded channel estimation results to optimize the IRS phase shift matrix. However since IRS can not supply a continuous phase shift, channel estimation processes in the IRS scenario are always not perfect [?]. Therefore, some algorithms are made available to study how to optimize the beamforming design of IRS under an inaccurate CSI [?], [?], [?], [?], [?].

The fourth method is an optimization algorithm based on the received signals [?], [?], [?], such as iterative optimization algorithm [?]. Because IRS is passive and it is difficult to obtain CSI, CSI based algorithms are not suitable for IRS assisted scenarios. There are three CSI-free methods which are exhaustive search, hierarchical search and heuristic algorithm, respectively. Exhaustive search method [?] first uses all the beams present in a sequence, then observes the power of the signal obtained by using each beam at the receiving end, and finally selects the corresponding beam with the highest receiving power. The disadvantage of exhaustive search method is that its search time is too long. Hierarchical method first searches with wide beams to determine the general direction, and then narrows down the search range with narrower beams, and finally identifies the most suitable pencil beam [?], [?]. The advantage of the hierarchical method is that it speeds up the beam search process, but the disadvantage is that the beamforming gain of the wide beams are small, and it is prone to misalignment. Heuristic algorithm [?], [?], [?], such as the genetic algorithm [?], [?], treats each beam as a gene, and then uses the principles of mutation, recombination and selection of genes to gradually find the optimal beam. This method also only suit for small arrays since there are too much pencil beams in large arrays scenario which makes it difficult for the algorithm to converge. In this paper, we use the hierarchical algorithm because of its low time complexity, and since we can stop beam searching at any time according to the required beamforming gain and directly use the intermediate results, it is more flexible. In order to implement the hierarchical search algorithm, we need to design a hierarchical codebook in IRS scenarios.

The main difficulty in the design of the IRS cascaded codebook is how to design the beams of different widths on IRS. Since the design principle of the hierarchical codebook on IRS is similar to that on the active array, we can learn from the wide beamforming methods on the active array. There are mainly three conventional wide beamforming methods. The first method is to turn off a portion of the elements. As the smaller the array, the wider is the beam formation [?]. This method has two major disadvantages. The first disadvantage is that the width of the beam cannot be controlled flexibly, and only a few specific widths can be used. Compared with a narrow beam, a wide beam loses not only the beamforming gain, but also some additional gain due to the reduction in the array size. The lose in such a gain can be effectively compensated in active arrays by power amplifier, but it cannot be effectively compensated in IRS. Therefore, this method is not easy to be applied in the IRS scenario. The second wide beamforming method is the beam combination [?]. In this method, the array is first divided into multiple areas, and then

each area forms a sub-beam, and these sub-beams can be used to combine the final desired wide beam. However, the number of divided areas needs to be manually selected according to the array size and targeted beams. So, this algorithm can be used only if the array size is known. When the array size changes, the beam combination algorithm needs to be redesigned. In the third method, nonlinear function is used to form IRS phase shift matrix, e.g., a quadratic function [?]. This method can form a more flexible wide beam. In order to address the above issues, in this study we propose to relax the IRS beamforming problem into a continuous function analysis problem, and then select the parameters of this continuous function to achieve a hierarchical codebook that can be used in arrays of different size. Another typical quadratic function used in wide beamforming is Zadoff-Chu sequence [?], [?]. Zadoff-Chu sequence is not suitable for small arrays, but it performs good on large arrays. However, Zadoff-Chu sequence can only form flat beams, it can not used to form more complicated beams.

The main contributions made in this paper are summarized as below:

- According to the characteristics of the cascaded channel via IRS, we propose a definition of the direction of the cascaded channel. Compared with the conventional form where the elevation and azimuth angles of incident and reflecting paths are used to describe the channel direction, our scheme uses fewer parameters to achieve the same purpose, in which only two parameters are used to replace the original four parameters. In addition, the values of the parameters of the cascaded channel direction defined in our scheme have a one-to-one mapping with the IRS phase shift configuration scheme. Therefore, the hierarchical search in the proposed cascaded channel direction reduces the time complexity.
- For the cascaded channel, we propose a transformed array factor defined in the direction of the cascaded channel mentioned above. Using the array factor of the cascaded channel, we can analyze the directions of both incident and reflected signals at IRS at the same time. In order to develop an IRS beamforming scheme applicable to arrays of any size, we transform the phase shift function of IRS into a continuous function. Then, the array factor is given after applying the continuous phase shift function. Using the beamforming scheme designed based on such an array factor, we need only to select different numbers of points on the continuous phase shift function to obtain the beamforming scheme on an array of the corresponding size. This avoids redesigning different phase shift configurations for arrays of different sizes.
- For designing the continuous phase shift function, we find the analytical solution of the

IRS beamforming problem by using the method of functional analysis. Using this analytical solution, we can find the IRS phase shift configuration for any target beam allowed by the array resolution, which reduces the time complexity of getting the phase shift function to  $\mathcal{O}(1)$ . For the wide beam required in the hierarchical codebook scenario, we propose a non-constant phase shift difference (NCPD) algorithm by applying the above analytical solution of IRS beamforming. Compared with the conventional algorithm that uses the beam combination method to realize the required wide beam, the NCPD algorithm is more flexible and the power distribution in the wide beam is more uniform. When the number of IRS array elements tends to be infinite, the array factor of the beam formed by the NCPD algorithm converges to the array factor of the target beam, which avoids the problem of deep depression of the beam in the beam combination methods. Using the wide beam formed by the NCPD algorithm, we implement a hierarchical codebook over one parameter in the direction of the cascaded channel.

- For the problem of having two parameters in the cascaded channel direction, we propose two beam training schemes, namely the joint searching (JS) scheme and direction-wise searching (DWS) scheme. The JS scheme searches for two parameters at the same time while the DWS scheme searches for one parameter first, and then for the other parameter. Applying the above two beam training schemes, we realize the hierarchical search of the cascaded channel direction on IRS.
- Simulation results verify our theoretical findings that the NCPD algorithm is significantly improved compared to the beam combination algorithm on arrays of different sizes and when forming beams of different widths. When the number of array elements increases, we can observe that the power distribution in the middle of the beam is very flat, and the Gibbs phenomenon appears at the edge of the beam, which shows that the accuracy of the NCPD algorithm is far superior to other similar algorithms.

The remainder of this paper is organized as follows. Section II introduces the system and channel models, followed by the proposal of array factor of IRS in Section III. Section IV explains the principle of IRS phase shift configuration. The design of the hierarchical codebook on IRS is proposed in Section V, while in Section VI we show the performance and time complexity of proposed NCPD algorithm. Simulation results are presented in Section VII. Finally, we conclude the paper in Section VIII.

*Notations:* In this paper, scalars, vectors and matrices are denoted by italic letters, bold-face

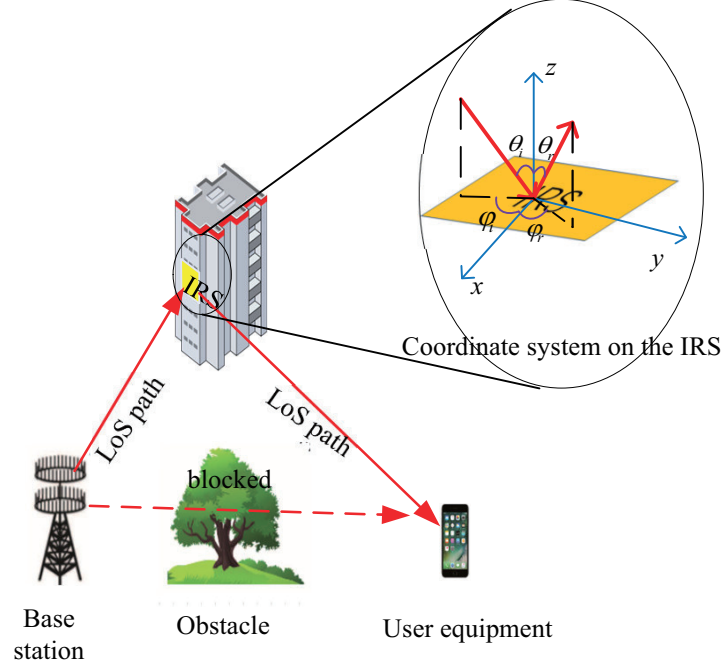


Fig. 1. An IRS assisted communication system.

lower-case and upper-case letters, respectively. The space of  $x \times y$  complex-valued matrices is denoted by  $\mathbb{C}^{x \times y}$ . For a complex-value vector  $\mathbf{x}$ ,  $\mathbf{x} \otimes \mathbf{y}$  denotes the kronecker product of  $\mathbf{x}$  and  $\mathbf{y}$  while  $|\mathbf{x}|$  denotes its modulus and  $\text{diag}(\mathbf{x})$  denotes a diagonal matrix with each diagonal entry being the corresponding entry in  $\mathbf{x}$ . For a function  $\mathbf{y} = H(\mathbf{x})$ ,  $H^{-1}(\mathbf{y})$  denotes its inverse function. For a general matrix  $\mathbf{A}$ ,  $\mathbf{A}^*$ ,  $\mathbf{A}^H$ , and  $\mathbf{A}[i, j]$  denote its conjugate, conjugate transpose, and the  $(i, j)$ th entry, respectively.  $j$  denotes the imaginary unit, i.e.,  $j^2 = -1$ .

## II. SYSTEM AND CHANNEL MODELS

For the purpose of exposition, we consider a basic IRS-aided single-user communication system in millimeter-wave (mmWave) band as illustrated in Fig. ??, where an IRS composed of a large number of  $N = N_{\text{ver}} \times N_{\text{hor}}$  passive reflecting elements, is deployed in proximity to a user for assisting its data transmission with an BS, both of which are equipped with a single antenna. The results in this paper can be readily extended to the more general system with multiple users served by the IRS (e.g., by applying orthogonal time/frequency division multiple access) and/or multiple antennas at the BS (by estimating their associated channels in parallel). In this paper, it assumed that the line of sight (LoS) path of BS-UE channel is blocked while

those of BS-IRS and IRS-UE channel are not blocked as shown in Fig. ?? . Therefore, the IRS establishes a virtual LoS path of BS-IRS-UE. The received signal of the UE is:

$$y = \mathbf{h}_r^H \Phi \mathbf{h}_i x + z, \quad (1)$$

where  $x$ ,  $y$  and  $z \sim N(0, \sigma^2)$  are transmitted signal of BS, received signal of UE and channel noise with zero mean and variance  $\sigma^2$ , respectively.  $\mathbf{h}_r$  and  $\mathbf{h}_i$  are channel of IRS-UE and BS-IRS respectively.  $\Phi$  is a diagonal matrix with elements' response value on its diagonal.  $\Phi$  can be expressed as:

$$\Phi = \text{diag}(e^{jkdg(1)}, e^{jkdg(2)}, \dots, e^{jkdg(n)}, \dots, e^{jkdg(N)}), \quad (2)$$

where  $n$ ,  $k$  and  $d$  are the index of IRS elements, wavenumber and spacing distance of IRS respectively.  $k = \frac{2\pi}{\lambda}$  where  $\lambda$  is the wave length.  $kdg(n)$  is the phase shift configuration of IRS elements.

