Simultaneous Multiuser Beam Training using Adaptive Hierarchical Codebook for mmWave Massive MIMO

Kangjian Chen*, Chenhao Qi*, Octavia A. Dobre† and Geoffrey Li‡
*School of Information Science and Engineering, Southeast University, Nanjing, China
†Faculty of Engineering and Applied Science, Memorial University, Canada
‡School of Electrical and Computer Engineering, Georgia Institute of Technology, USA
Email: {kjchen,qch}@seu.edu.cn, odobre@mun.ca, liye@ece.gatech.edu

***Abstract*—In this paper, a simultaneous multiuser hierarchical beam training scheme for multiuser mmWave massive MIMO systems is proposed based on the designed adaptive hierarchical codebook. Different from the existing work sequentially performing the beam training for different users with the same hierarchical codebook, in our work the hierarchical codebook is designed in an adaptive manner, where the codewords in the current layer are designed according to the beam training results of the previous layer. In particular, multi-mainlobe codewords are designed for simultaneously beam training with all the users, where each mainlobe of the multi-mainlobe codeword covers a spatial region that one or more users are probably in. Except for the bottom layer, there are only two codewords at each layer in the designed adaptive hierarchical codebook, which only requires two times of simultaneous beam training for all the users no matter how many users the BS serves. Simulation results verify the effectiveness of our scheme and show that our scheme can approach the performance of the beam scanning but with considerable reduction in training overhead.**

***Index Terms*—Millimeter wave (mmWave) communications, massive MIMO, beam training, hierarchical codebook.**

I. Introduction

Millimeter wave (mmWave) massive MIMO has been considered as a promising technology for future wireless communications due to its rich spectrum and spatial resources [1]–[4]. However, the transmission of mmWave signal experiences large path loss because of its high frequency. To overcome the shortcoming, large antenna arrays with hybrid precoding architecture have been introduced. For the hybrid precoding architecture, a small number of radio frequency (RF) chains are connected to a large number of antennas via phase shifters. It can save the energy consumption compared to the fully digital architecture [2]. On the other hand, a large number of antenna arrays with precoding can perform directional transmission and compensate the severe path loss of mmWave channel [5].

To acquire the channel state information (CSI) in the mmWave massive MIMO systems with a hybrid precoding structure, codebook-based beam training methods have been widely adopted [6]–[9]. Hierarchical codebook-based beam training can reduce the training overhead. It employs a predefined hierarchical codebook including several layers, where the spatial region covered by the codeword at the upper layer in the codebook is split into several smaller spatial regions covered by codewords at the lower layer [10]–[12]. Earlier work on hierarchical beam training focuses on peer-to-peer mmWave massive MIMO systems. In [13], a hierarchical codebook design method has been proposed, where the channel estimation is formulated as a sparse reconstruction problem. In [10], another hierarchical codebook design method, named joint sub-array and de-activation (JOINT), has been proposed. Recent work on hierarchical beam training focuses on multiuser mmWave massive MIMO systems. A straightforward extension of the above work to the multiuser scenario is time-division multiple access (TDMA) hierarchical beam training, where the BS sequentially performs the hierarchical beam training user by user and each user occupies a different part of time. However, the total training overhead grows linearly with the number of users. To reduce the overhead of beam training, a simultaneous hierarchical beam training for multiuser mmWave massive MIMO system has been proposed in [12]. However, it is only proposed for partially connected structure, where each RF chain is solely connected to an antenna subarray at the BS and the beam training is independently performed by each subarray.

In this paper, we propose a simultaneous multiuser hierarchical beam training scheme based on our designed adaptive hierarchical codebook for multiuser mmWave massive MIMO systems. Unlike the existing work sequentially performing the beam training for different users using the same hierarchical codebook, we design a hierarchical codebook in an adaptive manner, where the codewords in the current layer of the hierarchical codebook are designed according to the beam training results of the previous layer. In particular, we design multi-mainlobe codewords to perform the beam training simultaneously for all the users, where each mainlobe covers a spatial region that one or more users are probably in. Except for the bottom layer in the designed adaptive hierarchical codebook, there are only two codewords at each layer. Therefore, it only requires two times of simultaneous beam training for all the users no matter how many users the BS serves.

978-1-7281-0962-6/19/$31.00 ©2019 IEEE

The notations are defined as follows. Symbols for matrices (upper case) and vectors (lower case) are in boldface. $[\boldsymbol{a}]_n$, $[\boldsymbol{A}]_{:,n}$ and $[\boldsymbol{A}]_{m,n}$ denote the nth entry of a vector $\boldsymbol{a}$, the nth column of a matrix $\boldsymbol{A}$, and the entry on the mth row and nth column of $\boldsymbol{A}$. $\boldsymbol{I}$ denotes the identity matrix. $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and conjugate transpose (Hermitian), respectively. $\|\cdot\|_2$ and $\|\cdot\|_F$ denote ℓ_2-norm and Frobenius norm, respectively. $\mathbb{C}$, $\mathbb{E}\{\cdot\}$, $\circ$ and $\mathcal{CN}$ denote the set of complex number, operation of expectation, Kronecker product and complex Gaussian distribution, respectively.

II. System Model

We consider a multiuser mmWave massive MIMO system with a base station (BS) and K user equipments (UEs). The number of antennas at the BS and each UE is N_{BS} and N_{UE} ($N_{\mathrm{UE}} \leq N_{\mathrm{BS}}$), respectively. The number of RF chains at the BS and each UE is N_{RF} ($K \leq N_{\mathrm{RF}} \ll N_{\mathrm{BS}}$) and one, respectively. To simplify the analysis, both N_{BS} and N_{UE} are usually set as integer power of two. The BS employs hybrid precoding, including digital precoding and analog precoding while each UE employs analog combining. The antennas at both the BS and the UEs are placed into uniform linear arrays (ULAs) with half wavelength spacing.

