Sparse Channel Estimation and Hybrid Precoding Using Deep Learning for Millimeter Wave Massive MIMO

Wenyan Ma, *Student Member, IEEE*, Chenhao Qi, *Senior Member, IEEE*, Zaichen Zhang, *Senior Member, IEEE*, and Julian Cheng, *Senior Member, IEEE*

***Abstract*— Channel estimation and hybrid precoding are considered for multi-user millimeter wave massive multi-input multi-output system. A deep learning compressed sensing (DLCS) channel estimation scheme is proposed. The channel estimation neural network for the DLCS scheme is trained offline using simulated environments to predict the beamspace channel amplitude. Then the channel is reconstructed based on the obtained indices of dominant beamspace channel entries. A deep learning quantized phase (DLQP) hybrid precoder design method is developed after channel estimation. The training hybrid precoding neural network for the DLQP method is obtained offline considering the approximate phase quantization. Then the deployment hybrid precoding neural network (DHPNN) is obtained by replacing the approximate phase quantization with ideal phase quantization and the output of the DHPNN is the analog precoding vector. Finally, the analog precoding matrix is obtained by stacking the analog precoding vectors and the digital precoding matrix is calculated by zero-forcing. Simulation results demonstrate that the DLCS channel estimation scheme outperforms the existing schemes in terms of the normalized mean-squared error and the spectral efficiency, while the DLQP hybrid precoder design method has better spectral efficiency performance than other methods with low phase shifter resolution.**

***Index Terms*— Channel estimation, deep learning, hybrid precoding, massive MIMO, mmWave communications.**

I. Introduction

DUE to the rich bandwidth resources of the millimeter wave (mmWave), mmWave communication has attracted broad attention and become an important technology in future wireless communication systems [1], [2]. When operating at high frequency, the mmWave signal experiences high path loss. Fortunately, this challenge can be overcome by directional beamforming with a massive multi-input multi-output (MIMO) antenna array. Since mmWave bands have short wavelengths, large antenna arrays can be packed into small form factors [3].

Due to the large antenna arrays of mmWave communications, channel estimation requires a large number of time slots as overhead. Note that the mmWave channels have sparsity feature in the beamspace domain with hybrid precoding [4]. Although the beamspace is typically addressed in the mmWave lens antenna arrays, we can also obtain the beamspace channel with hybrid precoding by introducing a dictionary matrix consisting of column steering vectors. Several channel estimation schemes have been proposed to explore the beamspace channel sparsity. For examples, a distributed grid matching pursuit (DGMP) channel estimation scheme was proposed [4], where the dominant entries of the line-of-sight (LOS) channel path were detected and updated iteratively; an orthogonal matching pursuit (OMP) channel estimation scheme was proposed to detect the dominant entries of multiple channel paths [5]; a simultaneous weighted orthogonal matching pursuit (SWOMP) channel estimation scheme was proposed [6], where the frequency-selective mmWave channels were considered based on the OMP method. However, these compressed sensing (CS) channel estimation schemes estimate the dominant beamspace channel entries sequentially and greedily, which cannot guarantee the global optimality [7].

After the channel estimation of mmWave communications, hybrid precoding consisting of analog precoding and digital precoding is usually adopted. Analog precoding aims to form directional beams using phase shifter network, while digital precoding is designed to mitigate interference of multiple data streams. Several hybrid precoding methods have been proposed for single-user multi-stream mmWave communication systems. For examples, a hybrid precoding algorithm was proposed [8], where the analog precoding problem was formulated as a sparse reconstruction problem and the OMP method was adopted; to avoid the greed of the OMP method, the alternating minimization method was used [9], where the hybrid precoding problem was designed as a matrix decomposition problem and the analog precoder and digital precoder were optimized alternately; to reduce the computational complexity of the

Manuscript received December 7, 2019; revised January 17, 2020; accepted February 9, 2020. Date of publication February 17, 2020; date of current version May 15, 2020. This work was supported in part by National Natural Science Foundation of China under Grant 61871119 and 61960206005, by Natural Science Foundation of Jiangsu Province under Grant BK20161428, by National Key Research and Development Plan Project under Grant 2018YFB1801101, and by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this article and approving it for publication was Z. Qin. *(Corresponding author: Chenhao Qi.)*

Wenyan Ma, Chenhao Qi, and Zaichen Zhang are with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: qch@seu.edu.cn; zczhang@seu.edu.cn).

Julian Cheng is with the School of Engineering, The University of British Columbia, Kelowna, BC V1V 1V7, Canada (e-mail: julian.cheng@ubc.ca).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2020.2974457

alternating minimization method, the hierarchical codebook was used to obtain multiple beams and then form the analog precoding [10], [11].

In the multi-user multi-stream mmWave communication systems, the base station (BS) transmits multiple data steams to serve all users simultaneously. To improve the spectral efficiency, the beamsteering codebook based on steering vectors was used to formulate the analog precoder vectors and the digital precoder was designed [12]. To consider the hardware constraint of the limited phase shifter resolution, beam allocation for multiple users was considered [13], where the discrete fourier transformation (DFT) codebook was adopted for analog precoding and the phase shifter resolution must be proportional to the number of antennas. To remove the constraint that the phase shifter resolution was related to the number of antennas, a quantized angle linear search (QALS) precoding scheme was proposed [14], where the angular domain was quantized according to the limited resolution of phase shifters and a linear search method was used to obtain the optimal analog beamforming vectors aligning with the dominant channel paths. However, these hybrid precoding schemes design the analog precoder using the steering vectors of quantized angles, which is heavily constrained by the resolution of phase shifters. When the mmWave system is equipped with low resolution phase shifters, there is a small number of available steering vectors of quantized angles. Since the angles of arrival (AoAs) of channel paths are randomly distributed, it cannot guarantee that the precoding based on these limited steering vectors can always have the high beamforming gain. Therefore these hybrid precoding schemes may have unsatisfactory spectral efficiency performance if none of these limited steering vectors can be aligned with the AoAs well.

Recently, the application of deep learning to mmWave communications has received much attention owing to the capability of deep learning to solve complicated nonlinear problems [15]–[17]. For examples, a machine learning based beam prediction scheme was proposed [18], where the machine learning tools and situational awareness were combined to learn the beam information (power, optimal beam index, etc) from past observations; a learned denoising based approximate message passing network was proposed to estimate the mmWave communication system with lens antenna array [19], where the noise term was detected and removed to estimate the channel. However, channel estimation for mmWave massive MIMO systems with hybrid precoding was not considered [19]. Besides, a deep learning based beamforming design method was proposed [20], where a beamforming neural network was trained to learn how to optimize the beamformer for maximizing the spectral efficiency; a deep reinforcement learning hybrid precoding method was proposed [21]. However, both these two deep learning hybrid precoder design methods neglect the constraint of limit resolution of phase shifters.

In this paper, we investigate sparse channel estimation and hybrid precoding considering the limited resolution of phase shifters for multi-user mmWave massive MIMO systems. The paper has the following two main contributions.

1) We propose a deep learning compressed sensing (DLCS) channel estimation scheme for the multi-user mmWave massive MIMO systems. The DLCS scheme consists of beamspace channel amplitude estimation and channel reconstruction. In the offline training stage, we train the channel estimation neural network (CENN) using the simulated environment based on the mmWave channel model. Then in the online deployment stage, the correlation between the received signal vectors and the measurement matrix is fed into the trained CENN to predict the beamspace channel amplitude. Afterwards, the indices of dominant entries of beamspace channel are obtained, based on which the channel can be reconstructed. Unlike the existing work that estimates the dominant beamspace channel entries sequentially [4]–[6], we estimate dominant entries simultaneously, which will be shown to have better channel estimation performance.

2) We propose a deep learning quantized phase (DLQP) hybrid precoding method for the multi-user mmWave massive MIMO systems. In the DLQP method, we first design the analog precoder and then the digital precoder. In the offline training stage, we obtain the training hybrid precoding neural network (THPNN) using the estimated channel vector and real channel vector of each user, where the approximate phase quantization is considered. Then in the online deployment stage, we obtain the deployment hybrid precoding neural network (DHPNN) by replacing the approximate phase quantization in the THPNN with ideal phase quantization, where the estimated channel vector of each user is fed into the DHPNN to obtain the analog precoding vector. Afterwards, the analog precoding matrix is obtained by stacking the analog precoding vectors of all users, based on which the digital precoding matrix can be calculated by zero-forcing (ZF).