*Proposition 1:* The maximum spacing between the elements of IRS is  $\frac{\lambda}{4}$ .

*Proof:* See Appendix ??.

In this paper, we choose  $d = \frac{\lambda}{4}$ . Since mmWave channels are sparse, most of the signal power exists in the LoS paths. So, we ignore the NLoS path and consider only the LoS path. Then  $\mathbf{h}_r$  and  $\mathbf{h}_i$  can be expressed as:

$$\mathbf{h}_r = \varrho_r \mathbf{d}, \quad (3)$$

$$\mathbf{h}_i = \varrho_i \mathbf{a}, \quad (4)$$

where  $\varrho_r$  and  $\varrho_i$  are path loss of the IRS-UE and BS-IRS channels, respectively.  $\mathbf{d}$  and  $\mathbf{a}$  are the steering vectors corresponding to the angles-of-departure (AoDs) and angles-of-arrival (AoAs) of IRS, respectively. Therefore, (??) can be rewritten as:

$$y = \varrho_r \varrho_i \mathbf{d}^H \Phi \mathbf{a} x. \quad (5)$$

Then the change in variables  $\mathbf{d} \Phi \mathbf{a} = \boldsymbol{\nu}^H \mathbf{h}_{\text{vtl}}$  is applied, where  $\boldsymbol{\nu} = [e^{jkdg(1)}, e^{jkdg(2)}, \dots, e^{jkdg(N)}]^T$ .  $\mathbf{h}_{\text{vtl}} = \text{diag}(\mathbf{d}) \mathbf{a}$  is an equivalent virtual LoS path of BS-UE. Since IRS is an uniform planar array (UPA),  $\mathbf{d}$  and  $\mathbf{a}$  can be disassembled into the following form:

$$\mathbf{d} = \mathbf{d}_{\text{hor}} \otimes \mathbf{d}_{\text{ver}}, \quad (6)$$

$$\mathbf{a} = \mathbf{a}_{\text{hor}} \otimes \mathbf{a}_{\text{ver}}, \quad (7)$$

where  $\mathbf{d}_{\text{hor}}$ ,  $\mathbf{a}_{\text{hor}}$ ,  $\mathbf{d}_{\text{ver}}$  and  $\mathbf{a}_{\text{ver}}$  are steering vectors of horizontal and vertical direction of IRS array, which can be expressed as:

$$\mathbf{d}_{\text{hor}} = [1, e^{jkd \sin \theta_r \cos \varphi_r}, e^{jkd2 \sin \theta_r \cos \varphi_r}, \dots, e^{jkd(N_{\text{hor}}-1) \sin \theta_r \cos \varphi_r}]^T, \quad (8)$$

$$\mathbf{d}_{\text{ver}} = [1, e^{jkd \sin \theta_r \sin \varphi_r}, e^{jkd2 \sin \theta_r \sin \varphi_r}, \dots, e^{jkd(N_{\text{ver}}-1) \sin \theta_r \sin \varphi_r}]^T, \quad (9)$$

$$\mathbf{a}_{\text{hor}} = [1, e^{jkd \sin \theta_i \cos \varphi_i}, e^{jkd2 \sin \theta_i \cos \varphi_i}, \dots, e^{jkd(N_{\text{hor}}-1) \sin \theta_i \cos \varphi_i}]^T, \quad (10)$$

$$\mathbf{a}_{\text{ver}} = [1, e^{jkd \sin \theta_i \sin \varphi_i}, e^{jkd2 \sin \theta_i \sin \varphi_i}, \dots, e^{jkd(N_{\text{ver}}-1) \sin \theta_i \sin \varphi_i}]^T, \quad (11)$$

where  $\theta_i$ ,  $\theta_r$ ,  $\phi_i$  and  $\phi_r$  are elevation and azimuth angle of corresponding AoA and AoD of the IRS. Hence

$$\mathbf{h}_{\text{vtl}} = (\text{diag}(\mathbf{d}_{\text{hor}})\mathbf{a}_{\text{hor}}) \otimes (\text{diag}(\mathbf{d}_{\text{ver}})\mathbf{a}_{\text{ver}}). \quad (12)$$

We define

$$\mathbf{h}_{\text{hor}} = \text{diag}(\mathbf{d}_{\text{hor}})\mathbf{a}_{\text{hor}}, \quad (13)$$

$$\mathbf{h}_{\text{ver}} = \text{diag}(\mathbf{d}_{\text{ver}})\mathbf{a}_{\text{ver}}. \quad (14)$$

Thus  $\mathbf{h}_{\text{hor}}$  and  $\mathbf{h}_{\text{ver}}$  can be expressed as:

$$\mathbf{h}_{\text{hor}} = [1, e^{jkd\beta_{\text{hor}}}, e^{jkd2\beta_{\text{hor}}}, \dots, e^{jkd(N_{\text{hor}}-1)\beta_{\text{hor}}}]^T, \quad (15)$$

$$\mathbf{h}_{\text{ver}} = [1, e^{jkd\beta_{\text{ver}}}, e^{jkd2\beta_{\text{ver}}}, \dots, e^{jkd(N_{\text{ver}}-1)\beta_{\text{ver}}}]^T, \quad (16)$$

where  $\beta_{\text{hor}}$  and  $\beta_{\text{ver}}$  can be written as [?]:

$$\beta_{\text{hor}} = \sin \theta_r \cos \varphi_r + \sin \theta_i \cos \varphi_i, \quad (17)$$

$$\beta_{\text{ver}} = \sin \theta_r \sin \varphi_r + \sin \theta_i \sin \varphi_i. \quad (18)$$

From above we can see that the direction of  $\mathbf{h}_{\text{vtl}}$  can be determined by direction of  $\mathbf{h}_{\text{hor}}$  and  $\mathbf{h}_{\text{ver}}$  which are controlled by  $\beta_{\text{hor}}$  and  $\beta_{\text{ver}}$ . Simultaneously,  $\beta_{\text{hor}}$  and  $\beta_{\text{ver}}$  can be determined by  $\theta_r$ ,  $\theta_i$ ,  $\phi_r$  and  $\phi_i$ . Thus we define  $\beta_{\text{hor}}$  and  $\beta_{\text{ver}}$  are cascaded channel angles of the direction  $\mathbf{h}_{\text{hor}}$  and  $\mathbf{h}_{\text{ver}}$ , respectively. The following sections will use  $\beta_{\text{hor}}$  and  $\beta_{\text{ver}}$  to describe cascaded channel direction rather than  $\theta_r$ ,  $\theta_i$ ,  $\phi_r$  and  $\phi_i$ .



### III. IRS ARRAY FACTOR

In order to realize the hierarchical codebook on IRS, we need to design an algorithm that can flexibly control the orientation, width and shape of beam on the IRS. In this section we will introduce array factor of IRS and then analysis beamforming goal of hierarchical codebook on the proposed continuous array factor.

According to the array factor of a conventional array [?], the array factor modulus (AFM) on the IRS can be expressed as:

$$\begin{aligned}
 AFM(\beta_{\text{hor}}, \beta_{\text{ver}}) &= |\mathbf{d}^H \Phi \mathbf{a}| \\
 &= |\boldsymbol{\nu}^H \mathbf{h}_{\text{vtl}}| \\
 &= \left| \sum_{n_{\text{hor}}=0}^{N_{\text{hor}}-1} \sum_{n_{\text{ver}}=0}^{N_{\text{ver}}-1} e^{jkd[\beta_{\text{hor}}n_{\text{hor}}+\beta_{\text{ver}}n_{\text{ver}}+g[(n_{\text{hor}}-1)N_{\text{hor}}+n_{\text{ver}}]]} \right|.
 \end{aligned} \tag{19}$$

Since  $\beta_{\text{hor}}n_{\text{hor}}$  and  $\beta_{\text{ver}}n_{\text{ver}}$  are independent to each other, so we can restrict  $g[(n_{\text{hor}}-1)N_{\text{hor}}+n_{\text{ver}}]$  to the form in which  $g[(n_{\text{hor}}-1)N_{\text{hor}}+n_{\text{ver}}]$  is the sum of two independent parts as below:

$$g[(n_{\text{hor}}-1)N_{\text{hor}}+n_{\text{ver}}] = g_{\text{hor}}(n_{\text{hor}}) + g_{\text{ver}}(n_{\text{ver}}), \tag{20}$$

where  $n_{\text{hor}}$  and  $n_{\text{ver}}$  are elements' index along horizontal and vertical on IRS array and  $g_{\text{hor}}(n_{\text{hor}})$  and  $g_{\text{ver}}(n_{\text{ver}})$  are the IRS phase shift functions along the horizontal and vertical directions on IRS array, respectively. Thus (??) can be rewritten as:

$$\begin{aligned}
 AFM(\beta_{\text{hor}}, \beta_{\text{ver}}) &= \left| \sum_{n_{\text{hor}}=0}^{N_{\text{hor}}-1} \sum_{n_{\text{ver}}=0}^{N_{\text{ver}}-1} e^{jkd[\beta_{\text{hor}}n_{\text{hor}}+\beta_{\text{ver}}n_{\text{ver}}+g_{\text{hor}}(n_{\text{hor}})+g_{\text{ver}}(n_{\text{ver}})]} \right| \\
 &= \left| \sum_{n_{\text{hor}}=0}^{N_{\text{hor}}-1} e^{jkd[\beta_{\text{hor}}n_{\text{hor}}+g_{\text{hor}}(n_{\text{hor}})]} \sum_{n_{\text{ver}}=0}^{N_{\text{ver}}-1} e^{jkd[\beta_{\text{ver}}n_{\text{ver}}+g_{\text{ver}}(n_{\text{ver}})]} \right| \\
 &= \left| \sum_{n_{\text{hor}}=0}^{N_{\text{hor}}-1} e^{jkd[\beta_{\text{hor}}n_{\text{hor}}+g_{\text{hor}}(n_{\text{hor}})]} \right| \cdot \left| \sum_{n_{\text{ver}}=0}^{N_{\text{ver}}-1} e^{jkd[\beta_{\text{ver}}n_{\text{ver}}+g_{\text{ver}}(n_{\text{ver}})]} \right|.
 \end{aligned} \tag{21}$$

We can see in (??) that the IRS reflected power is the product of the reflected powers in the  $\mathbf{h}_{\text{hor}}$  and  $\mathbf{h}_{\text{ver}}$  directions. The reflected powers of IRS along the  $\mathbf{h}_{\text{hor}}$  and  $\mathbf{h}_{\text{ver}}$  directions are two independent parts, and they have duality. So we take only the  $\mathbf{h}_{\text{hor}}$  direction as an example to explain how to determine the optimal  $g_{\text{hor}}(n_{\text{hor}})$ , and then the  $\mathbf{h}_{\text{ver}}$  direction can be obtained in the same way. The array factor on  $\mathbf{h}_{\text{hor}}$  direction can be expressed as:

$$AFM(\beta_{\text{hor}}) = \left| \sum_{n_{\text{hor}}=0}^{N_{\text{hor}}-1} e^{jkd[\beta_{\text{hor}}n_{\text{hor}}+g_{\text{hor}}(n_{\text{hor}})]} \right|. \tag{22}$$

We normalize (??) to get the following:

$$AFM_{\text{norm}}(\beta_{\text{hor}}) = \frac{1}{N_{\text{hor}}} \left| \sum_{n_{\text{hor}}=0}^{N_{\text{hor}}-1} e^{jk d [\beta_{\text{hor}} n_{\text{hor}} + g_{\text{hor}}(n_{\text{hor}})]} \right|. \quad (23)$$

*Proposition 2:* As  $d \rightarrow 0$ , we have:

$$AFM_{\text{norm}}(\beta_{\text{hor}}) = \left| \int_0^1 e^{jk N_{\text{hor}} d [\beta_{\text{hor}} \ddot{u} - \int_0^{\ddot{u}} \dot{f}(\mu) d\mu]} d\ddot{u} \right|. \quad (24)$$

*Proof:* See Appendix ??.