During the downlink signal transmission from the BS to the UEs, the received signal by the kth UE, for $k = 1, 2, \ldots, K$, can be expressed as

$$y_k = \boldsymbol{w}_k^H \boldsymbol{H}_k \boldsymbol{F}_{\mathrm{RF}} \boldsymbol{F}_{\mathrm{BB}} \boldsymbol{s} + \boldsymbol{w}_k^H \boldsymbol{n}_k, \tag{1}$$

where y_k, $\boldsymbol{w}_k \in \mathbb{C}^{N_{\mathrm{UE}}}$, $\boldsymbol{H}_k \in \mathbb{C}^{N_{\mathrm{UE}} \times N_{\mathrm{BS}}}$, $\boldsymbol{F}_{\mathrm{RF}} \in \mathbb{C}^{N_{\mathrm{BS}} \times N_{\mathrm{RF}}}$, $\boldsymbol{F}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{RF}} \times K}$, $\boldsymbol{s} \in \mathbb{C}^K$ and $\boldsymbol{n}_k \in \mathbb{C}^{N_{\mathrm{UE}}}$ denote the received signal, the analog combiner of the kth UE, the channel matrix between the BS and the kth UE, the analog precoder of the BS, the digital precoder of the BS, the transmitted signal vector, and the additive white Gaussian noise vector obeying $\boldsymbol{n}_k \sim \mathcal{CN}(0, \sigma^2 \boldsymbol{I}_{N_{\mathrm{UE}}})$, respectively. Note that the hybrid precoder, including the analog precoder and the digital precoder, does not provide power gain, i.e., $\|\boldsymbol{F}_{\mathrm{RF}} \boldsymbol{F}_{\mathrm{BB}}\|_F^2 = K$. Moreover, $\boldsymbol{s}$ subjects to the constraint of total transmit power P, i.e., $\mathbb{E}\{\boldsymbol{s}\boldsymbol{s}^H\} = \frac{P}{K}\boldsymbol{I}_K$.

According to the Saleh-Valenzuela channel model [1], the mmWave MIMO channel matrix $\boldsymbol{H}_k \in \mathbb{C}^{N_{\mathrm{UE}} \times N_{\mathrm{BS}}}$ between the BS and the kth UE is expressed as

$$\boldsymbol{H}_k = \sqrt{\frac{N_{\mathrm{BS}} N_{\mathrm{UE}}}{L_k}} \sum_{l=1}^{L_k} \lambda_l \boldsymbol{\alpha}\big(N_{\mathrm{UE}}, \theta_{\mathrm{UE}}^l\big) \boldsymbol{\alpha}^H\big(N_{\mathrm{BS}}, \theta_{\mathrm{BS}}^l\big) \tag{2}$$

where L_k, λ_l, θ_{UE}^l and θ_{BS}^l denote the number of multipath, the channel gain, the channel angle-of-arrival (AoA), and the channel angle-of-departure (AoD) of the lth path, respectively. In fact, $\theta_{\mathrm{UE}}^l = \cos\big(\omega_{\mathrm{UE}}^l\big)$ and $\theta_{\mathrm{BS}}^l = \cos\big(\omega_{\mathrm{BS}}^l\big)$, where ω_{UE}^l and ω_{BS}^l denote the physical AoA and AoD of the lth path, respectively. It is obvious that $\theta_{\mathrm{UE}}^l \in [-1, 1]$ and $\theta_{\mathrm{BS}}^l \in [-1, 1]$. The channel steering vector $\boldsymbol{\alpha}(\cdot)$ in (2) is defined as

$$\boldsymbol{\alpha}\left(N, \theta\right) = \frac{1}{\sqrt{N}} \left[1, e^{j\pi\theta}, \cdots, e^{j(N-1)\pi\theta}\right]^T \tag{3}$$

where N is the number of antennas and θ is the channel AoA or AoD.

We denote the codebooks at the BS and each UE as $\mathcal{F} = \{\boldsymbol{f}_{\mathrm{c}}^1, \boldsymbol{f}_{\mathrm{c}}^2, \cdots, \boldsymbol{f}_{\mathrm{c}}^{N_{\mathrm{BS}}}\}$ and $\mathcal{W} = \{\boldsymbol{w}_{\mathrm{c}}^1, \boldsymbol{w}_{\mathrm{c}}^2, \cdots, \boldsymbol{w}_{\mathrm{c}}^{N_{\mathrm{UE}}}\}$, respectively, where

$$\begin{aligned} \boldsymbol{f}_{\mathrm{c}}^n &= \boldsymbol{\alpha}(N_{\mathrm{BS}}, -1 + (2n-1)/N_{\mathrm{BS}}), \\ \boldsymbol{w}_{\mathrm{c}}^m &= \boldsymbol{\alpha}(N_{\mathrm{UE}}, -1 + (2m-1)/N_{\mathrm{UE}}). \end{aligned} \tag{4}$$

The objective of beam training is to select K codewords from $\mathcal{F}$ for the BS and K codewords from $\mathcal{W}$ for the K UEs. Since designing $\boldsymbol{F}_{\mathrm{RF}}$ is essentially to find $\boldsymbol{f}_k \triangleq [\boldsymbol{F}_{\mathrm{RF}}]_{:,k}$, our objective turns to find a pair of $\boldsymbol{f}_k$ and $\boldsymbol{w}_k$ best fit for $\boldsymbol{H}_k$, which can be expressed as

$$\max_{\boldsymbol{f}_k, \boldsymbol{w}_k} \ |\boldsymbol{w}_k^H \boldsymbol{H}_k \boldsymbol{f}_k| \ \ \text{s.t.} \ \ \boldsymbol{f}_k \in \mathcal{F}, \ \boldsymbol{w}_k \in \mathcal{W}. \tag{5}$$