The rest of the paper is organized as follows. In Section II, we introduce the system model and formulate the problem of channel estimation for the multi-user mmWave massive MIMO systems with hybrid precoding. In Sections III, we propose the DLCS channel estimation scheme. In Section IV, we develop the DLQP hybrid precoder design method. The simulation results are provided in Section V. Finally, Section VI concludes the paper.

We use the following notations. Symbols for vectors (lower case) and matrices (upper case) are in boldface. $(\cdot)^T$, $(\cdot)^*$, $(\cdot)^H$, and $(\cdot)^{-1}$ denote the transpose, conjugate, conjugate transpose (Hermitian), and inverse, respectively. We use $\boldsymbol{I}_K$ to represent identity matrix of order K. The set of $P \times Q$ complex-valued matrices and real-valued matrices are denoted by $\mathbb{C}^{P\times Q}$ and $\mathbb{R}^{P\times Q}$, respectively. We use $\mathbb{E}\{\cdot\}$ to represent expectation. The l_2-norm of a vector and Frobenius norm of a matrix are denoted by $\|\cdot\|_2$ and $\|\cdot\|_F$, respectively. We use $\boldsymbol{a}[p]$ to denote the pth entry of $\boldsymbol{a}$. Complex Gaussian distribution is denoted by $\mathcal{CN}$. We use $|\cdot|$ to denote the absolute value. $\mathrm{Im}(\boldsymbol{a})$ and $\mathrm{Re}(\boldsymbol{a})$ denote the imaginary and real parts of $\boldsymbol{a}$, respectively.

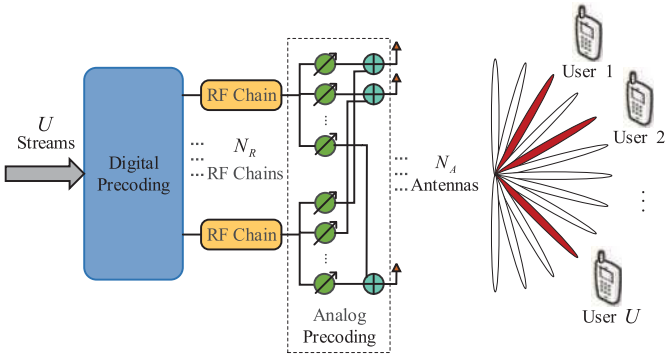


Fig. 1. Block diagram of downlink transmission in the multi-user mmWave massive MIMO system.

II. System Model and Problem Formulation

We first introduce the system model of multi-user mmWave massive MIMO. Then the channel estimation problem is formulated as a CS problem to estimate the sparse channel in the beamspace.

A. System Model

We consider a downlink multi-user mmWave massive MIMO communication system that comprises a BS and U users with single antenna, as shown in Fig. 1. The BS is equipped with a uniform linear array (ULA) [1]. Note that the present method can be generalized to other array structures. Let N_A and N_R denote the numbers of antennas and RF chains at the BS, respectively. Hybrid precoding is typically adopted, where the number of antennas is much larger than that of RF chains, i.e., $N_A \gg N_R$ [2]. We consider the orthogonal multiple access, where the number of active users simultaneously connected with the BS is no larger than the number of RF chains, i.e., $U \leq N_R$ [10]. If $U < N_R$, the BS will only turn on U RF chains to serve the U users simultaneously and turn off $N_R - U$ RF chains, which will save the power consumed at the BS.

For downlink transmission, the BS performs hybrid precoding, which consists of baseband digital precoding and RF analog precoding [13]. The received signal of all U users, denoted by $\boldsymbol{y}^{\mathrm{dl}} \in \mathbb{C}^U$, can be represented as

$$\boldsymbol{y}^{\mathrm{dl}} = \boldsymbol{H}\boldsymbol{F}_R\boldsymbol{F}_B\boldsymbol{s} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{F}_R \in \mathbb{C}^{N_A \times U}$ and $\boldsymbol{F}_B \in \mathbb{C}^{U \times U}$ denote the analog precoder and digital precoder, respectively. To normalize the power of the hybrid precoder, we set $\|\boldsymbol{F}_R\boldsymbol{F}_B\|_F^2 = U$. We denote the signal vector by $\boldsymbol{s} \in \mathbb{C}^U$ satisfying $\mathrm{E}\{\boldsymbol{s}\boldsymbol{s}^H\} = \boldsymbol{I}_U$ and additive white Gaussian noise (AWGN) vector by $\boldsymbol{n} \in \mathbb{C}^U$ satisfying $\boldsymbol{n} \sim \mathcal{CN}(0, \sigma^2\boldsymbol{I}_U)$. The channel matrix for the BS and all users is denoted by

$$\boldsymbol{H} \triangleq [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_U]^T \in \mathbb{C}^{U \times N_A}. \tag{2}$$

There are different kinds of channel model in mmWave systems, such as the clustered mmWave channel model and the Saleh-Valenzuela mmWave channel model [2], [22]. We choose the Saleh-Valenzuela mmWave channel model in our paper. The channel vector $\boldsymbol{h}_u \in \mathbb{C}^{N_A}$ for the BS and the uth user is represented as

$$\boldsymbol{h}_u = \sqrt{\frac{N_A}{L_u}} \sum_{i=1}^{L_u} \boldsymbol{h}_{u,i} = \sqrt{\frac{N_A}{L_u}} \sum_{i=1}^{L_u} g_{u,i}\boldsymbol{\alpha}(N_A, \theta_{u,i}) \tag{3}$$

where the channel vector, number of multiple channel paths, and complex gain of the ith path are denoted by $\boldsymbol{h}_{u,i}$, L_u, and $g_{u,i}$, respectively. Typically $\boldsymbol{h}_u$ consists of one LOS path (the 1st channel path), and $L_u - 1$ non-line-of-sight (NLOS) paths (the ith channel path for $2 \leq i \leq L_u$). The steering vector $\boldsymbol{\alpha}(N, \theta)$ can be expressed as

$$\boldsymbol{\alpha}(N, \theta) = \frac{1}{\sqrt{N}}\left[1, e^{j\pi\theta}, \ldots, e^{j\pi\theta(N-1)}\right]^T. \tag{4}$$

Denote the AoA for the ith path of the uth user by $\vartheta_{u,i}$, which is uniformly distributed over $[-\pi, \pi)$ [4], [23]. Then we have $\theta_{u,i} \triangleq \sin\vartheta_{u,i}$ if the distance between adjacent two antennas at the BS is half-wave length [4].

B. Problem Formulation

To design $\boldsymbol{F}_B$ and $\boldsymbol{F}_R$ for downlink data transmission, $\boldsymbol{H}$ should be estimated. Based on channel reciprocity, the estimate of downlink channel can be obtained by employing uplink channel estimation to estimate $\boldsymbol{H}$. Note that the proposed DLCS channel estimation scheme can also be used for the downlink channel estimation. Since the BS usually has more computing power than each user in practice, we consider the uplink channel estimation where the neural network (NN) is trained and utilized for prediction at the BS. For uplink channel estimation, mutually orthogonal pilot sequences are transmitted by all users to distinguish different signals from different users for K times. Denote the pilot matrix consisted of the U mutually orthogonal pilot sequences from U users by $\boldsymbol{P} \in \mathbb{C}^{U \times U}$. For the uplink pilot transmission, we use K different analog precoding matrices and digital precoding matrices, denoted by $\boldsymbol{F}_R^k \in \mathbb{C}^{N_A \times N_R}$ and $\boldsymbol{F}_B^k \in \mathbb{C}^{N_R \times N_R}$, respectively, for $k = 1, 2, \ldots, K$. The pilot sequences received at the BS for the kth sending are given by

$$\boldsymbol{Y}_k^{\mathrm{ul}} = (\boldsymbol{F}_R^k\boldsymbol{F}_B^k)^T\boldsymbol{H}^T\boldsymbol{P} + (\boldsymbol{F}_R^k\boldsymbol{F}_B^k)^T\boldsymbol{N}_k \tag{5}$$

where the AWGN matrix for the kth transmission is denoted by $\boldsymbol{N}_k$. Each entry of $\boldsymbol{N}_k$ obeys $\mathcal{CN}(0, \sigma^2)$. Based on the orthogonality of U mutually orthogonal pilot sequences, i.e., $\boldsymbol{P}\boldsymbol{P}^H = \boldsymbol{I}_U$, we multiply $\boldsymbol{Y}_k^{\mathrm{ul}}$ by $\boldsymbol{P}^H$ and obtain

$$\boldsymbol{R}_k \triangleq \boldsymbol{Y}_k^{\mathrm{ul}}\boldsymbol{P}^H = (\boldsymbol{F}^k)^T\boldsymbol{H}^T + \widetilde{\boldsymbol{N}}_k \tag{6}$$

where

$$\begin{aligned} \boldsymbol{F}^k &\triangleq \boldsymbol{F}_R^k\boldsymbol{F}_B^k \in \mathbb{C}^{N_A \times N_R}, \\ \widetilde{\boldsymbol{N}}_k &\triangleq (\boldsymbol{F}_R^k\boldsymbol{F}_B^k)^T\boldsymbol{N}_k\boldsymbol{P}^H \in \mathbb{C}^{N_R \times U}. \end{aligned} \tag{7}$$