Next, we first get the corresponding  $\dot{f}(\mu)$  according to the goal of beamforming, and then get the final  $g_{\text{hor}}(n_{\text{hor}})$ .

The direction of the narrow beam is a certain cascaded channel direction. If the targeted cascaded channel direction is  $\psi_{\text{hor},0}$ , the goal of the narrow beam can be expressed as:

$$AFM_{\text{norm}}(\beta_{\text{hor}}) = \begin{cases} 1, & \text{if } \beta_{\text{hor}} = \psi_{\text{hor},0}, \\ 0, & \text{otherwise.} \end{cases} \quad (25)$$

The maximum of AFM is 1 for the reason that:

$$\left| \int_0^1 e^{jk N_{\text{hor}} d [\beta_{\text{hor}} \ddot{u} - \int_0^{\ddot{u}} \dot{f}(\mu) d\mu]} d\ddot{u} \right| \leq \int_0^1 \left| e^{jk N_{\text{hor}} d [\beta_{\text{hor}} \ddot{u} - \int_0^{\ddot{u}} \dot{f}(\mu) d\mu]} \right| d\ddot{u} \leq 1. \quad (26)$$

When  $\dot{f}(\mu) = \psi_{\text{hor},0}$ ,  $AFM_{\text{norm}}(\psi_{\text{hor},0}) = 1$ . So one of the solution of (??) is  $\dot{f}(\mu) = \psi_{\text{hor},0}$ .

The wide beam refers to the beam oriented towards a range of the cascaded channel directions. Suppose that cascaded channel direction range of the target wide beam is  $(\psi_{\text{hor},a}, \psi_{\text{hor},b})$ , and the amplitude distribution of the wide beam is  $h(\beta_{\text{hor}})$ . Then, the wide beam array factor goal can be expressed as:

$$AFM_{\text{norm}}(\beta_{\text{hor}}) = \begin{cases} h(\beta_{\text{hor}}), & \beta_{\text{hor}} \in [\psi_{\text{hor},a}, \psi_{\text{hor},b}], \\ 0, & \beta_{\text{hor}} \notin [\psi_{\text{hor},a}, \psi_{\text{hor},b}]. \end{cases} \quad (27)$$

The following sections will elaborate how to design  $\dot{f}(\mu)$  to satisfy this wide beam requirement and how to use wide beams to form hierarchical codebook on IRS.

#### IV. CONFIGURATION OF IRS PHASE SHIFT FOR BEAMFORMING OF WIDE BEAM

In this section, we will discuss how to design  $\dot{f}(\mu)$  by satisfying the goal of wide beam as in (??). In order to simplify the problem, we specify that  $\dot{f}(\mu)$  is an increasing differentiable function in the range of  $[\dot{f}(0), \dot{f}(1)]$ . We take a infinitesimal interval  $(\mu_0, \mu_0 + \Delta\mu_0)$  and replace  $\dot{f}(\mu)$  with its first-order Taylor approximation on this element:  $\dot{f}(\mu) = \dot{f}(\mu_0) + \dot{f}'(\mu_0)(\mu - \mu_0)$ .

In the infinitesimal interval  $(\mu_0, \mu_0 + \Delta\mu_0)$  at  $\mu_0$ , (??) can be written as:

$$\begin{aligned}
& AFM_{\text{norm}}(\beta_{\text{hor}})|_{\ddot{u} \in (\mu_0, \mu_0 + \Delta\mu_0)} \\
&= \left| \int_{\mu_0}^{\mu_0 + \Delta\mu_0} e^{\mathcal{J}kN_{\text{hor}}d[\beta_{\text{hor}}\ddot{u} - \int_0^{\ddot{u}} \dot{f}(\mu_0) + \dot{f}'(\mu_0)(\mu - \mu_0)d\mu]} d\ddot{u} \right| \\
&= \left| \int_{\mu_0}^{\mu_0 + \Delta\mu_0} e^{\mathcal{J}kN_{\text{hor}}d[\beta_{\text{hor}}\mu_0 - \int_0^{\mu_0} \dot{f}(\mu)d\mu]} \cdot e^{\mathcal{J}kN_{\text{hor}}d[\beta_{\text{hor}}(\ddot{u} - \mu_0) - \dot{f}(\mu_0)(\ddot{u} - \mu_0) - \frac{1}{2}\dot{f}'(\mu_0)(\ddot{u} - \mu_0)^2]} d\ddot{u} \right| \\
&= \left| e^{\mathcal{J}kN_{\text{hor}}d[\beta_{\text{hor}}\mu_0 - \int_0^{\mu_0} \dot{f}(\mu)d\mu]} \cdot \int_{\mu_0}^{\mu_0 + \Delta\mu_0} e^{\mathcal{J}kN_{\text{hor}}d[\beta_{\text{hor}}(\ddot{u} - \mu_0) - \dot{f}(\mu_0)(\ddot{u} - \mu_0) - \frac{1}{2}\dot{f}'(\mu_0)(\ddot{u} - \mu_0)^2]} d\ddot{u} \right| \quad (28) \\
&= \left| e^{\mathcal{J}kN_{\text{hor}}d[\beta_{\text{hor}}\mu_0 - \int_0^{\mu_0} \dot{f}(\mu)d\mu]} \right| \cdot \left| \int_0^{\Delta\mu_0} e^{\mathcal{J}kN_{\text{hor}}d[\beta_{\text{hor}}\ddot{u} - \dot{f}(\mu_0)\ddot{u} - \frac{1}{2}\dot{f}'(\mu_0)\ddot{u}^2]} d\ddot{u} \right| \\
&= \left| \int_0^{\Delta\mu_0} e^{\mathcal{J}kN_{\text{hor}}d[\beta_{\text{hor}}\ddot{u} - \dot{f}(\mu_0)\ddot{u} - \frac{1}{2}\dot{f}'(\mu_0)\ddot{u}^2]} d\ddot{u} \right|.
\end{aligned}$$

Since the array resolution increases with an increase in  $kN_{\text{hor}}d$ , we should design  $\dot{f}(\mu)$  on the condition that  $kN_{\text{hor}}d \rightarrow \infty$ . If  $\beta_{\text{hor}} \neq \dot{f}(\mu_0)$ , the following holds as the  $kN_{\text{hor}}d$  tends to infinity:

$$\begin{aligned}
& \lim_{kN_{\text{hor}}d \rightarrow \infty} \left| \int_0^{\Delta\mu_0} e^{\mathcal{J}kN_{\text{hor}}d[\beta_{\text{hor}}\ddot{u} - \dot{f}(\mu_0)\ddot{u} - \frac{1}{2}\dot{f}'(\mu_0)\ddot{u}^2]} d\ddot{u} \right| \\
& \stackrel{\textcircled{1}}{=} \lim_{kN_{\text{hor}}d \rightarrow \infty} \frac{1}{kN_{\text{hor}}d} \left| \int_0^{kN_{\text{hor}}d\Delta\mu_0} e^{\mathcal{J}[\beta_{\text{hor}}t - \dot{f}(\mu_0)t - \frac{1}{2}\dot{f}'(\mu_0)\frac{t^2}{kN_{\text{hor}}d}]} dt \right| \\
&= \lim_{kN_{\text{hor}}d \rightarrow \infty} \frac{1}{kN_{\text{hor}}d} \left| \int_0^{kN_{\text{hor}}d\Delta\mu_0} e^{\mathcal{J}[\beta_{\text{hor}}t - \dot{f}(\mu_0)t]} dt \right| \quad (29) \\
&\leq \lim_{kN_{\text{hor}}d \rightarrow \infty} \frac{2}{kN_{\text{hor}}d} \\
&= 0,
\end{aligned}$$

where  $\textcircled{1}$  is  $t = kN_{\text{hor}}d\ddot{u}$ . So  $\lim_{kN_{\text{hor}}d \rightarrow \infty} AFM_{\text{norm}}(\beta_{\text{hor}})|_{\ddot{u} \in (\mu_0, \mu_0 + \Delta\mu_0)} = 0$  when  $\beta_{\text{hor}} \neq \dot{f}(\mu_0)$ .

If  $\beta_{\text{hor}} = \dot{f}(\mu_0)$ , then the following holds:

$$\begin{aligned}
& \lim_{kN_{\text{hor}}d \rightarrow \infty} AFM_{\text{norm}}(\beta_{\text{hor}})|_{\ddot{u} \in (\mu_0, \mu_0 + \Delta\mu_0)} \\
&= \lim_{kN_{\text{hor}}d \rightarrow \infty} \left| \int_0^{\Delta\mu_0} e^{\mathcal{J}kN_{\text{hor}}d[-\frac{1}{2}\dot{f}'(\mu_0)\ddot{u}^2]} d\ddot{u} \right| \\
& \stackrel{\textcircled{2}}{=} \lim_{kN_{\text{hor}}d \rightarrow \infty} \frac{1}{\sqrt{\dot{f}'(\mu_0)}} \left| \int_0^{\sqrt{\dot{f}'(\mu_0)\Delta\mu_0}} e^{\mathcal{J}kN_{\text{hor}}dt^2} dt \right| \quad (30) \\
&= \frac{K}{\sqrt{\dot{f}'(\mu_0)}}
\end{aligned}$$

where  $\textcircled{2}$  is  $t = \sqrt{\dot{f}'(\mu_0)}\ddot{u}$ , and  $K$  is a variable independent to  $\dot{f}(\mu)$ .