A straightforward method to solve (5) is the exhaustive beam training, also known as beam scanning [8], which tests all possible pairs of $\boldsymbol{f}_k$ and $\boldsymbol{w}_k$ to find the best one. However, such a method takes a long time and therefore has a large overhead. If we denote the time period of each test of a pair of $\boldsymbol{f}_k$ and $\boldsymbol{w}_k$ as a time slot, the exhaustive beam training needs totally $N_{\mathrm{BS}} N_{\mathrm{UE}}$ time slots. Note that the number of total time slots is independent of K since the UEs can simultaneously test the power of their received signal and eventually feed back the indices of the best codewords to the BS.

To reduce the overhead of exhaustive beam training, hierarchical beam training is widely adopted [10]. The hierarchical beam training usually first tests the mmWave channel with some low-resolution codewords at the upper layer and then narrows down the beam width layer by layer until a codeword pair at the bottom layer is obtained. We denote the hierarchical codebooks employed at the BS and UEs as $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}$ and $\boldsymbol{\mathcal{V}}_{\mathrm{UE}}$, respectively. The mth codeword at the sth layer of $\boldsymbol{\mathcal{V}}_{\mathrm{UE}}$ for $s = 1, \cdots, T$ and $m = 1, 2, \cdots 2^s$ is denoted as $\boldsymbol{\mathcal{V}}_{\mathrm{UE}}(s, m)$, where $T = \log_2 N_{\mathrm{UE}}$ is the number of layers of $\boldsymbol{\mathcal{V}}_{\mathrm{UE}}$. Note that the codewords at the bottom layer of $\boldsymbol{\mathcal{V}}_{\mathrm{UE}}$ are exactly the same as the codewords in $\mathcal{W}$, which implies that the motivation of hierarchical beam training is to adopt the merits of binary tree to improve the efficiency of beam training. Compared with the exhaustive beam training, the TDMA hierarchical beam training sequentially performed user by user in the TDMA fashion can reduce the overhead from $N_{\mathrm{BS}} N_{\mathrm{UE}}$ to $2K(\log_2 N_{\mathrm{BS}} + \log_2 N_{\mathrm{UE}})$. Therefore, the TDMA hierarchical beam training can reduce the training overhead by 92.2% compared to the exhaustive beam training if $K = 4$, $N_{\mathrm{BS}} = 64$ and $N_{\mathrm{UE}} = 16$.

III. Simultaneous Beam Training based on Adaptive Hierarchical Codebook

In this section, we will propose a simultaneous multiuser beam training scheme based on adaptive hierarchical codebook, which aims at considerably reducing the training overhead compared to the TDMA hierarchical beam training. Note that we will design a new hierarchical codebook for the BS, while for the UE we use the existing hierarchical codebook such as [10].

A. Simultaneous Multiuser Hierarchical Beam Training

The new hierarchical codebook for the BS is denoted by $\boldsymbol{\mathcal{C}}$ to distinguish it from the existing hierarchical codebook $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}$.

As shown in Fig. 1, the adaptive hierarchical codebook with totally $S = \log_2 N_{\mathrm{BS}}$ layers can be divided into three parts

1) The top layer of the codebook: In this layer, we equally divide the channel AoD $[-1, 1]$ into two codewords, so that the beam width of each codeword is one.
2) The bottom layer: It is essentially the Sth layer of the hierarchical codebook. This layer is exactly the same as the bottom layer of the existing hierarchical codebook and can be designed according to (4).
3) The intermediate layers: It includes the second layer to the $(S-1)$th layer of the codebook. Different from the existing codebook, each intermediate layer only includes two codewords, no matter how large K is. The beam coverage of codewords in the intermediate layers is intermittent. In particular, the codewords in the current layer are adaptively designed according to the beam training results of the previous layer. Note that the union of the beam coverage of two codewords in the same layer may not be $[-1, 1]$ because some spatial regions may not have any channel path and we do not need to waste signal beam to cover.

Now we focus on designing codewords in the first and intermediate layers of the adaptive hierarchical codebook. Note that the beam coverage of a codeword at these layers can be considered as the union of several codewords at the bottom layer. We can design the codewords in the first and intermediate layers by combining several codewords from the bottom layer of $\boldsymbol{\mathcal{C}}$. Then, the mth codeword at the sth layer of $\boldsymbol{\mathcal{C}}$, denoted as $\boldsymbol{\mathcal{C}}(s, m)$, for $s = 1, \cdots, S-1$ and $m = 1, 2, \cdots 2^s$, can be represented as

$$\boldsymbol{\mathcal{C}}(s, m) = \sum_{n \in \boldsymbol{\Psi}_{s,m}} e^{j\psi_n} \boldsymbol{f}_{\mathrm{c}}^n, \tag{6}$$

which is essentially a weighted summation of several channel steering vectors.

The indices of the codewords in $\mathcal{F}$ involved in the weighted summation form an integer set $\boldsymbol{\Psi}_{s,m}$. Here we introduce the phase ψ_n to explore the additional degree of freedom to avoid low beam gain within the beam coverage [14]. Based on our previous work [11], we can set ψ_n as

$$\psi_n = n\pi(-1 + 1/N_{\mathrm{BS}}). \tag{7}$$

To fairly compare different codewords in each beam training, we usually normalize $\boldsymbol{\mathcal{C}}(s, m)$ so that $\|\boldsymbol{\mathcal{C}}(s, m)\|_2 = 1$.