After each user repeatedly transmits orthogonal pilot sequences K times, $\boldsymbol{R}_k$ for $k = 1, 2, \ldots, K$ can be stacked as

$$\boldsymbol{R} = [\boldsymbol{R}_1^T, \ldots, \boldsymbol{R}_K^T]^T = \boldsymbol{F}^T\boldsymbol{H}^T + \widetilde{\boldsymbol{N}} \tag{8}$$

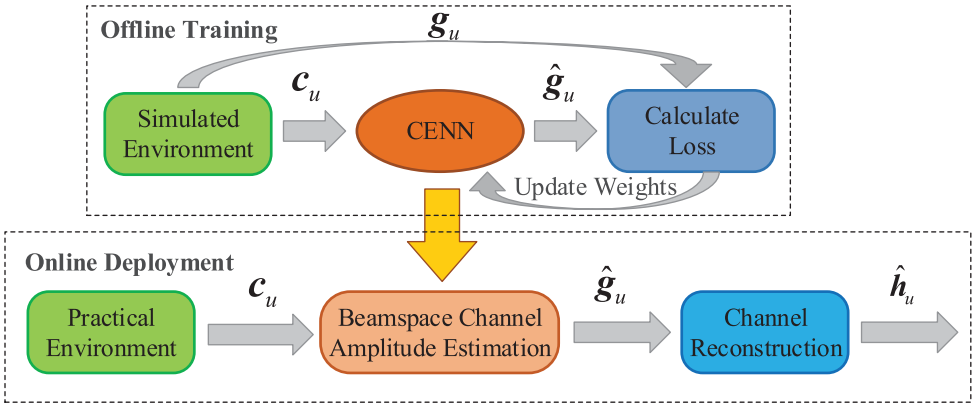


Fig. 2. Block diagram of the DLCS channel estimation scheme: offline training and online deployment.

where

$$\begin{aligned} \boldsymbol{F} &\triangleq [\boldsymbol{F}^1, \ldots, \boldsymbol{F}^K] \in \mathbb{C}^{N_A \times N_R K}, \\ \widetilde{\boldsymbol{N}} &\triangleq [\widetilde{\boldsymbol{N}}_1^T, \ldots, \widetilde{\boldsymbol{N}}_K^T]^T \in \mathbb{C}^{N_R K \times U}. \end{aligned} \tag{9}$$

Note that $N_A > N_R K$ since $N_A \gg N_R$ and we need a small number of time slots for channel training. Denote the uth column of $\boldsymbol{R}$ by $\boldsymbol{r}_u$ for $u = 1, 2, \ldots, U$. Then $\boldsymbol{r}_u$ can be represented as

$$\boldsymbol{r}_u = \boldsymbol{F}^T \boldsymbol{h}_u + \widetilde{\boldsymbol{n}}_u \tag{10}$$

where $\widetilde{\boldsymbol{n}}_u$ is the uth column of $\widetilde{\boldsymbol{N}}$.

Note that the mmWave channels have sparsity feature in the beamspace domain [4], [6]. We define

$$\boldsymbol{h}_u^b = \boldsymbol{A}\boldsymbol{h}_u \tag{11}$$

as a beamspace channel vector where $\boldsymbol{A} \in \mathbb{C}^{G \times N_A}$ is the dictionary matrix consisted of G column vectors $\boldsymbol{\alpha}(N_A, \phi_t)$, with $\phi_t \triangleq -1 + 2(t-1)/G$ representing the tth point of the angle grid. Note that the range of AoAs is quantified into G grids for $t = 1, 2, \ldots, G$. Based on the fact that $\boldsymbol{A}^H \boldsymbol{A} = G\boldsymbol{I}_{N_A}/N_A$, eq. (11) can be further rewritten as

$$\boldsymbol{r}_u = \frac{N_A}{G} \boldsymbol{F}^T \boldsymbol{A}^H \boldsymbol{h}_u^b + \widetilde{\boldsymbol{n}}_u. \tag{12}$$

Due to the sparse property of $\boldsymbol{h}_u^b$, eq. (12) is essentially a sparse recovery problem, which can be tackled by CS techniques [24]. Note that the sparsity of $\boldsymbol{h}_u^b$ can be impaired by channel power leakage caused by the limited beamspace resolution of $\boldsymbol{A}$ [25], which indicates that $\boldsymbol{h}_u^b$ is not perfectly sparse and many entries of $\boldsymbol{h}_u^b$ are small but nonzero. Sparse channel estimation schemes such as OMP and DGMP estimate the dominant beamspace channel entries in a sequential and greedy manner. However, they cannot guarantee the global optimality. Therefore, in the following we will propose a DLCS channel estimation scheme to estimate dominant beamspace channel entries simultaneously.

III. DLCS Channel Estimation

The proposed DLCS channel estimation scheme consists of beamspace channel amplitude estimation and channel reconstruction. The main idea of the DLCS scheme is to estimate first the beamspace channel amplitude using an offline-trained CENN, and then sort the estimated beamspace channel amplitude in descending order to select the indices of dominant entries, and finally reconstruct the channel according to the selected indices. The block diagram of the DLCS scheme is illustrated in Fig. 2. The detailed steps of the DLCS scheme are summarized in Algorithm 1.

A. Beamspace Channel Amplitude Estimation

We define

$$\boldsymbol{\Phi} \triangleq \frac{N_A}{G} \boldsymbol{F}^T \boldsymbol{A}^H \in \mathbb{C}^{N_R K \times G} \tag{13}$$

as the measurement matrix in (12). As shown in Algorithm 1, we feed $\boldsymbol{\Phi}$ and $\boldsymbol{r}_u$ to obtain the estimate of $\boldsymbol{h}_u$, denoted by $\hat{\boldsymbol{h}}_u$, for $u = 1, 2, \ldots, U$. The correlation vector between $\boldsymbol{\Phi}$ and $\boldsymbol{r}_u$, denoted by $\boldsymbol{c}_u \in \mathbb{C}^G$, can be expressed as

$$\boldsymbol{c}_u = \boldsymbol{\Phi}^H \boldsymbol{r}_u. \tag{14}$$

The sparse channel estimation schemes sequentially select the atoms, i.e., column vectors of $\boldsymbol{\Phi}$, which yield the greatest correlation with $\boldsymbol{r}_u$. However, such greedy algorithms cannot guarantee the global optimality, which motivates us to use the NN to estimate the atoms simultaneously instead of sequentially.

As shown in Fig. 2, the beamspace channel amplitude estimation has two stages: the offline training of the CENN and its online deployment. The CENN is first trained offline and then used as the kernel of the beamspace channel amplitude estimation. The input of the CENN is $\boldsymbol{c}_u$. The amplitude of $\boldsymbol{h}_u^b$ can be denoted by

$$\boldsymbol{g}_u \triangleq \left[\left| \boldsymbol{h}_u^b[1] \right|, \left| \boldsymbol{h}_u^b[2] \right|, \ldots, \left| \boldsymbol{h}_u^b[G] \right| \right]^T \in \mathbb{R}^G. \tag{15}$$

The output of the CENN is denoted by $\hat{\boldsymbol{g}}_u$ and is expected to be $\boldsymbol{g}_u$.