Since  $\dot{f}(\mu)$  is an increasing differentiable function in the range of  $[\dot{f}(0), \dot{f}(1)]$ , we can find only one  $\mu_0$  satisfies  $\dot{f}(\mu_0) = \beta_{\text{hor}}$  when  $\beta_{\text{hor}} \in [\dot{f}(0), \dot{f}(1)]$ . Thus we can get:

$$\lim_{kN_{\text{hor}}d \rightarrow \infty} AFM_{\text{norm}}(\beta_{\text{hor}}) = \begin{cases} \frac{K}{\sqrt{\dot{f}'(\mu_0)}}, & \beta_{\text{hor}} \in [\dot{f}(0), \dot{f}(1)], \mu_0 = \dot{f}^{-1}(\beta_{\text{hor}}), \\ 0, & \beta_{\text{hor}} \notin [\dot{f}(0), \dot{f}(1)]. \end{cases} \quad (31)$$

Comparing (??) and (??), in the wide beamforming problem we can get  $\dot{f}(0) = \psi_{\text{hor},a}$   $\dot{f}(1) = \psi_{\text{hor},b}$  and:

$$K \sqrt{\left. \frac{d\mu}{d\dot{f}(\mu)} \right|_{\mu=\mu_0}} = h(\beta_{\text{hor}}), \quad (32)$$

where  $K$  is a coefficient and  $\beta_{\text{hor}} = \dot{f}(\mu_0)$ . We do not need to normalize the total power when selecting  $h(\beta_{\text{hor}})$ , as the selection of  $K$  can automatically achieve normalization. So the value of  $h(\beta_{\text{hor}})$  does not actually represent the true value of  $AFM_{\text{norm}}(\beta_{\text{hor}})$ , but rather represents the ratio relationship between the values of  $AFM_{\text{norm}}(\beta_{\text{hor}})$  at different  $\beta_{\text{hor}}$ . Hence, the following differential equation holds:

$$K\mu = \int h(\dot{f}(\mu)) d\dot{f}(\mu). \quad (33)$$

Suppose that the solution of (??) is:

$$K\mu = H(\dot{f}(\mu)) + C. \quad (34)$$

Putting  $\dot{f}(0) = \psi_{\text{hor},a}$  and  $\dot{f}(1) = \psi_{\text{hor},b}$  into (??), we can get:

$$\mu = \frac{1 - 0}{H(\psi_{\text{hor},b}) - H(\psi_{\text{hor},a})} (H(\dot{f}(\mu)) - H(\psi_{\text{hor},a})) + 0. \quad (35)$$

Hence,  $\dot{f}(\mu)$  becomes:

$$\dot{f}(\mu) = H^{-1}((H(\psi_{\text{hor},b}) - H(\psi_{\text{hor},a}))\mu + H(\psi_{\text{hor},a})). \quad (36)$$

When  $N_{\text{hor}}$  tends to be infinity,  $\dot{f}(\mu)$  can be considered to have a constant value  $\dot{f}(\frac{n_{\text{hor}}}{N_{\text{hor}}})$  in  $(n_{\text{hor}} - 1, n_{\text{hor}})$ . Accordingly, (??) can be simplified as:

$$g_{\text{hor}}(n_{\text{hor}}) = - \sum_{\tau=1}^{n_{\text{hor}}} \dot{f}(\frac{\tau}{N_{\text{hor}}}). \quad (37)$$

Intuitively, for the hierarchical codebook, the best wide beam power distribution is the power evenly distributed in the beam range. Hence,  $h(\beta_{\text{hor}})$  is a constant function, and  $\dot{f}(\mu)$  can be simplified as:

$$\dot{f}(\mu) = (\psi_{\text{hor},b} - \psi_{\text{hor},a})\mu + \psi_{\text{hor},a}. \quad (38)$$

Since (??) is a linear function, we call this algorithm as the none-constant phase shift difference (NCPD) algorithm. Therefore  $g_{\text{hor}}(n_{\text{hor}})$  can be expressed as:

$$g_{\text{hor}}(n_{\text{hor}}) = -\left(\sum_{\tau=1}^{n_{\text{hor}}} ((\psi_{\text{hor},b} - \psi_{\text{hor},a}) \frac{\tau}{N_{\text{hor}}} + \psi_{\text{hor},a})\right) = -\left(\frac{n_{\text{hor}}(n_{\text{hor}} + 1)}{2N_{\text{hor}}}(\psi_{\text{hor},b} - \psi_{\text{hor},a}) + n_{\text{hor}}\psi_{\text{hor},a}\right). \quad (39)$$

**Remark 1.** If  $\dot{f}(\mu)$  is not monotonic, then it can be viewed as a piecewise function, where  $\dot{f}(\mu)$  is monotonic in each segment, and the conclusions described in this paper can still be applied. Therefore, in order to simplify the expression, this paper discusses only the case where  $\dot{f}(\mu)$  is a monotonic function.

**Remark 2.** The above algorithm is developed based on the constant modulus constraint. So, the conclusions work not only on IRS but also on analog beamforming of active equipments.

## V. DESIGN OF HIERARCHICAL CODEBOOK ON IRS

Since the cascaded channel direction is determined by using  $\beta_{\text{hor}}$  and  $\beta_{\text{ver}}$ , and the hierarchical codebook on them are the same, in this section we first introduce the hierarchical codebook in a single direction, and then propose two beam training algorithms.

### A. Hierarchical Codebook on One Direction

Suppose that the hierarchical codebook has  $S$  layers, and  $s$  is its layer index, where  $s = 1, 2, \dots, S$ . We use  $\Xi_s$  to denote the codebook of layer  $s$ .  $\Xi_S$  is the bottom layer with narrow beams, other layers use wide beams.  $\Xi_1$  has two beams and  $\Xi_2$  has four beams, and  $\Xi_s$  has  $2^s$  beams. From [?], we know that there are  $2N_{\text{hor}}$  narrow beams in the  $\mathbf{h}_{\text{hor}}$  direction, and  $2^S = 2N_{\text{hor}}$ . We can find that the hierarchical codebook needs  $\lceil \log_2 N_{\text{hor}} \rceil$  layers. The  $s$ th layer of the hierarchical codebook can be expressed as:

$$\Xi_s = [\xi_1^s, \xi_2^s, \dots, \xi_{2^s}^s], \quad (40)$$

where  $\xi_i^s$  is a  $N_{\text{hor}}$ -dimensional column vector. For the bottom layer, the beams can be expressed as:

$$\xi_i^S[n_{\text{hor}}] = e^{jkd(n_{\text{hor}}-1)\frac{2(i-N_{\text{hor}})-1}{N_{\text{hor}}}}. \quad (41)$$

For any other layer  $s$  ( $s \neq S$ ), we first need to confirm the range of the directions of the beams. Suppose that  $(\psi_{a,\xi_i^s}, \psi_{b,\xi_i^s})$  is the directional range of beam  $\xi_i^s$

$$\psi_{a,\xi_i^s} = -2 + \frac{4}{2^s}(i-1), \quad \psi_{b,\xi_i^s} = -2 + \frac{4}{2^s}i. \quad (42)$$

Putting them into (??), we can get the beam  $\xi_i^s$  ( $s \neq S$ ) as:

$$\xi_i^s[n_{\text{hor}}] = \exp\left\{-jkd\left[\frac{n_{\text{hor}}(n_{\text{hor}} + 1)}{2N_{\text{hor}}} \cdot \frac{4}{2^s} - 2 + \frac{4}{2^s}(i - 1)\right]\right\}. \quad (43)$$

$$s = 1, 2, \dots, S - 1, \quad i = 1, 2, \dots, 2^s.$$

For the  $\mathbf{h}_{\text{ver}}$  direction, the hierarchical codebook are exactly the same as that in the  $\mathbf{h}_{\text{hor}}$  direction. Fig. ?? shows the structure of the codebook.

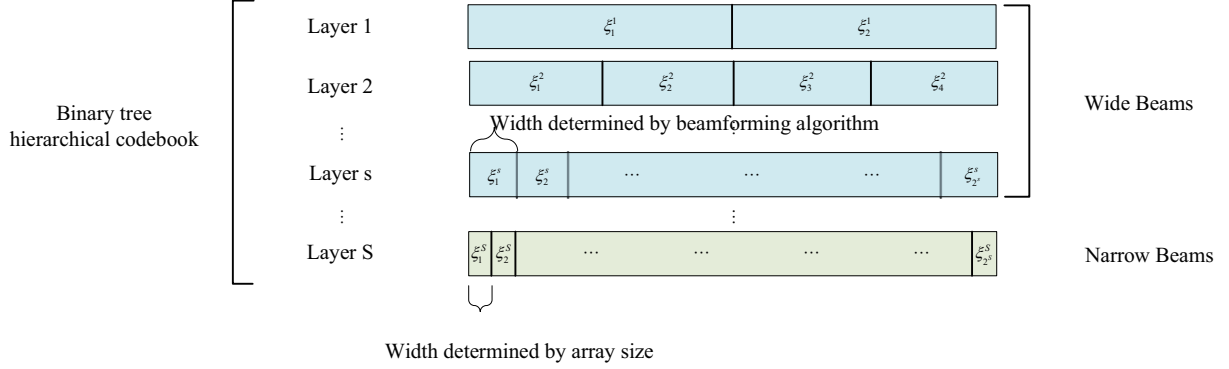


Fig. 2. Hierarchical codebook structure.

### B. Beam Training Scheme for Jointment Searching

In this sub-section, we will propose a searching scheme to search  $\mathbf{h}_{\text{hor}}$  direction and  $\mathbf{h}_{\text{ver}}$  direction at the same time which is called jointment searching (JS) scheme. The phase shift matrix of the IRS is searched by using the above-described hierarchical codebook. According to (??), the beam direction of the IRS can be divided into the  $\mathbf{h}_{\text{hor}}$  direction and the  $\mathbf{h}_{\text{ver}}$  direction, and the JS search scheme will search in these two directions synchronously. In the  $s$ th layer search, codewords in  $\Xi^1$  are used in the  $\mathbf{h}_{\text{hor}}$  and  $\mathbf{h}_{\text{ver}}$  directions. According to (??), the phase shift matrix of IRS in the  $s$ th layer search can be written as:

$$\Phi_{p,q} = \text{diag}(\xi_p^s \otimes \xi_q^s). \quad (44)$$

Then, according to (??), the received signal is obtained, and the received signal with the largest power is selected, that is, the indices of the best code of  $s$ th layer is selected. According to (??) we know:

$$\{p^*, q^*\} = \arg \max_{p,q} \mathbf{d}^H \Phi_{p,q} \mathbf{a}. \quad (45)$$

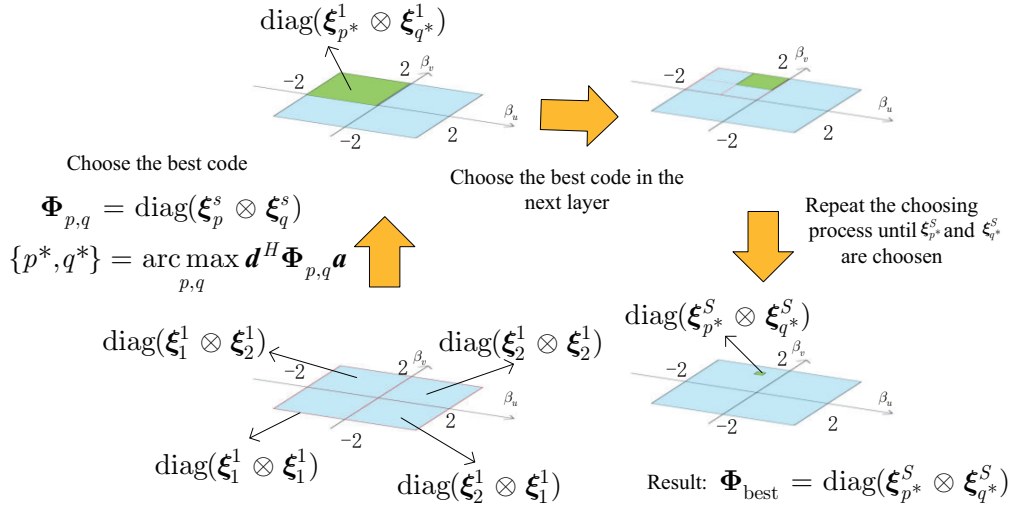


Fig. 3. JS beam training process.