We design $\boldsymbol{\Psi}_{s,m}$ for $s = 1, \cdots, S-1$ and $m = 1, 2, \cdots 2^s$, as follows. We denote the beam coverage of $\boldsymbol{\mathcal{C}}(s, m)$ as $B_{s,m}$. Then, $B_{1,m}$ at the top layer can be expressed as

$$B_{1,m} = [m - 2, m - 1] \tag{8}$$

for $m = 1, 2$. We determine $\boldsymbol{\Psi}_{1,m}, m = 1, 2$ by

$$\boldsymbol{\Psi}_{1,m} = \{n | R_n \subseteq B_{1,m},\ n = 1, 2, \cdots, N_{\mathrm{BS}}\} \tag{9}$$

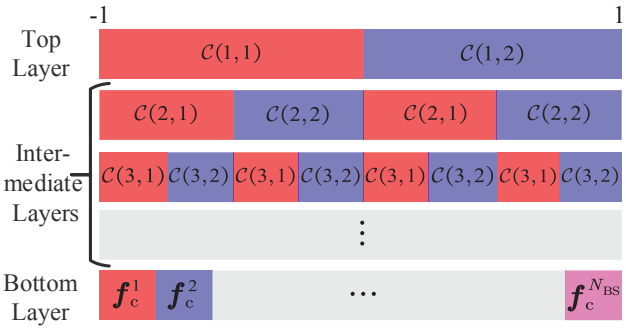

Fig. 1. Illustration of adaptive hierarchical codebook $\boldsymbol{\mathcal{C}}$.

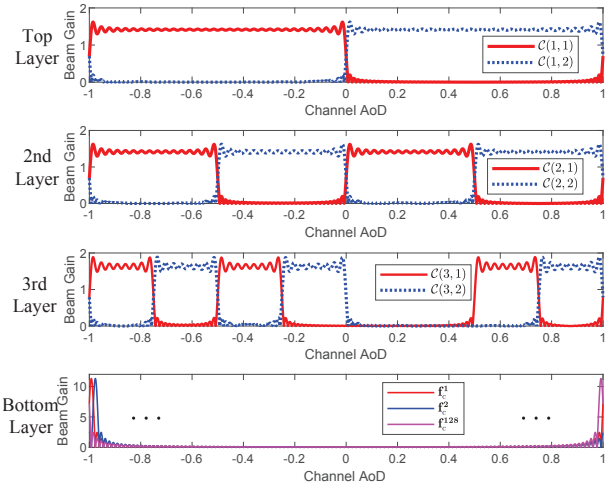

Fig. 2. Beam gain of different codewords in adaptive hierarchical codebook $\boldsymbol{\mathcal{C}}$ with $N_{\mathrm{BS}} = 128$, $N_{\mathrm{UE}} = 16$ and $K = 4$.

where $R_n \triangleq [-1 + (2n-2)/N_{\mathrm{BS}}, -1 + 2n/N_{\mathrm{BS}}]$ denotes the beam coverage of $\boldsymbol{f}_{\mathrm{c}}^n$.

The BS sequentially transmits $\boldsymbol{\mathcal{C}}(1, 1)$ and $\boldsymbol{\mathcal{C}}(1, 2)$ to all K UEs and each UE receives the signal with $\boldsymbol{\mathcal{V}}_{\mathrm{UE}}(1, 1)$ and $\boldsymbol{\mathcal{V}}_{\mathrm{UE}}(1, 2)$. Then, each UE compares the received signal power of $\boldsymbol{\mathcal{C}}(1, 1)$ and $\boldsymbol{\mathcal{C}}(1, 2)$ and individually feeds back the index of the larger codeword to the BS. We define $\mathcal{K} \triangleq \{1, 2, \ldots, K\}$ and a vector $\boldsymbol{\Gamma}_1$ of length of K, where $[\boldsymbol{\Gamma}_1]_k$ corresponds to the index of the larger codewords from the kth UE, i.e., $[\boldsymbol{\Gamma}_1]_k \in \{1, 2\}$.

Denote the index set of the selected codewords after beam training at the $(s-1)$th layer to be $\boldsymbol{\Gamma}_{s-1}$ for $s = 2, 3, \cdots, S-1$. According to the existing hierarchical beam training, at the sth layer, the BS will test $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}(s, 2[\boldsymbol{\Gamma}_{s-1}]_k - 1)$ and $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}(s, 2[\boldsymbol{\Gamma}_{s-1}]_k)$, which are the refined codewords of $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}(s-1, [\boldsymbol{\Gamma}_{s-1}]_k)$. Therefore, it requires totally $2K$ times of beam training to test all K UEs. To reduce the training overhead, we consider the following two cases:

1) If $[\boldsymbol{\Gamma}_{s-1}]_i = [\boldsymbol{\Gamma}_{s-1}]_q$ $(i \neq q, i, q \in \mathcal{K})$, which means that the ith UE and the qth UE share the same AoD at the $(s-1)$th layer of $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}$, we can perform beam training for them simultaneously because they will have the same refined codewords at the sth layer of $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}$.

2) If $[\mathbf{\Gamma}_{s-1}]_i \neq [\mathbf{\Gamma}_{s-1}]_q$ $(i \neq q, i, q \in \mathcal{K})$, which means that the ith UE and the qth UE have different AoD at the $(s-1)$th layer of $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}$, the ith UE cannot receive the signal transmitted from the beam coverage of $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}(s-1, [\mathbf{\Gamma}_{s-1}]_q)$ because the AoD of the ith UE is located in the beam coverage of $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}(s-1, [\mathbf{\Gamma}_{s-1}]_i)$. Therefore, we can distinguish different UEs based on their different AoDs. We will show subsequently that the BS can also simultaneously perform the beam training for these two UEs.