As illustrated in Fig. 3, the adopted CENN in this work consists of three hidden layers and a fully connected (FC) layer. Since the NN can only deal with the real number, the input of the CENN is a real-valued vector having $2G$ entries composed by the imaginary and real parts of $\boldsymbol{c}_u$. Each hidden layer

Algorithm 1 DLCS Channel Estimation

Input: $\boldsymbol{\Phi}$, $\boldsymbol{r}_u$, J.
Initialization: $\hat{\boldsymbol{h}}_u^b \leftarrow \mathbf{0}^G$.
(Beamspace Channel Amplitude Estimation)
Obtain $\boldsymbol{c}_u$ via (14).
Input $\boldsymbol{c}_u$ to the offline-trained CENN to get $\hat{\boldsymbol{g}}_u$.
(Channel Reconstruction)
Obtain $\boldsymbol{\Gamma}$ based on J dominant entries of $\hat{\boldsymbol{g}}_u$.
Compute $\hat{\boldsymbol{h}}_u^b[\boldsymbol{\Gamma}]$ via (17).
Obtain $\hat{\boldsymbol{h}}_u$ according to (18).
Output: $\hat{\boldsymbol{h}}_u$.

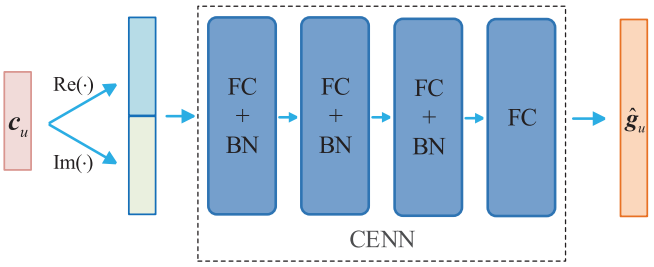


Fig. 3. Illustration of the CENN.

includes an FC layer and a batch normalization (BN) layer. The numbers of neurons in these three hidden layers are set as 1,024, 512, and 256. The activation function adopted in the FC layer is the ReLU function, which can be represented as $f_{\mathrm{Re}}(x) = \max(0, x)$.

During the offline training of the CENN, we generate the dataset of $\boldsymbol{c}_u$ and $\boldsymbol{g}_u$ based on the simulated mmWave channel environment. With the beamspace channel amplitude in (15) and the correlation of the received signals and the measurement matrix in (14), the training data of $\boldsymbol{c}_u$ and $\boldsymbol{g}_u$ can be obtained. In fact, the process to obtain $\boldsymbol{c}_u$ and $\boldsymbol{g}_u$ involves the following four steps: **i)** we randomly generate a channel vector based on the mmWave channel model in (3); **ii)** we obtain $\boldsymbol{g}_u$ based on (15); **iii)** we compute the received signal vector $\boldsymbol{r}_u$ based on (10); **iv)** we obtain the correlation vector $\boldsymbol{c}_u$ based on (14). We divide the data set into the training set and the validation set randomly, where the size of the training set is nine times the size of the validation set. The output of the CENN is $\hat{\boldsymbol{g}}_u$.

The training of the CENN aims to minimize the difference between $\hat{\boldsymbol{g}}_u$ and $\boldsymbol{g}_u$. The difference, typically named as the loss in machine learning, can be calculated in several ways. In this work, we calculate the loss by measuring the mean square error as [15]

$$f_{\mathrm{LossCS}}(\boldsymbol{g}_u, \hat{\boldsymbol{g}}_u) = \frac{1}{G}\sum_{n=1}^{G}\left(\boldsymbol{g}_u[n] - \hat{\boldsymbol{g}}_u[n]\right)^2. \tag{16}$$

We adopt the adaptive moment estimation (Adam) optimizer to train the CENN by TensorFlow. The CENN is trained for 1,000 epochs, where 50 mini-batches are utilized in each epoch. The learning rate is set to be a step function, which decreases with training epochs. The learning rate is initialized with the value of 0.01 and decreases 5-fold every 400 epochs.

During the online deployment of the CENN, we obtain the real measured $\boldsymbol{r}_u$ from practical mmWave channel environments. We compute $\boldsymbol{c}_u$ based on (14), which is then fed to the offline-trained CENN. The prediction of $\boldsymbol{g}_u$ by the CENN is $\hat{\boldsymbol{g}}_u$.

B. Channel Reconstruction

Note that the sparsity of $\boldsymbol{h}_u^b$ can be impaired by channel power leakage caused by the limited beamspace resolution of $\boldsymbol{A}$ [25], which indicates that $\boldsymbol{h}_u^b$ is not perfectly sparse and many entries of $\boldsymbol{h}_u^b$ have small but nonzero values. Denote the number of dominant entries of $\boldsymbol{g}_u$ by J, which is the beamspace channel sparse level. In the online deployment stage, we sort $\hat{\boldsymbol{g}}_u$ in descending order according to the absolute value of $\hat{\boldsymbol{g}}_u$. Then we obtain the indices of the first J entries, which are the prediction of the indices of J dominant entries in $\boldsymbol{g}_u$.

We denote the prediction of these J indices by $\boldsymbol{\Gamma} \in \mathbb{R}^J$. We further let $\hat{\boldsymbol{h}}_u^b$ denote an estimate of $\boldsymbol{h}_u^b$. We initialize $\hat{\boldsymbol{h}}_u^b$ to be zero. Then the J dominant entries of $\hat{\boldsymbol{h}}_u^b$ can be computed via the least squares (LS) estimation as

$$\hat{\boldsymbol{h}}_u^b[\boldsymbol{\Gamma}] = (\boldsymbol{\Phi}_{\Gamma}^H \boldsymbol{\Phi}_{\Gamma})^{-1} \boldsymbol{\Phi}_{\Gamma}^H \boldsymbol{r}_u \tag{17}$$

where $\boldsymbol{\Phi}_{\Gamma}$ consists of J columns of $\boldsymbol{\Phi}$ and the column indices are denoted by $\boldsymbol{\Gamma}$. Then using the result $\boldsymbol{A}^H\boldsymbol{A} = G\boldsymbol{I}_{N_A}/N_A$, we obtain the estimated channel vector for the uth user based on (11) as

$$\hat{\boldsymbol{h}}_u = \frac{N_A}{G}\boldsymbol{A}^H \hat{\boldsymbol{h}}_u^b. \tag{18}$$

It is shown that the proposed DLCS channel estimation scheme can avoid the greedy search that is commonly adopted by the existing sparse channel estimation schemes based on CS, since the DLCS scheme estimates dominant entries simultaneously instead of sequentially.

IV. DLQP Hybrid Precoder Design

Hybrid precoding is usually required for downlink data transmission after the channel estimation. In the proposed DLQP hybrid precoder design method, we first design the analog precoder and then the digital precoder. The main idea of the DLQP scheme is to first train the THPNN using the estimated channel vectors, where the approximate phase quantization is considered. Then the DHPNN is obtained by replacing the approximate phase quantization in THPNN with ideal phase quantization, where the estimated channel vectors are fed into the DHPNN to obtain the analog precoder vectors. Finally the analog precoding matrix is obtained by stacking the analog precoding vectors of all users and the digital precoding matrix can be calculated by ZF. The block diagram of the DLQP method is illustrated in Fig. 4. The detailed steps of the DLQP method is summarized in Algorithm 2.

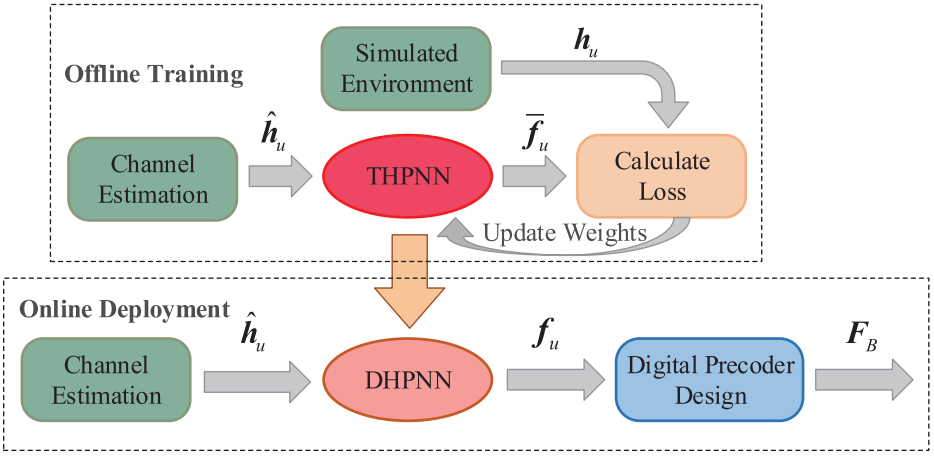


Fig. 4. Block diagram of the DLQP hybrid precoder design: offline training and online deployment.