At this point, the searching procedure of the  $s$ th layer is finished, and then we can search the codes in the next layer. Finally we can get the best phase shift matrix of IRS after the bottom layer of the hierarchical codebook are searched. The details of the proposed algorithm are summarized in Algorithm ??.

---

**Algorithm 1** JS codebook design

---

- 1: Initialize the layer index  $s = 1$  and  $p^* = 1$ .
  - 2: **repeat**
  - 3:   Compute  $\Phi_{p,q}$  by (??), where  $p \in \{2p^* - 1, 2p^*\}$ ,  $q \in \{2q^* - 1, 2q^*\}$ .
  - 4:   Update  $p^*$  by choosing the indices of the best codes of the  $s$ th layer by (??).
  - 5:   Update  $s = s + 1$ .
  - 6: **until**  $s > S$ .
  - 7:  $\Phi_{\text{best}} = \text{diag}(\xi_{p^*}^S \otimes \xi_{q^*}^S)$ .
- 

### C. Direction Wise Searching Scheme

In this section, we propose another beam searching scheme on IRS named direction wise searching (DWS) scheme. The difference between the DWS and the JS scheme is that the DWS

scheme first searches the  $\mathbf{h}_{\text{hor}}$  direction, and then starts searching in the  $\mathbf{h}_{\text{ver}}$  direction after determining the codeword in the  $\mathbf{h}_{\text{hor}}$  direction. Therefore, the whole search process is divided into two stages. In the first stage, the  $\Xi$  codebook is used for searching in the  $\mathbf{h}_{\text{hor}}$  direction, and an omnidirectional codeword  $\xi_{\text{ver}}$  is used in the  $\mathbf{h}_{\text{ver}}$  direction. According to formula (??), we can get:

$$\xi_{\text{ver}}[n_{\text{ver}}] = -\frac{2n_{\text{ver}}(n_{\text{ver}} + 1)}{N_{\text{ver}}} + 2n_{\text{ver}}, n_{\text{ver}} = 1, \dots, N_{\text{ver}}. \quad (46)$$

In the  $s$ th layer search, codewords of  $\Xi^s$  are used in the  $\mathbf{h}_{\text{hor}}$  direction, and  $\xi_{\text{ver}}$  is used in the  $\mathbf{h}_{\text{ver}}$  direction. Thus the codes of  $s$  layer search in stage 1 can be written as:

$$\Phi_{p,0} = \text{diag}(\xi_p^s \otimes \xi_{\text{ver}}). \quad (47)$$

Then we select the index of the best codewords by the following:

$$p^* = \arg \max_p \mathbf{d}^H \Phi_{p,0} \mathbf{a}. \quad (48)$$

Then repeat the above searching procedure until the bottom layer of  $\Xi$ . At this time, the searching procedure in the  $\mathbf{h}_{\text{hor}}$  direction is completed, and the second stage begins to search for the  $\mathbf{h}_{\text{ver}}$  direction.

---

**Algorithm 2** DWS codebook design

---

- 1: Initialize the layer index  $s = 1$  and  $p^* = 1$  for stage 1.
  - 2: **repeat**
  - 3:   Compute  $\Phi_{p,0}$ , where  $p \in \{2p^* - 1, 2p^*\}$  by (??).
  - 4:   Update  $p^*$  by choosing the index of the best codes of  $s$ th layer by (??).
  - 5:   Update the layer index:  $s = s + 1$ .
  - 6: **until**  $s > S$ .
  - 7: Initialize the layer index  $s = 1$  and  $q^* = 1$  for stage 2.
  - 8: **repeat**
  - 9:   Compute  $\Phi_{p^*,q}$ , where  $q \in \{2q^* - 1, 2q^*\}$  by (??).
  - 10:   Update  $q^*$  by choosing the index of the best codes of  $s$ th layer by (??).
  - 11:   Update the layer index:  $s = s + 1$ .
  - 12: **until**  $s > S$ .
  - 13:  $\Phi_{\text{best}} = \text{diag}(\xi_{p^*}^S \otimes \xi_{q^*}^S)$ .
-



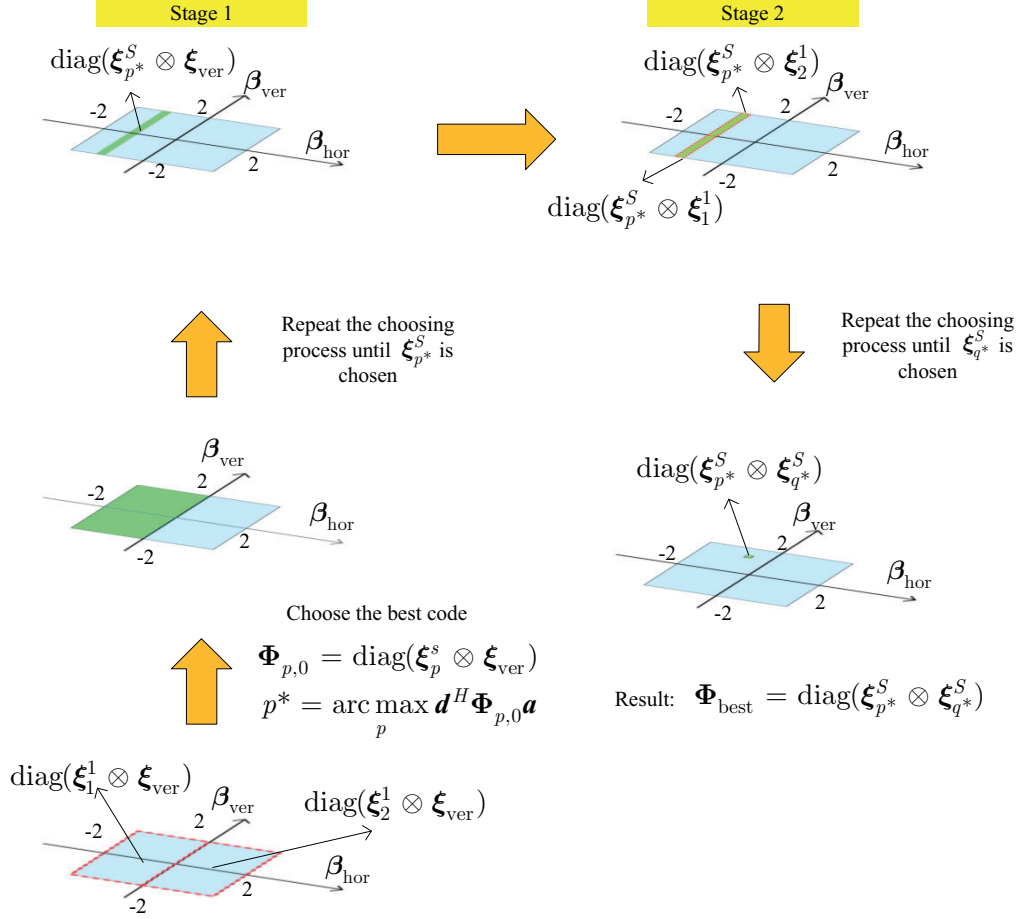


Fig. 4. DWS beam training process.

In the stage 2, the  $\mathbf{h}_{\text{hor}}$  direction uses the selected codewords in the stage 1, that is  $\xi_{p^*}^S$ , and the  $\mathbf{h}_{\text{ver}}$  direction uses  $\Xi$  for hierarchical search. For the  $s$ th layer search, the phase shift matrices of IRS can be written as:

$$\Phi_{p^*,q} = \text{diag}(\xi_{p^*}^S \otimes \xi_q^S). \quad (49)$$

In the same way, the optimal  $q$  is selected according to the received signal:

$$q^* = \arg \max_q \mathbf{d}^H \Phi_{p^*,q} \mathbf{a}. \quad (50)$$

Repeat the above searching procedure until the codes in the last layer of the hierarchical codebook are searched. Finally we get the best codeword  $\Phi_{\text{best}} = \text{diag}(\xi_{p^*}^S \otimes \xi_{q^*}^S)$ . The details of the proposed algorithm are summarized in Algorithm ??.

#### D. Performance Comparison of Algorithm and Analysis of Time Complexity

In this section, the JS and DWS schemes are compared in terms of their time complexity and designing flexibility.

When the IRS array is a square matrix, i.e.,  $N_{\text{ver}} = N_{\text{hor}}$ , and both JS and DWS schemes use binary tree structured codebook for searching, their time complexities become the same, which is  $\mathcal{O}(\lceil \log_2 N \rceil)$ . However, if the hierarchical codebook are changed from a binary tree structure to an  $M$ -ary tree structure, the time complexity of the JS scheme becomes  $\mathcal{O}(M^2 \lceil \log_M N \rceil)$ , and that of the DWS scheme becomes  $\mathcal{O}(2M \lceil \log_M N \rceil)$ . Accordingly, the DWS scheme takes less time.

In terms of the design freedom and algorithmic flexibility, the DWS algorithm is better than the JS algorithm. In the JS scheme, although the codebooks used in the  $\mathbf{h}_{\text{hor}}$  and  $\mathbf{h}_{\text{ver}}$  directions are binary tree structures, the entire JS algorithm actually searches in a quad-tree structure as the  $\mathbf{h}_{\text{hor}}$  and  $\mathbf{h}_{\text{ver}}$  directions are searched at the same time. Hence, we can conclude that the JS scheme cannot implement a search structure with the branch number less than 4 or an odd number of branches. The DWS scheme can realize the search structure of any number of branches. In addition, the JS scheme can be used only when the IRS array is a square array, while the DWS scheme does not have this limitation. Therefore, the DWS scheme is more flexible.

However, in some special scenarios, it is not necessary to complete all the searching processes, i.e., it is not necessary to search for any narrow beam at the bottom layer of the hierarchical codebook. For example, when the array size is practically large, the power required for communication can already be achieved by using wide beams in the middle layer of the hierarchical codebook. In this situation, completing the entire searching process will waste time and increases delay. In addition, if UEs moves slowly, the use of a wider beam can avoid the problem of frequent beam switching. In the above scenarios, we need to terminate the searching progress before the beam training process is completed, and then directly use the beam of the intermediate result of the beam training process for communication. The searching process of the JS scheme needs to simultaneously narrow the scopes of the  $\mathbf{h}_{\text{hor}}$  and  $\mathbf{h}_{\text{ver}}$  directions, while that of the DWS scheme needs to first narrow the scope of the  $\mathbf{h}_{\text{hor}}$  direction, and then that of the  $\mathbf{h}_{\text{ver}}$  direction. Therefore, the intermediate results of the DWS search are not suitable for the above scenarios, while the JS scheme can be interrupted at any time.

It is worth mentioning that the JS and DWS schemes are not "either-or" choices. Sometimes

TABLE I  
TIME COMPLEXITY COMPARISON.