Based on the above discussion, in either case, there are at most K different integers in $\mathbf{\Gamma}_{s-1}$, i.e., $K' \leq K$, which correspond to K' different codewords at the $(s-1)$th layer and $2K'$ refined codewords at the sth layer of $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}$. In the proposed simultaneous multiuser hierarchical beam training scheme, we divide these $2K'$ refined codewords into two groups. The union of beam coverage of K' codewords in the first and second group can be expressed respectively as

$$\begin{cases} B_{s,1} = \bigcup_m D_{s,m}, \text{ if } \frac{m+1}{2} \in \mathbf{\Gamma}_{s-1}, \ m = 1, 2, \cdots, 2^s, \\ B_{s,2} = \bigcup_m D_{s,m}, \text{ if } \frac{m}{2} \in \mathbf{\Gamma}_{s-1}, \ m = 1, 2, \cdots, 2^s \end{cases} \tag{10}$$

where $D_{s,m} = [-1+(m-1)/2^{s-1}, -1+m/2^{s-1}]$ is the beam coverage of $\boldsymbol{\mathcal{V}}_{\mathrm{BS}}(s, m)$. It is seen that the beam coverage of $B_{s,1}$ and $B_{s,2}$ is intermittent, which is the motivation of our work to design multi-mainlobe codewords. We can obtain $\mathbf{\Psi}_{s,1}$ and $\mathbf{\Psi}_{s,2}$ based on $B_{s,1}$ and $B_{s,2}$ respectively via

$$\mathbf{\Psi}_{s,m} = \{n | R_n \subset B_{s,m}, \ n = 1, 2, \cdots, N_{\mathrm{BS}}\} \tag{11}$$

for $m = 1, 2$. Given $\mathbf{\Psi}_{s,1}$ and $\mathbf{\Psi}_{s,2}$, we can design $\boldsymbol{\mathcal{C}}(s, 1)$ and $\boldsymbol{\mathcal{C}}(s, 2)$, respectively via (6). Therefore, by using (10), (11) and (6), we can design $\boldsymbol{\mathcal{C}}(s, 1)$ and $\boldsymbol{\mathcal{C}}(s, 2)$ based on the beam training results of the $(s-1)$th layer, i.e., $\mathbf{\Gamma}_{s-1}$. Note that either $\boldsymbol{\mathcal{C}}(s, 1)$ or $\boldsymbol{\mathcal{C}}(s, 2)$ is a multi-mainlobe codeword, where each mainlobe covers a spatial region that one or more users are probably in.

At the sth layer of $\boldsymbol{\mathcal{C}}$, for $s = 2, 3, \ldots, S-1$, the BS sequentially transmits $\boldsymbol{\mathcal{C}}(s, 1)$ and $\boldsymbol{\mathcal{C}}(s, 2)$ to all K UEs. Note that there are only two codewords at each intermediate layer, which only requires two times of simultaneous beam training for all the UEs no matter how many UEs the BS serves. Then each UE compares the received signal power of $\boldsymbol{\mathcal{C}}(s, 1)$ and $\boldsymbol{\mathcal{C}}(s, 2)$ and individually feeds back the index of the larger codeword to the BS. We define $\mathbf{\Phi}_s$ as a vector of length of K to keep the indices fed back by all K UEs, where $[\mathbf{\Phi}_s]_k$ is the index fed back by the kth UE. In fact, $[\mathbf{\Phi}_s]_k \in \{1, 2\}$. Then we can obtain $\mathbf{\Gamma}_s$ by

$$[\mathbf{\Gamma}_s]_k = 2\big([\mathbf{\Gamma}_{s-1}]_k - 1\big) + [\mathbf{\Phi}_s]_k, \tag{12}$$

which can be used to determine $B_{s+1,1}$ and $B_{s+1,2}$ via (10) for $k = 1, 2, \cdots, K$. We iteratively perform these steps until arriving at the bottom layer of $\boldsymbol{\mathcal{C}}$.

At the bottom layer of $\boldsymbol{\mathcal{C}}$, unlike the downlink beam training in the top and intermediate layers, we perform the uplink beam training so that each entry of the effective channel matrix in (15) can be obtained. During the uplink beam training between the kth UE and the BS, the latter sequentially uses $\boldsymbol{f}_{\mathrm{c}}^{2[\mathbf{\Gamma}_{S-1}]_k - 1}$ and $\boldsymbol{f}_{\mathrm{c}}^{2[\mathbf{\Gamma}_{S-1}]_k}$ to receive the signal and selects one with the larger received signal power. In this way, the best BS codeword in $\mathcal{F}$ for the kth UE can be determined. Note that the BS has N_{RF} RF chains, which implies that the BS can use multiple RF chains for parallel signal receiving to improve the efficiency [9]. Therefore, totally $2K$ times of beam training are required at the bottom layer of $\boldsymbol{\mathcal{C}}$. Similar to (12), we can obtain $\mathbf{\Gamma}_S$ as

$$[\mathbf{\Gamma}_S]_k = 2\big([\mathbf{\Gamma}_{S-1}]_k - 1\big) + [\mathbf{\Phi}_S]_k \tag{13}$$

for $k = 1, 2, \cdots, K$.

Finally, the kth column of the designed analog precoder $\widehat{\boldsymbol{F}}_{\mathrm{RF}}$, denoted as $\widehat{\boldsymbol{f}}_k$, can be obtained via

$$\widehat{\boldsymbol{f}}_k = \boldsymbol{f}_{\mathrm{c}}^{[\mathbf{\Gamma}_S]_k}. \tag{14}$$

The detailed steps of the proposed simultaneous multiuser hierarchical beam training are summarized in **Algorithm** 1.