Algorithm 2 DLQP Hybrid Precoder Design

Input: $\hat{\boldsymbol{h}}_u$.
(Analog Precoder Design)
Replace the AQ layer of the offline-trained THPNN by the IQ layer to obtain the DHPNN.
Input $\hat{\boldsymbol{h}}_u$ to the DHPNN to get $\boldsymbol{f}_u$.
Obtain $\boldsymbol{F}_R$ according to (27).
(Digital Precoder Design)
Compute $\hat{\boldsymbol{H}}$ based on (28).
Obtain $\boldsymbol{H}_{\text{eff}}$ according to (29).
Compute $\boldsymbol{F}_B$ via (30).
Normalize each column of $\boldsymbol{F}_B$ via (31).
Output: $\boldsymbol{F}_R$, $\boldsymbol{F}_B$.

A. Analog Precoder Design

Denote the analog precoder vector and approximate analog precoder vector by $\boldsymbol{f}_u \triangleq [f_{u,1}, f_{u,2}, \ldots, f_{u,N_A}]^T \in \mathbb{C}^{N_A}$ and $\bar{\boldsymbol{f}}_u \triangleq [\bar{f}_{u,1}, \bar{f}_{u,2}, \ldots, \bar{f}_{u,N_A}]^T \in \mathbb{C}^{N_A}$, respectively, for $u = 1, 2, \ldots, U$. As shown in Fig. 5, $\hat{\boldsymbol{h}}_u$ is fed to the THPNN to obtain $\bar{\boldsymbol{f}}_u$, while $\hat{\boldsymbol{h}}_u$ is fed to the DHPNN to obtain $\boldsymbol{f}_u$. Note that the difference between the THPNN and DHPNN is that we use approximate phase quantization in the THPNN so that the NN can be trained, while we use ideal phase quantization in the DHPNN to meet the practical constraint of limited phase shifter resolution.

We define B as the quantization bit number of the phase shifters used at the BS, where the RF phase is quantized into $Q \triangleq 2^B$ discrete values. Each entry of $\boldsymbol{f}_u$ is randomly drawn from the set $\{e^{j2\pi n/Q}, n = 1, 2, \ldots, Q\}$. The hybrid precoder design schemes based on beamsteering codebooks design the analog precoder vector as the steering vector of the LOS channel path [12]–[14]. However, such schemes require that $Q \geq N_A$ to obtain high beamforming gain, which will have unsatisfactory performance when $Q < N_A$. This requirement motivates us to use the NN to design the analog precoder when $Q < N_A$.

As shown in Fig. 4, the hybrid precoder design has two stages: the offline training of the THPNN and online deployment of the DHPNN, where the DHPNN is obtained based on the THPNN by replacing one layer of the THPNN. The THPNN is first trained offline and then the DHPNN is obtained, which is used as the kernel of the hybrid precoder design. The input of the THPNN and DHPNN is $\hat{\boldsymbol{h}}_u$. The outputs of the DHPNN and the THPNN are the analog precoder vector $\boldsymbol{f}_u$ and approximate analog precoder vector $\bar{\boldsymbol{f}}_u$, respectively.

As illustrated in Fig. 5, both the adopted THPNN and DHPNN consist of six layers, where five of them are shared. Since the NN can only deal with the real number, the input of the THPNN and DHPNN is a real-valued vector having $2N_A$ entries composed by the imaginary and real parts of $\hat{\boldsymbol{h}}_u$. Each of the first four layers consists of a convolutional (Conv) layer and a pooling (Pool) layer. The kernel size and strides of each Conv layer are set to be five and one, respectively. The number of filters of these three Conv layers are set as 16, 32, and 64, respectively. Both the pool size and strides of each Pool layer are set to be two. The activation function adopted in the first three layers is the ReLU function, while that adopted in the fourth layer is the Sigmoid function, which can be represented as $f_{\text{Sig}}(x) = \frac{1}{1+e^{-x}}$. Since the output of the FC and Pool layers can only be real number, we cannot directly obtain the complex-valued $\boldsymbol{f}_u$. Then the output of the fourth layer is the phase of analog precoder vector, which is denoted by

$$\boldsymbol{\phi} \triangleq [\phi_1, \phi_2, \ldots, \phi_{N_A}]^T \in \mathbb{R}^{N_A} \tag{19}$$

where $\phi_n \in [0, 2\pi)$, for $n = 1, 2, \ldots, N_A$. Since the RF phase is quantized into Q discrete values, in the DHPNN we use the ideal quantization (IQ) layer to quantize the continuous phase vector $\boldsymbol{\phi}$ into the discrete phase vector. Denote the IQ function and the step function by $\Lambda(\cdot)$ and $\varepsilon(\cdot)$, respectively, where $\varepsilon(x) = \begin{cases} 0, x < 0, \\ 1, x \geq 0. \end{cases}$ Then $\Lambda(\cdot)$ can be written as

$$\Lambda(x) \triangleq \frac{2\pi}{Q} \sum_{q=1}^{Q} \varepsilon\left(x - \frac{2\pi q}{Q}\right), \quad x \in [0, 2\pi). \tag{20}$$

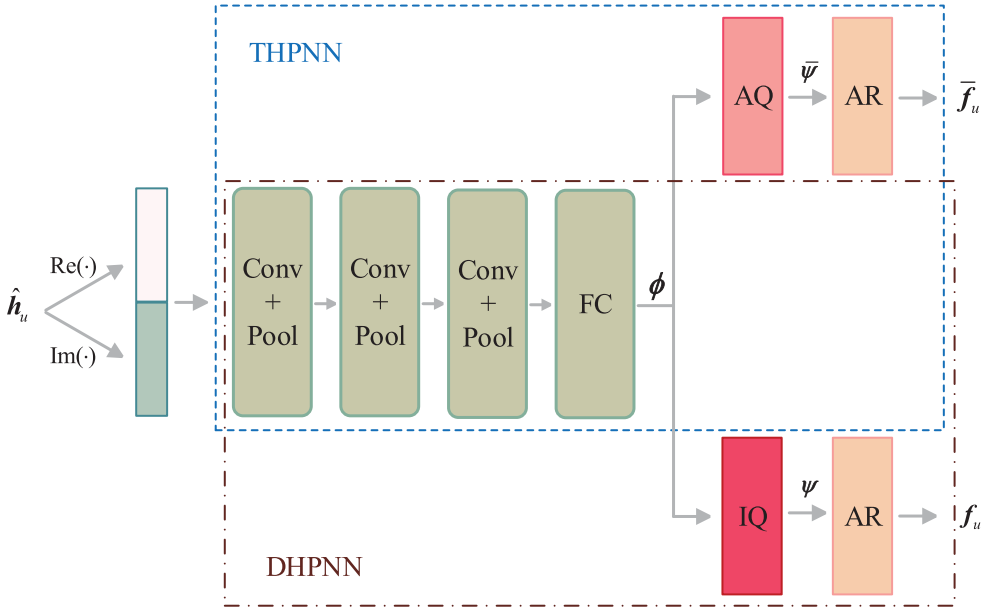


Fig. 5. Illustration of the THPNN for offline training and DHPNN for online deployment.