Algorithm	Time complexity
Iterative optimization method [?]	$\mathcal{O}(N^{6.5})$
Riemannian optimization [?]	$\mathcal{O}(N^4)$
Proposed NCPD (continuous phase shift function)	$\mathcal{O}(1)$
Proposed NCPD (final result)	$\mathcal{O}(N^2)$

we can use the JS and DWS schemes simultaneously for beam training to combine the individual advantages of the two schemes. For example, we can perform the JS search first, and stop when the JS search is not completed, and then switch to the DWS scheme for continuing the beam training in the area to be searched. This not only takes advantage of the feature that that can interrupt the JS scheme in the middle, but also takes advantage of the lower time complexity of the DWS scheme.

## VI. PERFORMANCE AND TIME COMPLEXITY ANALYSIS

In this section, we will explain the difference between NCPD algorithm and traditional beamforming algorithm from the perspective of design principle, performance and time complexity, and introduce the main advantages of NCPD algorithm on large arrays.

The traditional beamforming design algorithms are designed directly based on the discrete phase shift matrix, while the NCPD algorithm designs the continuous phase shift function first, and then obtains the final discrete beamforming solution through sampling progress. The time spent in obtaining discrete beamforming solutions from continuous phase shift function sampling can often be ignored. However, in order to show the performance of NCPD more comprehensively, we have written down the time complexity of both forms of solutions in Table ???. Since (??) is irrelevant to  $N$ , the time complexity of continuous phase shift function is  $\mathcal{O}(1)$ , while the discrete result in (??) is  $\mathcal{O}(N^2)$ .

In fact, the optimized beamforming design method and the NCPD algorithm are two completely different design logics. From Table ??, we can see that on large arrays, optimization algorithms and NCPD algorithms have high flexibility and performance, while on-off elements algorithm, beam combination algorithm, and NCPD algorithm have lower computational time. In general, NCPD algorithm has obvious advantages on large arrays. Among the above algorithms,

only NCPD algorithm has both high beam design freedom and low time complexity. Moreover, the larger the array, the better is the performance of the NCPD algorithm, and the more obvious is the advantages.

TABLE II  
COMPARISON OF STRUCTURE BETWEEN OPTIMIZATION ALGORITHM AND NCPD ALGORITHM.

Algorithm	Performance	Time
Optimization methods [?], [?]	Depends on the construction of the objective function and constraint function, as well as the setting of iteration step size and number of iterations.	The larger the array, the longer is the operation time.
Sub-array and beam combination methods [?], [?], [?]	There are three factors that determine performance: the first is the number of sections to be divided, the second is the array size, and the third is the width and shape of the target beam.	The difference in computation time between large and small arrays is not significant.
On-off elements	RF chains can be used on active arrays to compensate for the losses caused by closing elements, while passive arrays cannot compensate for this part of the losses, resulting in poor performance. Meanwhile, this algorithm has low design flexibility.	Operation time is independent of array size.
Proposed NCPD	Performance is only determined by the size of the array, and the larger the array, the better is the performance.	The time of a continuous form solution is not related to the size of the array, but only to the complexity of the target beam. The discrete form solution only increases the sampling time above it, and the sampling time can often be ignored.

## VII. SIMULATION RESULTS

In this section, the theoretical results described above are verified, and the performance of the NCPD algorithm is show in forming beams of different widths on IRS of different sizes. Since the  $\mathbf{h}_{\text{hor}}$  and  $\mathbf{h}_{\text{ver}}$  directions are independent of each other and the beamforming designs in these two directions are completely equivalent, we show the array factor in one direction only. Therefore, the number of elements refers to the value of either  $N_{\text{ver}}$  or  $N_{\text{hor}}$ . Since the value of  $\beta_{\text{hor}}$  is  $(-2, 2)$ , we use  $\frac{\pi}{2}\beta_{\text{hor}}$  as the independent variable for forming the array factor in the polar coordinate system. But the concept of "beam width" still refers to the width of range of  $\beta_{\text{hor}}$ . For example if the range of  $\beta_{\text{hor}}$  in a wide beam is  $(0, 0.5)$ , then the width of that wide beam will be 0.5.

In this section, we use the beam combination method as a benchmark algorithm for wide beamforming problems. In the beam combination method, IRS is divided into many sub-arrays, each of which forms a narrow beam, and then these narrow beams are superimposed to form a wide beam. In this experiment, we choose two beam combination methods, in which four sub-beams are used to synthesize wide beams of 4-combination and sixteen sub-beams are used to synthesize wide beams of 16-combination.

#### A. Comparison under Different Array Sizes.

Fig. ?? shows the plots of the results of the beam combination and NCPD algorithms obtained in forming beams of width 1 at different array sizes. Figs. ??(a) and ??(b) show those of the 4-combination and 16-combination algorithms and NCPD algorithm, respectively. As can be seen from Fig. ??(a), the performance of the 4-combination algorithm is the best when the total number of elements is 64. Due to the insufficient elements in the sub-arrays when the total number of elements is less than 64, the formed sub-beam is not accurate enough, which means that the main lobe of the sub-beam is relatively wide and the side lobes are relatively high, thus enabling the main lobes of the four sub-beams to get connected with each other. However, when the total number of elements is greater than 64, it can be seen that each sub-beam is more accurate and its main lobe is narrower, resulting in a deep depression among several sub-beams.

From Fig. ??(b), we can see that when the total number of elements is 64, the performance of the 4-combination algorithm is better than that of the 16-combination algorithm. It means that when the total number of elements is small, we need to use fewer sub-beams to achieve a beam combination. The main reason of this phenomenon is that when the total number of elements is small, the number of elements allocated to each sub-beam in the 4-combination algorithm is more than that in the 16-combination algorithm, which makes each sub-beam in the combination algorithm more accurate and to have smaller side lobes. When the total number of elements increases to 128, we can see that the 16-combination algorithm does not have the deep depression phenomenon that the 4-combination algorithm does. When the total number of elements increases to 1024, the performance of the 16-combination algorithm becomes the best, which means that the more the total number of elements, the more sub-arrays are needed to form more sub-beams for combining wide beams. However, when the total number of elements is greater than 1024, the 16-combination algorithm also has a deep depression phenomenon. Since the more the number of elements, the more the power of the sub-beams is concentrated in their

center direction, no matter how many sub-beams are used for combination, but the phenomenon of deep depression cannot be avoided.

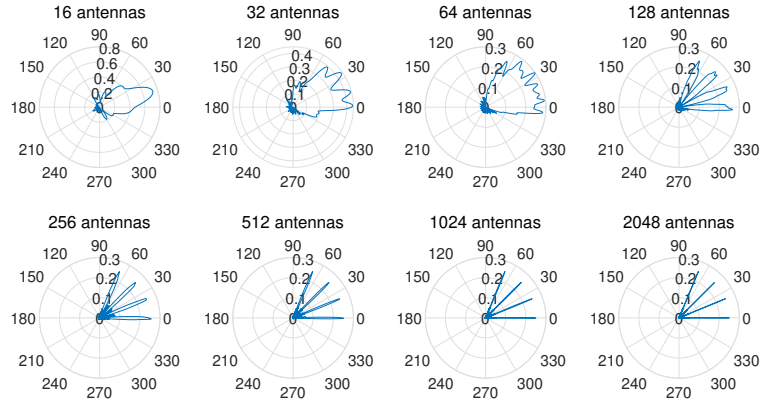
From Fig. ??(c), we can see that the NCPD algorithm is slightly better than the beam combination algorithm when the number of elements is small. However, when the number of elements increases, the NCPD algorithm does not achieve deep depression, and its performance is much better than that of the beam combination algorithm. With the increase in the number of elements, the side lobes of the beam formed by the NCPD algorithm become smaller, the main lobe becomes flatter and the edge of the main lobe becomes sharper, which gradually tends to our design goals. This shows that the NCPD algorithm can solve the problem of deep recesses when the number of elements increases. The NCPD algorithm is applicable to arrays of any size, and there is no need to artificially change the beamforming algorithm for the array size. Therefore, we can conclude that the NCPD algorithm is robust to the number of elements. Therefore, the time complexity of the NCPD algorithm is  $\mathcal{O}(1)$ , and the performance and speed in solving the IRS phase shift matrix in different scenarios are significantly better than those of the beam combination algorithm.

#### B. Comparison under Different Beam Width.

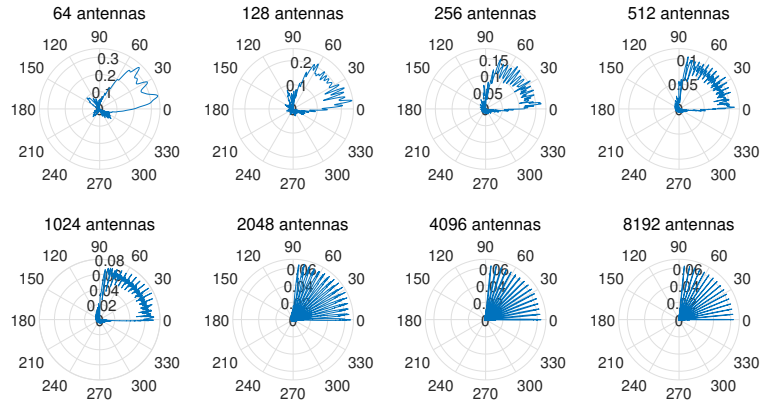
Fig. ?? shows the performance comparison when the beam combination and NCPD algorithms form wide beams with widths of 0.25, 0.5, 1, and 2 on 256 elements, respectively. Figs. ??(a)-??(c) show the performance of the 4-combination, the 16-combination, and the NCPD algorithms, respectively. We can see that both 4-combination and 16-combination algorithms have deep depressions when they synthesize beams with large widths, and the deep depression of the 4-combination algorithm is more serious. The performance of the NCPD algorithm is slightly better than those of the 4-combination and 16-combination algorithms when they form a narrow beam. However, when a wide beam is formed, there is no deep depression, and the effect is much better than those of the two beam combination algorithms. From this, it can be concluded that the NCPD algorithm is also robust when forming beams of different widths.

#### C. Performance of NCPD Algorithm on Giant IRS

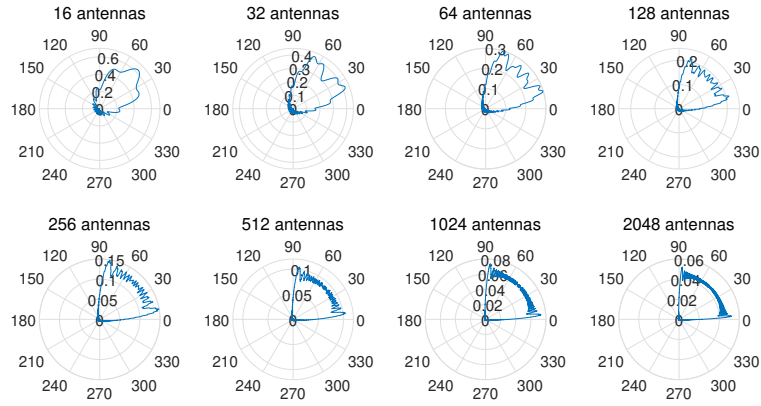
Fig. ?? shows the results of the NCPD algorithm forming beams of different widths for the number of elements of 10000, where we can see that the beam has almost no side lobes, its edge is sharply cut off, and the power in it is flatly distributed. This shows the superiority of the



(a) 4-combination algorithm.