Algorithm 1 Simultaneous Multiuser Hierarchical Beam Training

1: **Input:** N_{BS}, N_{UE} and K.
2: Obtain $\boldsymbol{\mathcal{C}}(1, 1)$ and $\boldsymbol{\mathcal{C}}(1, 2)$ via (9) and (6).
3: Obtain $\mathbf{\Gamma}_1$ by the top layer beam training.
4: Set $S = \log_2 N_{\mathrm{BS}}$.
5: **for** $s = 2, 3, \ldots, S-1$ **do**
6: Obtain $B_{s,1}$ and $B_{s,2}$ via (10).
7: Obtain $\mathbf{\Psi}_{s,1}$ and $\mathbf{\Psi}_{s,2}$ via (11).
8: Generate $\boldsymbol{\mathcal{C}}(s, 1)$ and $\boldsymbol{\mathcal{C}}(s, 2)$ via (6).
9: Obtain $\mathbf{\Gamma}_s$ via (12).
10: **end for**
11: Obtain $\mathbf{\Gamma}_S$ via (13).
12: Obtain $\widehat{\boldsymbol{f}}_k$ via (14).
13: **Output:** $\{\widehat{\boldsymbol{f}}_k, k = 1, 2, \ldots, K\}$.

When designing codewords at the top and intermediate layers of $\boldsymbol{\mathcal{C}}$, we first obtain the ideal codewords by the weighted summation of channel steering vectors as (6) and then obtain the practical codewords regarding the number of RF chains and the resolution of phase shifters to approximate the ideal codewords based on the method in [15].

B. Design Example

Consider a multiuser mmWave massive MIMO system with $N_{\mathrm{BS}} = 128$, $N_{\mathrm{UE}} = 16$ and $K = 4$. As shown in Fig. 2, we illustrate the beam gain of different codewords in $\boldsymbol{\mathcal{C}}$. To improve the readability of our scheme, each layer in Fig. 2 corresponds to that in Fig. 1. At the top layer of $\boldsymbol{\mathcal{C}}$, the BS sequentially transmits $\boldsymbol{\mathcal{C}}(1, 1)$ and $\boldsymbol{\mathcal{C}}(1, 2)$. The union of beam coverage of $\boldsymbol{\mathcal{C}}(1, 1)$ and $\boldsymbol{\mathcal{C}}(1, 2)$ is the full space of $[-1, 1]$ since the BS has no knowledge of the UEs at the beginning. After the top layer beam training, we suppose that the indices fed back from the four UEs form a set $\mathbf{\Gamma}_1 = \{1, 1, 2, 2\}$, which indicates that the channel AoDs of the first and second UEs happen to locate in the same beam coverage of $\boldsymbol{\mathcal{C}}(1, 1)$, and the channel AoDs of the third and fourth UEs happen to locate in the same beam coverage of $\boldsymbol{\mathcal{C}}(1, 2)$. Based on $\mathbf{\Gamma}_1$, we can obtain $\mathbf{\Psi}_{2,1} = \{1, 2 \cdots, 32, 65, 66, \cdots, 96\}$ and

$\boldsymbol{\Psi}_{2,2} = \{33, 34 \cdots, 64, 97, 98, \cdots, 128\}$ via (11). Based on $\boldsymbol{\Psi}_{2,1}$ and $\boldsymbol{\Psi}_{2,2}$, we can design multi-mainlobe codewords $\boldsymbol{\mathcal{C}}(2,1)$ and $\boldsymbol{\mathcal{C}}(2,2)$ via (6). Note that both $\boldsymbol{\mathcal{C}}(2,1)$ and $\boldsymbol{\mathcal{C}}(2,2)$ have two mainlobes. During the second layer beam training, the BS sequentially transmits $\boldsymbol{\mathcal{C}}(2,1)$ and $\boldsymbol{\mathcal{C}}(2,2)$. Suppose the indices fed back from the four UEs form a set $\boldsymbol{\Phi}_2 = \{1,2,2,2\}$, where each entry denotes the codeword index $\{i|\boldsymbol{\mathcal{C}}(2,i), i = 1,2\}$ with the larger received signal power at the UEs. Then we can obtain $\boldsymbol{\Gamma}_2 = \{1,2,4,4\}$ via (12). Based on $\boldsymbol{\Gamma}_2$, we can design $\boldsymbol{\Psi}_{3,1} = \{1,2,\cdots,16,33,34,\cdots,48,97,98\cdots,112\}$ and $\boldsymbol{\Psi}_{3,2} = \{17,18,\cdots,32,49,50,\cdots,64,113,114,\cdots,128\}$ via (11). Based on $\boldsymbol{\Psi}_{3,1}$ and $\boldsymbol{\Psi}_{3,2}$, we can design multi-mainlobe codewords $\boldsymbol{\mathcal{C}}(3,1)$ and $\boldsymbol{\mathcal{C}}(3,2)$ via (6). Note that both $\boldsymbol{\mathcal{C}}(3,1)$ and $\boldsymbol{\mathcal{C}}(3,2)$ have three mainlobes. We repeat the procedures until arriving at the bottom layer of $\boldsymbol{\mathcal{C}}$.