It is shown in Fig. 6 that $\Lambda(x)$ is not differentiable when $x = 2\pi q/Q$, $q = 1, 2, \ldots, Q$, which indicates that standard deep learning training algorithms, such as stochastic gradient descent (SGD), cannot be directly applied to train the NN. To overcome this problem, we use the approximate quantization (AQ) layer in the THPNN for offline training instead of the IQ layer. Therefore the DLQP hybrid precoder design method uses two NNs. However, the DLCS channel estimation scheme needs no quantization. Therefore the DLCS channel estimation scheme uses only one NN. Denote the AQ function by $\Gamma(x)$ [26], which can be represented as

$$\Gamma(x) \triangleq \frac{\pi}{Q}\sum_{q=1}^{Q}\tanh\left(\eta\left(x-\frac{2\pi q}{Q}\right)\right)+1, \quad x\in[0,2\pi) \quad (21)$$

where η a constant to represent the degree of approximation. As shown in Fig. 6, it is more accurate for $\Gamma(x)$ to approximate $\Lambda(x)$ if we set η as a larger number. It is also shown that $\Gamma(x)$ is differentiable for $x \in [0, 2\pi)$. Then we use $\Gamma(x)$ to quantize $\boldsymbol{\phi}$ in the THPNN for offline training. Denote the phase vector after quantization by $\bar{\boldsymbol{\psi}} \triangleq [\bar{\psi}_1, \bar{\psi}_2, \ldots, \bar{\psi}_{N_A}]^T \in \mathbb{R}^{N_A}$, which can be represented as

$$\bar{\psi}_n = \Gamma(\phi_n), \quad n = 1, 2, \ldots, N_A. \quad (22)$$

Based on the phase vector $\bar{\boldsymbol{\psi}}$, the approximate analog precoder vector $\bar{\boldsymbol{f}}_u$ can be obtained in the analog precoder reconstruction (AR) layer. By setting $\bar{\psi}_n$ as the phase of $\bar{f}_{u,n}$, $\bar{f}_{u,n}$ can be represented as

$$\bar{f}_{u,n} = e^{j\bar{\psi}_n}, \quad n = 1, 2, \ldots, N_A. \quad (23)$$

During the offline training of the THPNN, we generate the dataset of $\hat{\boldsymbol{h}}_u$ and $\boldsymbol{h}_u$ based on the output of the CENN and simulated mmWave channel environment. With the channel vector in (3) and the estimated channel vector in (18), the training data of $\hat{\boldsymbol{h}}_u$ and $\boldsymbol{h}_u$ can be obtained. In fact, the process to obtain $\hat{\boldsymbol{h}}_u$ and $\boldsymbol{h}_u$ involves the following five steps: **i)** we randomly generate the channel vector $\boldsymbol{h}_u$ based on the mmWave channel model in (3); **ii)** we compute the received signal vector $\boldsymbol{r}_u$ based on (10); **iii)** we obtain the correlation vector $\boldsymbol{c}_u$ based on (14); **iv)** we feed $\boldsymbol{c}_u$ to the offline-trained CENN for the DLCS channel estimation to get $\hat{\boldsymbol{g}}_u$; **v)** we obtain the estimated channel vector $\hat{\boldsymbol{h}}_u$ based on the channel reconstruction in (18). The output of the THPNN is $\bar{\boldsymbol{f}}_u$.

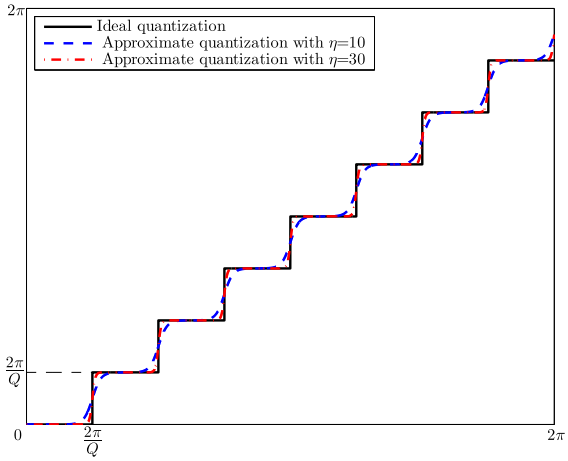


Fig. 6. Illustration of the ideal quantization and approximate quantization.

The training of the THPNN aims to maximize the beamforming gain, i.e., the inner product of $\bar{\boldsymbol{f}}_u$ and $\boldsymbol{h}_u$. Since the THPNN is trained to minimize the loss, we calculate the loss as the opposite number of the inner product, which can be

represented as [18]

$$f_{\text{LossQP}}(\bar{\boldsymbol{f}}_u, \boldsymbol{h}_u) = -\left|\bar{\boldsymbol{f}}_u^T \boldsymbol{h}_u\right|. \tag{24}$$

The training of the CENN aims to minimize the difference between $\hat{\boldsymbol{g}}_u$ and $\boldsymbol{g}_u$, while the training of the THPNN aims to maximize the beamforming gain. Therefore the loss function in (16) is different from that in (24). Note that the output of the NN is the analog precoder vector $\bar{\boldsymbol{f}}_u$, while we need to calculate the analog precoding matrix $\boldsymbol{F}_R$ so that the spectral efficiency can be obtained. However, the computational process from $\bar{\boldsymbol{f}}_u$ to $\boldsymbol{F}_R$ is not differentiable, which cannot be applied to the NN. Therefore we do not set the spectral efficiency as the loss. We adopt the adaptive moment estimation (Adam) optimizer to train the THPNN by TensorFlow. The THPNN is trained for 6,000 epochs, where 200 mini-batches are utilized in each epoch. The learning rate is set to be a step function. The learning rate is initialized with the value of 0.01 and decreases 2-fold every 2000 epochs.

During the online deployment stage, we obtain the DHPNN based on the offline-trained THPNN by replacing the AQ layer in the THPNN with IQ layer. To obtain the input of the THPNN $\hat{\boldsymbol{h}}_u$, we obtain the real measured $\boldsymbol{r}_u$ from practical mmWave channel environments. We compute $\boldsymbol{c}_u$ based on (14), which is then fed to the offline-trained CENN for the DLCS channel estimation to get $\hat{\boldsymbol{g}}_u$. We obtain the estimated channel vector $\hat{\boldsymbol{h}}_u$ based on the channel reconstruction in (18). We then feed $\hat{\boldsymbol{h}}_u$ to the DHPNN for the DLQP hybrid precoder design to get $\boldsymbol{f}_u$. Note that different from the offline training of the THPNN, we use the IQ function $\Lambda(x)$ to quantize $\boldsymbol{\phi}$ in the DHPNN, which ensures that each entry of $\boldsymbol{f}_u$ is drawn from the set $\{e^{j2\pi n/Q}, n = 1, 2, \ldots, Q\}$. Denote the phase vector after quantization by $\boldsymbol{\psi} \triangleq [\psi_1, \psi_2, \ldots, \psi_{N_A}]^T \in \mathbb{R}^{N_A}$, which can be represented as

$$\psi_n = \Lambda(\phi_n), \quad n = 1, 2, \ldots, N_A. \tag{25}$$

Based on the phase vector $\boldsymbol{\psi}$, the analog precoder vector $\boldsymbol{f}_u$ can be obtained. By setting ψ_n as the phase of $f_{u,n}$, $f_{u,n}$ can be represented as

$$f_{u,n} = e^{j\psi_n}, \quad n = 1, 2, \ldots, N_A. \tag{26}$$

After obtaining $\boldsymbol{f}_u$ in the online deployment stage, the analog precoding matrix $\boldsymbol{F}_R$ can be represented as

$$\boldsymbol{F}_R = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_U]. \tag{27}$$

It is shown in (24) that the analog precoder is designed to maximize the beamforming gain, where the quantization of the RF phase is considered. Note that although we use the AQ layer for offline training, we adopt the IQ layer in the online deployment stage, which guarantees the consistency of our adopted NN and the practical hardware constraint of limited phase shifter resolution.

B. Digital Precoder Design

We denote the estimated channel matrix for the BS and all users by

$$\hat{\boldsymbol{H}} \triangleq \left[\hat{\boldsymbol{h}}_1, \ldots, \hat{\boldsymbol{h}}_U\right]^T \in \mathbb{C}^{U \times N_A}. \tag{28}$$

We further denote the effective channel matrix by

$$\boldsymbol{H}_{\text{eff}} \triangleq \hat{\boldsymbol{H}} \boldsymbol{F}_R. \tag{29}$$

Analog precoding aims to form directional beams using phase shifter network, while digital precoding is designed to mitigate interference of multiple data streams after analog precoding. Then the ZF digital precoding matrix can be represented by

$$\boldsymbol{F}_B = \boldsymbol{H}_{\text{eff}}^H (\boldsymbol{H}_{\text{eff}} \boldsymbol{H}_{\text{eff}}^H)^{-1}. \tag{30}$$

To satisfy the total power constraint, each column of the designed digital precoder, denoted by $\boldsymbol{f}_{B,u}$, should be normalized, i.e.,

$$\boldsymbol{f}_{B,u} = \boldsymbol{f}_{B,u} / \left\|\boldsymbol{F}_R \boldsymbol{f}_{B,u}\right\|_2 \tag{31}$$

such that $\left\|\boldsymbol{F}_R \boldsymbol{f}_{B,u}\right\|_2^2 = 1$, $u = 1, 2, \ldots, U$.