(b) 16-combination algorithm.



(c) NCPD algorithm.

Fig. 5. Comparison of beam combination algorithm and NCPD algorithm in different sizes of IRS.

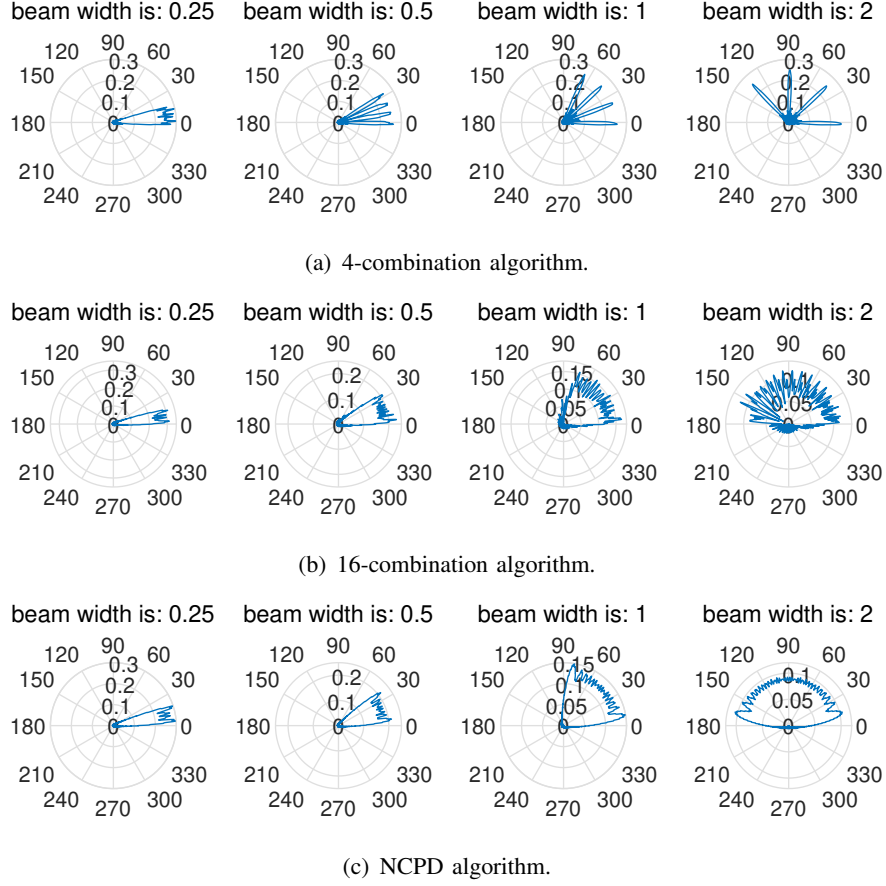


Fig. 6. Comparison of beam combination algorithm and NCPD algorithm when forming different width beams.

NCPD algorithm in dealing with beamforming problems on large-scale arrays. We found that at the edge of each beam, there is an overshoot phenomenon, which is the Gibbs phenomenon in beamforming. The Gibbs phenomenon in beamforming arises as IRS consists of discrete elements rather than a continuous reflector. Therefore, the beam-forming error caused by the Gibbs phenomenon is an inherent error of this problem, which cannot be eliminated by changing the algorithm.

Figure ?? shows the NCPD algorithm forming a trapezoidal beam on IRS with 10000 elements. The control function for beam shape is:  $h(\beta) = \beta, \beta \in (0.5, 1)$ . Trapezoidal beams cannot be used for hierarchical codebooks. This simulation is only to demonstrate the ability of the NCPD algorithm in controlling beam shapes. It can be seen that this trapezoidal beam still has the characteristic of low sidelobes, and its main lobe shape is completely consistent with the pre-set shape.



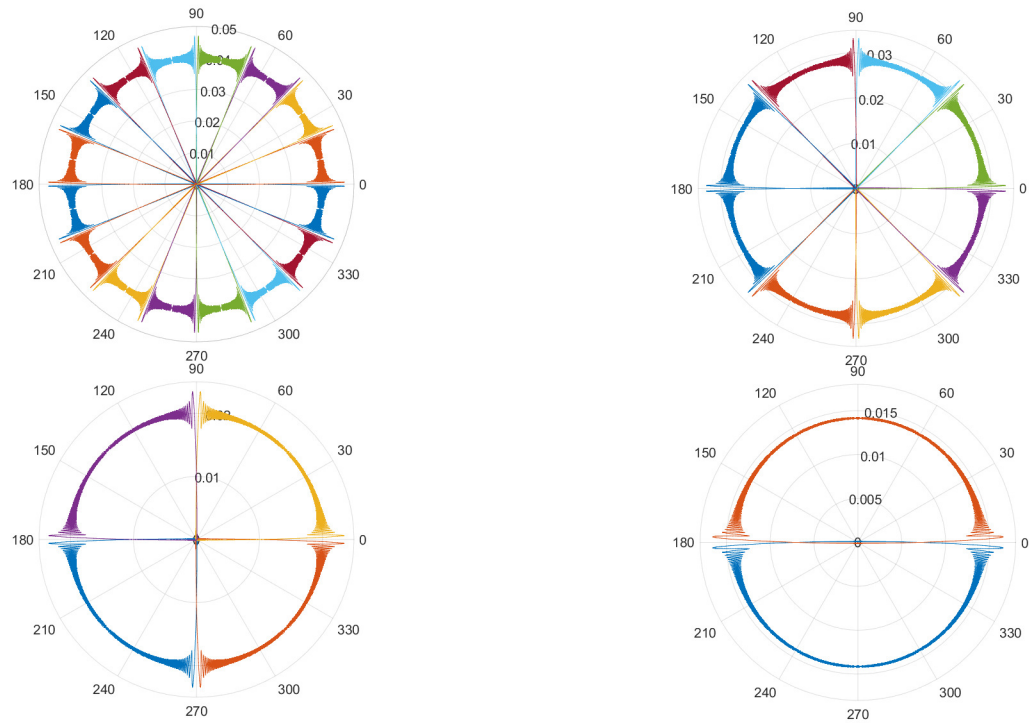


Fig. 7. The beams formed by NCPD algorithm on 10000 element IRS.

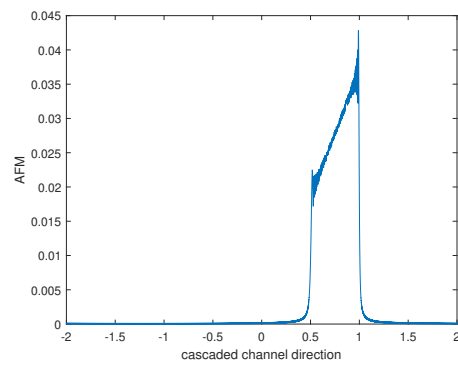


Fig. 8. The trapezoidal beams formed by NCPD algorithm on 10000 element IRS.

#### D. Comparison of Misalignment Rates between NCPD and Beam Combination Algorithms

Next, we compare the misalignment rates of wide beams formed by the NCPD, beam combination algorithms and the beam in the ideal case. The beam combination algorithm is the BMW-SS algorithm proposed in [?]. In the BMW-SS algorithm, a binary tree is also used for implementing a hierarchical codebook, in which the bottom layer also uses a narrow beam. In the case of wide beams, the BMW-SS algorithm uses  $2^{\lfloor \frac{l+1}{2} \rfloor}$  sub-beams for synthesis, where  $l$  represents the number of layers away from the bottom layer. For example, when forming a wide beam of the penultimate layer,  $l = 1$ , that is, the array is divided into two sub-arrays, and two sub-beams are respectively realized, and then the two sub-beams are synthesized into the required wide beam. Since the wider the beam, the smaller is the average power in each direction, the codeword of the top layer has the highest probability of misalignment so under the same signal-to-noise ratio (SNR). Therefore, the misalignment rate of the hierarchical codebook mainly depends on that of the first layer of codewords. Hence, we show the comparison of the misalignment only of the first layer of codewords. The ideal case means the ideal beams in the coverage of  $(-2, 0)$  and  $(0, 2)$  with the power of  $2N_{ver}$  or  $2N_{hor}$ , which are the target wide beams of the first layer of the hierarchical codebook. It is worth mentioning that there're no codewords corresponding to ideal case. We just suppose the formed beams are in the ideal shape and test the misalignment rate with the use of beams in this shape.

We conducted two experiments to show the comparison between the NCPD and the BMW-SS algorithms under different IRS received SNRs and different IRS array sizes. The IRS received SNR refers to the ratio of the total received power of IRS to the noise power of the cascaded channel. Since we choose the total received SNR of IRS rather than the received SNR on a single element, we can exclude the gain caused by the increase in the IRS array size, and show only the beamforming gain of IRS.

Fig. ?? shows the misalignment curves of the first-layer codewords of the NCPD and the BMW-SS algorithms and the ideal case under different IRS received SNRs on the uniform linear array (ULA) of 256 elements. We can see that when SNR is less than 0dB, the misalignment rates of the two algorithms are high. This is because the noise power at this stage is too large, and IRS cannot use the beamforming algorithm to offset the noise interference. But when SNR is greater than 0dB, the misalignment rate of the NCPD algorithm is significantly lower than that of the BMW-SS algorithm as the power distribution in the beam formed by the NCPD

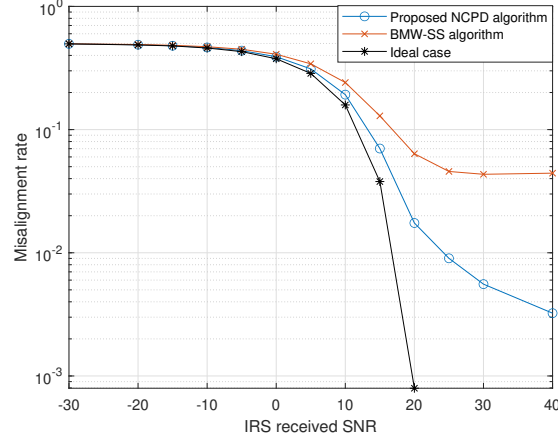


Fig. 9. Comparison of beam misalignment rates under different IRS received SNR.

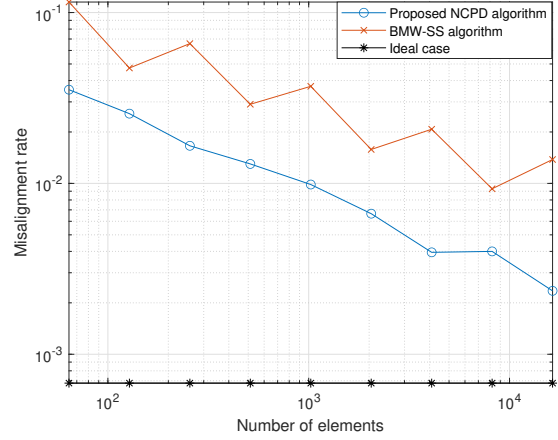


Fig. 10. The misalignment rates on different IRS size when IRS received SNR is 20 dB.

algorithm is more uniform. At the same time, we noticed that when SNR is greater than 30dB, the misalignment rate of the NCPD algorithm is still declining, while that in the BMW-SS algorithm is almost unchanged. This is because when SNR is high, the deep depression problem becomes the main reason for limiting the performance of the BMW-SS algorithm, while the performance of the NCPD algorithm is gradually improved with increasing SNR as it has no deep depression problem.