C. Digital Precoding

Since the designed analog combiner, $\widehat{\boldsymbol{w}}_k$, for the kth UE can be obtained by the existing hierarchical beam training method, the details are omitted in this work due to the page limitation. Stacking $\{y_k, k = 1,2,\ldots,K\}$ in (1) together as $\boldsymbol{y} = [y_1, y_2, \ldots, y_K]^T$, we have $\boldsymbol{y} = \boldsymbol{H}_{\rm e}\boldsymbol{F}_{\rm BB}\boldsymbol{s}$, where

$$\boldsymbol{H}_{\rm e} = \begin{bmatrix} \widehat{\boldsymbol{w}}_1^H\boldsymbol{H}_1\widehat{\boldsymbol{f}}_1 & \widehat{\boldsymbol{w}}_1^H\boldsymbol{H}_1\widehat{\boldsymbol{f}}_2 & \cdots & \widehat{\boldsymbol{w}}_1^H\boldsymbol{H}_1\widehat{\boldsymbol{f}}_K \\ \widehat{\boldsymbol{w}}_2^H\boldsymbol{H}_2\widehat{\boldsymbol{f}}_1 & \widehat{\boldsymbol{w}}_2^H\boldsymbol{H}_2\widehat{\boldsymbol{f}}_2 & \cdots & \widehat{\boldsymbol{w}}_2^H\boldsymbol{H}_2\widehat{\boldsymbol{f}}_K \\ \vdots & \cdots & \ddots & \vdots \\ \widehat{\boldsymbol{w}}_K^H\boldsymbol{H}_K\widehat{\boldsymbol{f}}_1 & \widehat{\boldsymbol{w}}_K^H\boldsymbol{H}_K\widehat{\boldsymbol{f}}_2 & \cdots & \widehat{\boldsymbol{w}}_K^H\boldsymbol{H}_K\widehat{\boldsymbol{f}}_K \end{bmatrix} \quad (15)$$

is defined as the effective channel matrix. Note that each entry of $\boldsymbol{H}_{\rm e}$ can be obtained via the uplink beam training at the bottom layer of $\boldsymbol{\mathcal{C}}$. The designed digital precoder under the zero forcing (ZF) criterion can be expressed as $\widehat{\boldsymbol{F}}_{\rm BB} = \boldsymbol{H}_{\rm e}^H(\boldsymbol{H}_{\rm e}\boldsymbol{H}_{\rm e}^H)^{-1}$. Once the digital precoder is designed, we can figure out the average sum-rate over all K users [9].

D. Overhead Analysis

At the top layer of $\boldsymbol{\mathcal{C}}$, the BS sequentially transmits two codewords and each UE receives the signal with two codewords, which occupies 4 time slots. At the intermediate layers of $\boldsymbol{\mathcal{C}}$ from $s = 2$ to $s = \log_2 N_{\rm UE}$, the BS sequentially transmits two codewords and each UE receives signal with two codewords, which occupies totally $4(\log_2 N_{\rm UE} - 1)$ time slots. At the intermediate layers of $\boldsymbol{\mathcal{C}}$ from $s = \log_2 N_{\rm UE} + 1$ to $s = \log_2 N_{\rm BS} - 1$, where the UEs have reached the bottom layer of $\boldsymbol{\mathcal{V}}_{\rm UE}$, the BS transmits two codewords and each UE receives signal with a single codeword, which occupies totally $2(\log_2 N_{\rm BS} - \log_2 N_{\rm UE} - 1)$ time slots. At the bottom layer of $\boldsymbol{\mathcal{C}}$, the BS uses two codewords to receive the signal from each UE, which results in totally $2K$ time slots. In all, our proposed scheme needs totally $(2K + 2\log_2 N_{\rm UE}N_{\rm BS} - 2)$ time slots.

As shown in Table I, we compare the training overhead of different schemes. For example, if $N_{\rm BS} = 128$, $N_{\rm UE} = 16$, $K = 8$, three different schemes, including our scheme, the scheme in [8] and the TDMA hierarchical beam training scheme, require 36, 2048 and 176 time slots, respectively. Compared to the latter two schemes, our scheme can reduce the training overhead by 98.2% and 79.6%, respectively.

TABLE I
COMPARISONS OF OVERHEAD FOR DIFFERENT SCHEMES.

Schemes	Training overhead	Feedback overhead
Our scheme	$2(K + \log_2 N_{\rm UE}N_{\rm BS} - 1)$	$K(\log_2 N_{\rm BS} - 1)$
Scheme in [8]	$N_{\rm BS}N_{\rm UE}$	K
TDMA hierarchical beam training	$2K(\log_2 N_{\rm BS} + \log_2 N_{\rm UE})$	$K\log_2 N_{\rm BS}$

We also compare the feedback overhead from the UEs to the BS. For the scheme in [8], each UE needs only one time of feedback after finishing the beam scanning, which results in totally K times of feedback. Since we do need feedback at the bottom layer in our scheme, the feedback times of our scheme is K less than that of the TDMA hierarchical beam training.

IV. Simulation Results

We consider a multiuser mmWave massive MIMO system, where the BS equipped with $N_{\rm BS} = 128$ antennas serves $K = 8$ UEs. Each UE is equipped with $N_{\rm UE} = 16$ antennas. The mmWave MIMO channel matrix is setup with $L_k = 3$ channel paths with one line-of-sight (LOS) path and two non-line-of-sight (NLOS) paths, where the channel gain of the LOS path obeys $\lambda_1 \sim \mathcal{CN}(0,1)$ and the NLOS paths obey $\lambda_2 \sim \mathcal{CN}(0, 0.01)$ and $\lambda_3 \sim \mathcal{CN}(0, 0.01)$. Both the channel AoA $\theta_{\rm UE}^l$ and channel AoD $\theta_{\rm BS}^l$ of the lth channel path obey the uniform distribution between $[-1, 1]$.