It is shown that the proposed DLQP hybrid precoder design method can obtain the analog precoder considering the quantized phase constraint, which is of great value in practical mmWave systems.

V. Simulation Results

In the following we will present the performance evaluation for the proposed DLCS channel estimation scheme and the proposed DLQP hybrid precoder design method. Considering a multi-user mmWave massive MIMO communication system, the BS equipped with $N_R = 4$ RF chains and $N_A = 64$ antennas serves $U = 3$ users with single antenna. We set $G = 128$ according to [6], and we set the number of multiple paths in mmWave channel as $L_u = 2$, where $g_{u,1} \sim \mathcal{CN}(0, 1)$ and $g_{u,2} \sim \mathcal{CN}(0, 0.5)$ [6], [11]. For the uplink pilot transmission, we set $\boldsymbol{F}_B^k = \boldsymbol{I}_{N_R}$. Therefore the hybrid precoding matrix is equal to the analog precoding matrix and is also a random matrix. The quantization bit number of the phase shifters used at the BS is $B = 4$, leading to $Q = 16$ [6]. We set $\eta = 100$. Since $\boldsymbol{h}_u^b$ is not ideally sparse due to the power leakage, the beamspace channel sparse level should be larger than L_u, i.e., $J > 2$. We set $J = 6, 7$ in performance simulating. Note that the CENN is trained to predict the beamspace channel amplitude, where the training process of the CENN is independent of J. The proposed DLCS channel estimation scheme is compared with the existing OMP [5] and DGMP [4] channel estimation schemes, while the proposed DLQP hybrid precoder design method is compared with the existing QALS [14] hybrid precoder design method. We also compare the DLQP method with the Exhaustion hybrid precoder design method, i.e., we generate the analog precoding matrix $\boldsymbol{F}_R$ for $30,000$ times, where each entry of $\boldsymbol{F}_R$ is randomly drawn from the set $\{e^{j2\pi n/Q}, n = 1, 2, \ldots, Q\}$, and then the digital precoder is designed according to Algorithm 2. We select the hybrid precoder with the largest spectral efficiency as the output of the Exhaustion hybrid precoder design method.

We first evaluate the performance of the proposed DLCS channel estimation scheme from Fig. 7 to Fig. 10. As shown in Fig. 7, the channel estimation performance for the proposed DLCS scheme together with the existing schemes is compared in terms of SNR. The channel estimation performance is

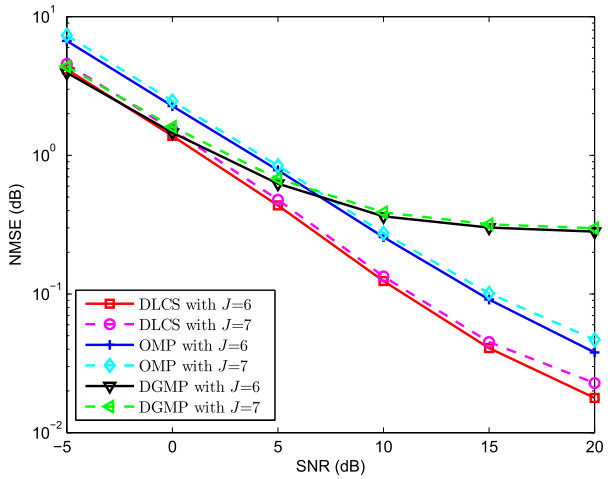


Fig. 7. Comparisons of channel estimation performance in terms of SNR for different schemes.

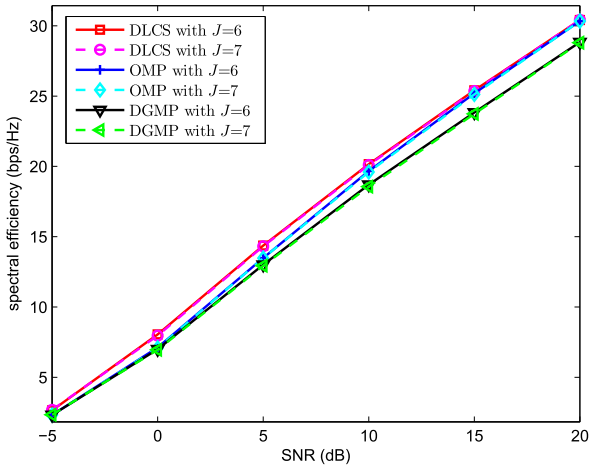


Fig. 8. Comparisons of spectral efficiency in terms of SNR for different channel estimation schemes.

measured by the normalized mean-squared error (NMSE), which is defined by

$$\text{NMSE} \triangleq \frac{\sum_{u=1}^{U} \|\hat{\boldsymbol{h}}_u - \boldsymbol{h}_u\|_2^2}{\sum_{u=1}^{U} \|\boldsymbol{h}_u\|_2^2}. \tag{32}$$

We use $K = 8$ time slots to transmit pilots for uplink channel estimation. To make a fair comparison, we fix the pilot training time slots to be eight for the OMP and DGMP schemes. It is shown that the DLCS scheme has better channel estimation performance than existing schemes. When SNR $=$ 10 dB, the DLCS scheme with $J = 6$ has 51.7% and 65.8% performance improvements over the OMP and DGMP schemes, respectively, while the DLCS scheme with $J = 7$ has 51.3% and 65.5% performance improvements over the OMP and DGMP schemes, respectively. We explain the reason for the performance improvements as follows. The OMP scheme estimates the beamspace channel dominant entries sequentially, which cannot guarantee global optimality. The DGMP scheme only estimates the LOS path, while the proposed DLCS scheme can simultaneously estimate all the dominant beamspace channel entries.

As shown in Fig. 8, we compare the spectral efficiency for the proposed DLCS scheme with the existing schemes in terms of SNR. Based on the estimated channel, there are various methods to design the hybrid precoding for mmWave downlink transmission. Similar to [6], in this work we wish to compare the upper bound of the downlink spectral efficiency, which can be simply measured by the fully-digital precoding. The ZF precoding matrix can be represented by $\boldsymbol{F}^{\text{dl}} \triangleq (\hat{\boldsymbol{H}}^* \hat{\boldsymbol{H}}^T)^{-1} \hat{\boldsymbol{H}}^*$. To meet the total power budget, the uth row of $\boldsymbol{F}^{\text{dl}}$, denoted by $\boldsymbol{f}_u^{\text{dl}}$, should be normalized, i.e., $\boldsymbol{f}_u^{\text{dl}} \leftarrow \boldsymbol{f}_u^{\text{dl}} / \|\boldsymbol{f}_u^{\text{dl}}\|_2$ such that $\|\boldsymbol{f}_u^{\text{dl}}\|_2 = 1$ for $u = 1, 2, \ldots, U$. Then the spectral efficiency is given by [6]

$$R \triangleq \sum_{u=1}^{U} \log_2 \left(1 + \frac{\frac{1}{U} \left| \boldsymbol{f}_u^{\text{dl}} \boldsymbol{h}_u \right|^2}{\frac{1}{U} \sum_{i \neq u} \left| \boldsymbol{f}_i^{\text{dl}} \boldsymbol{h}_u \right|^2 + \sigma^2} \right). \tag{33}$$

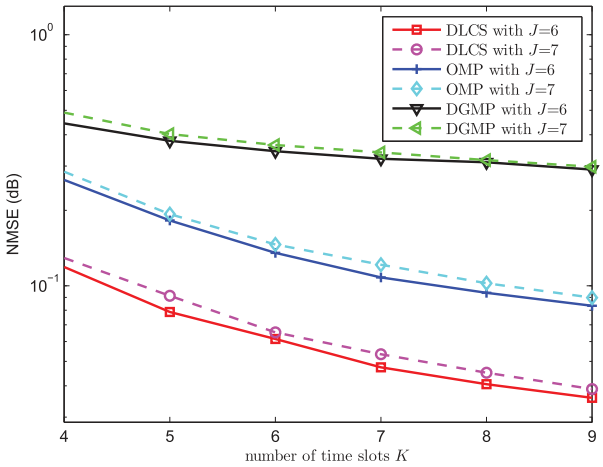


Fig. 9. Comparisons of channel estimation performance in terms of the number of time slots for channel training for different schemes.