Fig. ?? and Fig. ?? show the misalignment rates of the NCPD and BMW-SS algorithms on ULAs of different sizes when the IRS received SNR is 10dB and 20dB, respectively. The reason why the curve of the BMW-SS algorithm is a broken line is that the relationship between the

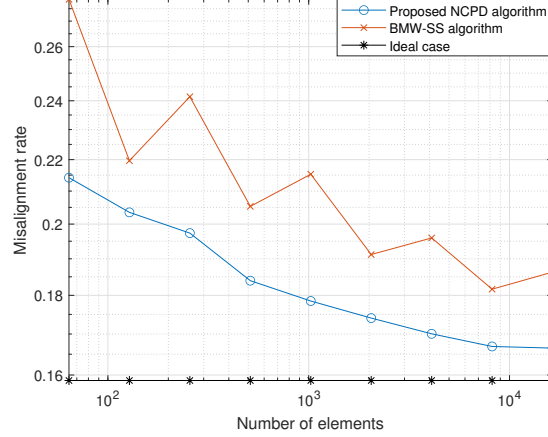


Fig. 11. The misalignment rates on different IRS size when IRS received SNR is 10 dB.

number of sub-arrays and number of array elements is not a continuous function. For example, when the number of elements is 128 or 256, the BMW-SS algorithm divides IRS into 16 sub-arrays. But when the number of elements is 512, it divides IRS into 32 sub-arrays. When the IRS array size is in the range from 128 to 256, the number of sub-arrays does not change. But due to the increase in the total number of elements, the number of elements in each sub-array increases, which leads to a more serious deep depression problem in the combined beam and an increase in the misalignment rate of the beam. This also shows that the deep depression problem is the performance bottleneck when dealing with large-scale array problems using the conventional beam combination algorithm. The performance of the NCPD algorithm is gradually improved with the increase in the array size. In all the experimental array sizes, the misalignment of the NCPD algorithm is lower than that of the BMW-SS algorithm, which proves that the NCPD algorithm is superior in solving the beamforming problem of large-scale IRS arrays.

## VIII. CONCLUSION

Aiming at the characteristics of IRS-aided communication scenarios, in this paper we proposed a new array factor model. In this model, we proposed the variables of the cascaded channel direction to replace the conventional elevation and azimuth angles for describing the AoA and AoD of IRS which can reduce the number of parameters. In addition, we used the integral form to replace the summation form in the conventional array factor. In order to control the shape of the array factor function of IRS, we proposed the NCPD algorithm and an analytical solution of

beamforming on IRS. The time complexity of the analytical solution was  $\mathcal{O}(1)$ , which solved the problem that beamforming design takes too much time on large-scale arrays. Using beams formed by the analytical solution, we proposed a hierarchical codebook and two beam training schemes, namely the JS scheme and DWS scheme. These two schemes could also be used in combination to form a more flexible beam training scheme for adapting to different scenarios.

The simulation results showed that the performance of the NCPD algorithm was better than those of the conventional beam combination algorithms under different array sizes, and the beam formed by the NCPD algorithm did not suffer from any deep depression problem. At the same time, the misalignment rate of the beam formed by the NCPD algorithm was lower than those of the beam combination algorithms under different array sizes and IRS received SNR. It could be seen that the NCPD algorithm could significantly reduce the misalignment rate of the beam.

There are many important issues not yet addressed in this paper, some of which are listed below for motivating future works. Firstly, we assumed that the phase of each element of IRS can be configured to any value. However, in practice, it is often not satisfied. Secondly, The scenario used in this paper was a single IRS and single UE scenario. In actual situations, multiple IRSs are often deployed together. Thirdly, this paper focused on scenarios of static UE. In a UE movement scenario, how to balance the beam training time and beamforming gain of IRS is worthy of further investigation. Finally, in this paper we considered the scenario where both BS and UE are in the far field of IRS. However, IRS often needs to be deployed close to one of BS and UE. Therefore, the near-field scenario is more common.

## APPENDIX A

### PROOF OF PROPOSITION 1

Suppose that there are two directions  $\beta_{\text{hor},1}$  and  $\beta_{\text{hor},2}$  satisfying:

$$kd\beta_{\text{hor},1} = kd\beta_{\text{hor},2} + 2m'\pi, \quad (51)$$

where  $m'$  is an integer. For any  $g_{\text{hor}}(n_{\text{hor}})$ , the following equation holds:

$$\begin{aligned}
 AFM_{\text{norm}}(\beta_{\text{hor},2}) &= \left| \sum_{n_{\text{hor}}=0}^{n_{\text{hor}}-1} e^{jk d [\beta_{\text{hor},2} n_{\text{hor}} + g_{\text{hor}}(n_{\text{hor}})]} \right| \\
 &= \left| \sum_{n_{\text{hor}}=0}^{N_{\text{hor}}-1} e^{jk d [\beta_{\text{hor},1} n_{\text{hor}} + g_{\text{hor}}(n_{\text{hor}})] + j 2 m \pi} \right| \\
 &= \left| \sum_{n_{\text{hor}}=0}^{n_{\text{hor}}-1} e^{jk d [\beta_{\text{hor},1} n_{\text{hor}} + g_{\text{hor}}(n_{\text{hor}})]} \cdot e^{j 2 m \pi} \right| \\
 &= \left| \sum_{n_{\text{hor}}=0}^{n_{\text{hor}}-1} e^{jk d [\beta_{\text{hor},1} n_{\text{hor}} + g_{\text{hor}}(n_{\text{hor}})]} \right| \\
 &= AFM_{\text{norm}}(\beta_{\text{hor},1}).
 \end{aligned} \tag{52}$$

From the (??) we can see that no matter what value  $g_{\text{hor}}(n_{\text{hor}})$  takes,  $AFM_{\text{norm}}(\beta_{\text{hor},1})$  is always equal to  $AFM_{\text{norm}}(\beta_{\text{hor},2})$ . This means that if the IRS has a beam in the  $\beta_{\text{hor},1}$  direction, there must also be a beam in the  $\beta_{\text{hor},2}$  direction, which will cause interference to other users. We need to choose suitable  $d$  to make sure there does not exist  $\beta_{\text{hor},1}$  and  $\beta_{\text{hor},2}$  satisfy (??). From (??) and (??) we know the range of  $\beta_{\text{hor}}$  is  $(-2, 2)$ , so  $kd \leq \frac{\pi}{2}$ ,  $d \leq \frac{\lambda}{4}$ . Thus, the max element spacing distance on IRS is  $\frac{\lambda}{4}$ . Without loss of generality, we choose  $d = \frac{\lambda}{4}$  in this paper.

**Remark 3.** *The conclusion of the maximum element spacing distance on the IRS is different from that on the active array. That is because beamforming on IRS serves cascaded channel while beamforming on active arrays serve one hop channel. For active array the range of  $\beta_{\text{hor}}$  is  $(-1, 1)$  while on IRS it is  $(-2, 2)$ , so the max element spacing distance on active array is  $\frac{\lambda}{2}$  while on IRS is  $\frac{\lambda}{4}$ .*

## APPENDIX B

### PROOF OF PROPOSITION 2

Since we need to find a general solution to the beamforming problem for arbitrary array sizes, we need to replace  $g_{\text{hor}}(n_{\text{hor}})$  with a continuous function. We let:

$$g_{\text{hor}}(n_{\text{hor}}) = \frac{1}{d} \int_0^{n_{\text{hor}} d} -f(u) du, \tag{53}$$

where  $f(u)$  is a continuously differentiable function. So (??) can be written as:

$$\begin{aligned}
 AFM_{\text{norm}}(\beta_{\text{hor}}) &= \frac{1}{N_{\text{hor}}} \left| \sum_{n_{\text{hor}}=0}^{N_{\text{hor}}-1} e^{jk [d \beta_{\text{hor}} n_{\text{hor}} - \int_0^{n_{\text{hor}} d} f(u) du]} \right| \\
 &= \frac{1}{N_{\text{hor}} d} \left| \sum_{n_{\text{hor}}=0}^{N_{\text{hor}}-1} e^{jk [d \beta_{\text{hor}} n_{\text{hor}} - \int_0^{n_{\text{hor}} d} f(u) du]} d \right|.
 \end{aligned} \tag{54}$$

Let  $\dot{u} = n_{\text{hor}}d$ . If the IRS array size tends to be infinite while IRS physical size is fixed, which means that  $N_{\text{hor}}d$  is a constant and  $N$  tends to be infinite, so  $d$  tends to 0 and (??) can be written as a integral form:

$$AFM_{\text{norm}}(\beta_{\text{hor}}) = \frac{1}{N_{\text{hor}}d} \left| \int_0^{N_{\text{hor}}d} e^{jk[\beta_{\text{hor}}\dot{u} - \int_0^{\dot{u}} f(u)du]} d\dot{u} \right|. \quad (55)$$

Let  $\ddot{u} = \frac{\dot{u}}{N_{\text{hor}}d}$ . Then, (??) can be written as:

$$AFM_{\text{norm}}(\beta_{\text{hor}}) = \left| \int_0^1 e^{jk[\beta_{\text{hor}}N_{\text{hor}}d\ddot{u} - \int_0^{\ddot{u}N_{\text{hor}}d} f(u)du]} d\ddot{u} \right|. \quad (56)$$

Let  $\mu = \frac{u}{N_{\text{hor}}d}$ . We further have:

$$\int_0^{\ddot{u}N_{\text{hor}}d} f(u)du = N_{\text{hor}}d \int_0^{\ddot{u}} f[J(\mu)]d\mu, \quad (57)$$

where  $J(\mu) = N_{\text{hor}}d\mu$ . Note that the (??) is an integral with respect to  $\mu$ . By setting  $\dot{f}(\cdot) = f[J(\cdot)]$ , we rewrite (??) as:

$$\int_0^{\ddot{u}N_{\text{hor}}d} f(u)du = N_{\text{hor}}d \int_0^{\ddot{u}} \dot{f}(\mu)d\mu. \quad (58)$$

Substituting (??) into (??), we can get:

$$AFM_{\text{norm}}(\beta_{\text{hor}}) = \left| \int_0^1 e^{jkN_{\text{hor}}d[\beta_{\text{hor}}\ddot{u} - \int_0^{\ddot{u}} \dot{f}(\mu)d\mu]} d\ddot{u} \right|. \quad (59)$$

From (??), we can get that the IRS array factor can be determined by a function  $\dot{f}(\mu)$  defined on  $(0, 1)$ . Thus (??) holds. Since  $\ddot{u} = \frac{n_{\text{hor}}}{N_{\text{hor}}}$ , the relationship between  $g_{\text{hor}}(n_{\text{hor}})$  and  $\dot{f}(\mu)$  is:

$$g_{\text{hor}}(n_{\text{hor}}) = -N_{\text{hor}} \int_0^{\frac{n_{\text{hor}}}{N_{\text{hor}}}} \dot{f}(\mu)d\mu. \quad (60)$$