As shown in Fig. 3, we compare the success rate of beam training for different schemes. The success rate is defined as follows. If the LOS path of the kth UE is correctly identified after beam training, we regard that the beam training of the kth UE is successful; otherwise, we regard that the beam training of the kth UE is failed. The ratio of the number of successful beam training over the total number of beam training is defined as the success rate. With totally K UEs served by the BS, the success rate shown in Fig. 3 is averaged over all K UEs. To make fair comparisons, we first extend the scheme in [12] from the partially connected structure to the fully connected structure. From the figure, the scheme in [8] can achieve better performance than the other three schemes, which lies in the fact that the beam scanning inherently performs better than the hierarchical beam training. Note that in order to clearly present our idea in this work, we start the hierarchical beam training from the top layer of the hierarchical codebook for both the BS and the UEs. In fact, we may start the hierarchical beam training from the lower layer of the hierarchical codebook to enlarge the beam gain of the codewords, which can improve the beam training performance. Since the training overhead of the scheme in [8] is much higher than the other schemes, our interest is indeed the comparisons of the three hierarchical beam training schemes. From the figure, the performance of

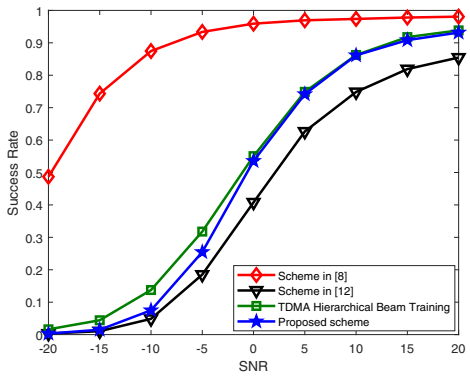

Fig. 3. Comparisons of success rate of beam training for different schemes.

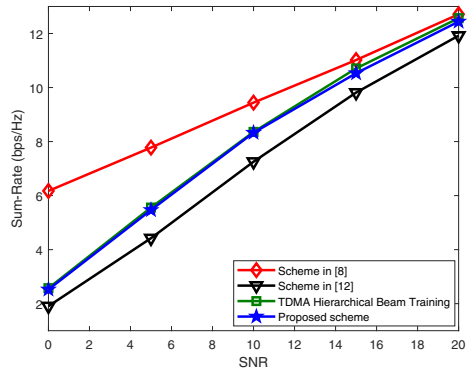

Fig. 4. Comparisons of the averaged sum-rate for different schemes.

our scheme is better than the scheme in [12] and almost the same as that of TDMA hierarchical beam training. At low signal-to-noise-ratio (SNR) region, e.g., SNR $= -5$dB, our scheme performs slightly worse than the TDMA hierarchical beam training, which is caused by the lower signal power averaged over all UEs in the simultaneous multiuser beam training of our scheme. However, the training overhead of our scheme is much smaller than the TDMA hierarchical beam training, i.e., 176 versus 36 with 79.6% reduction.

Fig. 4 compares the average sum-rate for different schemes. From the figure the curves of our scheme and the TDMA hierarchical beam training scheme are almost overlapped. Moreover, as the SNR increases, the performance gap between our scheme and the scheme in [8] gets smaller. At SNR $= 15$dB, the gap is no more than 0.5 bps/Hz. In summary, our scheme can approach the performance of the beam scanning with considerable reduction in training overhead.

V. Conclusion

In this paper, we have designed the hierarchical codebook in an adaptive manner, where the codewords in the current layer of the hierarchical codebook are designed according to the beam training results of the previous layer. In particular, we have designed multi-mainlobe codewords to perform the beam training simultaneously for all the users, where each mainlobe covers a spatial region that one or more users are probably in. Except for the bottom layer of the designed adaptive hierarchical codebook, there are only two codewords at each layer, which only requires twice of simultaneous beam training for all the users no matter how many users the BS serves. Simulation results have shown that our scheme can approach the performance of the beam scanning with considerable reduction in training overhead. Future work will include the efficient design of multi-mainlobe codewords as well as the corresponding multiuser hierarchical beam training schemes.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61871119 and by the Natural Science Foundation of Jiangsu Province under Grant BK20161428.

References

[1] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Top. Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[2] L. Zhao, G. Geraci, T. Yang, D. W. K. Ng, and J. Yuan, "A tone-based AoA estimation and multiuser precoding for millimeter wave massive MIMO," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5209–5225, Dec. 2017.

[3] C. Lin and G. Y. Li, "Energy-efficient design of indoor mmWave and sub-THz systems with antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4660–4672, Jul. 2016.

[4] C. Lin, G. Y. Li, and L. Wang, "Subarray-based coordinated beamforming training for mmWave and sub-THz communications," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2115–2126, Sep. 2017.

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

[6] J. Song, J. Choi, and D. J. Love, "Common codebook millimeter wave beam design: Designing beams for both sounding and communication with uniform planar arrays," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1859–1872, Apr. 2017.

[7] A. Ali, N. Gonzlez-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1038–1052, Feb. 2018.

[8] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[9] X. Sun, C. Qi, and G. Y. Li, "Beam training and allocation for multiuser millimeter wave massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1041–1053, Feb. 2019.

[10] Z. Xiao, T. He, P. Xia, and X. G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May. 2016.

[11] K. Chen and C. Qi, "Beam training based on dynamic hierarchical codebook for millimeter wave massive MIMO," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 132–135, Jan. 2019.

[12] R. Zhang, H. Zhang, W. Xu, and C. Zhao, "A codebook based simultaneous beam training for mmwave multi-user MIMO systems with split structures," in *2018 IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018, pp. 1–6.

[13] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[14] S. Noh, M. D. Zoltowski, and D. J. Love, "Multi-resolution codebook and adaptive beamforming sequence design for millimeter wave beam alignment," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5689–5701, Sep. 2017.

[15] K. Chen and C. Qi, "Beam design with quantized phase shifters for millimeter wave massive MIMO," in *2018 IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018, pp. 1–7.