It is seen from Fig. 8 that the DLCS scheme has better channel estimation performance than existing schemes. When SNR $=$ 10 dB, the DLCS scheme with $J = 6$ has 2.5% and 7.8% performance improvements over the OMP and DGMP schemes, respectively, while the DLCS scheme with $J = 7$ has 2.6% and 8.3% performance improvements over the OMP and DGMP schemes, respectively. The reason for the smaller spectral efficiency gap between different schemes than the NMSE gap is that the NMSE performance is much more sensitive to the success rate of the sparse recovery, while the spectral efficiency performance is determined by the beamforming gain and is less sensitive to the success rate of the sparse recovery.

In Fig. 9, the channel estimation performance for the DLCS, OMP, and DGMP schemes is compared in terms of the number of time slots for channel training. We use the same number of pilot training time slots for the DLCS, OMP, and DGMP schemes. SNR is fixed as 15 dB. From Fig. 9, it is shown that the DLCS scheme has the best channel estimation performance. When fixing the number of pilot training time

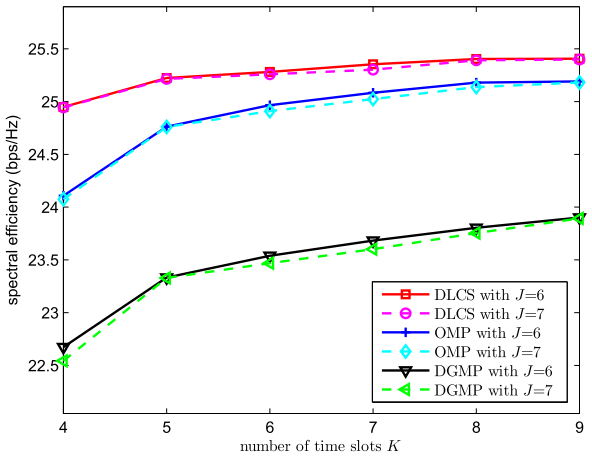


Fig. 10. Comparisons of spectral efficiency in terms of the number of time slots for channel training for different channel estimation schemes.

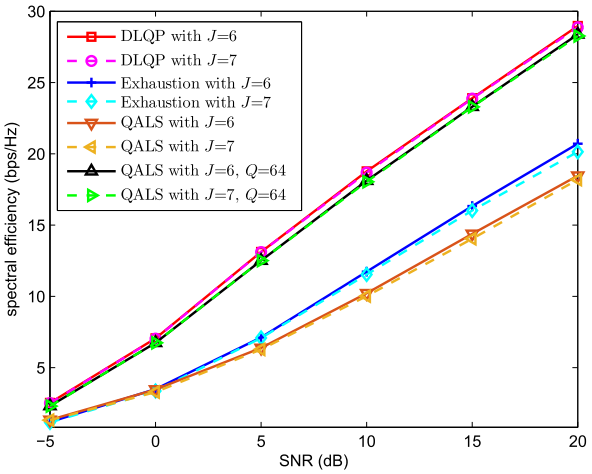


Fig. 11. Comparisons of spectral efficiency in terms of SNR for different hybrid precoder design methods.

slots to be $K = 7$, the DLCS scheme with $J = 6$ has 56.1% and 85.2% performance improvements over the OMP and DGMP schemes, respectively, while the DLCS scheme with $J = 7$ has 56.0% and 84.3% performance improvements over the OMP and DGMP schemes, respectively.

As shown in Fig. 10, we compare the spectral efficiency for different schemes in terms of the number of time slots for channel training. The system parameters for performance simulation are set to be the same as those for Fig. 9. It is shown that the DLCS scheme can have better channel estimation performance than the OMP and DGMP schemes. When the number of channel training time slots is more than eight, the spectral efficiency of the DLCS scheme remains constant, indicating that $K = 8$ is sufficient to obtain the full channel state information.

In the following, we evaluate the performance of the proposed DLQP hybrid precoder design method in Fig. 11 and Fig. 12. Fig. 11 compares of the spectral efficiency for the proposed DLQP hybrid precoder design method together with the existing methods in terms of SNR. Since the QALS method needs high phase shifter resolution to obtain analog

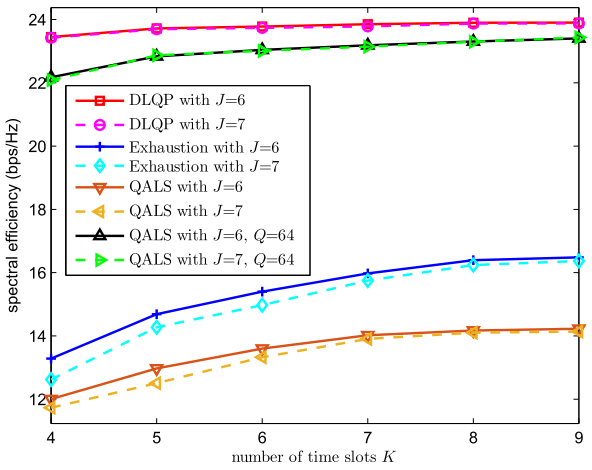


Fig. 12. Comparisons of spectral efficiency in terms of the number of time slots for channel training for different hybrid precoder design methods.

beamforming vectors aligning with the dominant channel paths, we also simulate the QALS method with $Q = 64$. It is seen from Fig. 11 that the DLQP method has better spectral efficiency performance than existing methods. When SNR = 10 dB, the DLQP method with $J = 6$ has 59.9%, 83.6% and 3.5% performance improvements over the Exhaustion, QALS with $Q = 16$ and QALS with $Q = 64$ methods, respectively, while the DLQP method with $J = 7$ has 62.0%, 86.5% and 3.6% performance improvements over the Exhaustion, QALS with $Q = 16$ and QALS with $Q = 64$ methods, respectively. We explain the reason for the performance gap as follows. In the Exhaustion method, although we generate the analog precoding matrix $\boldsymbol{F}_R$ for $30,000$ times, the number of the total possible $\boldsymbol{F}_R$ should be $Q^{N_A U} = 1.55 \times 10^{231}$, which is far more than the acceptable computational complexity. In the QALS method, the AoA of the LOS channel path cannot be aligned with well with the small number of available steering vectors of quantized angles.

As shown in Fig. 12, we compare the spectral efficiency for different hybrid precoding methods in terms of the number of time slots for channel training. The system parameters for performance simulation are set to be the same as those for Fig. 9. It is shown that the DLQP method can have better spectral efficiency performance than the Exhaustion and QALS methods. When fixing the number of pilot training time slots to be $K = 7$, the DLQP method with $J = 6$ has 49.3%, 70.1% and 2.9% performance improvements over the Exhaustion, QALS with $Q = 16$ and QALS with $Q = 64$ methods, respectively, while the DLQP method with $J = 7$ has 51.0%, 71.0% and 2.7% performance improvements over the Exhaustion, QALS with $Q = 16$ and QALS with $Q = 64$ methods, respectively.

VI. Conclusion

We proposed a DLCS channel estimation scheme and a DLQP hybrid precoder design method for the multi-user mmWave massive MIMO communication systems. The proposed DLCS scheme and DLQP method were compared with the existing works in the aspect of NMSE and spectral

efficiency. Simulation results showed that the proposed DLCS scheme has better channel estimation performance than existing schemes and the proposed DLQP method has high spectral efficiency with low resolution of phase shifters. As a future work, it is worth developing the channel estimation and hybrid precoding design for wideband multi-user mmWave massive MIMO transmission adopting deep learning.

REFERENCES

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

[2] P. Wang, Y. Li, L. Song, and B. Vucetic, "Multi-gigabit millimeter wave wireless communications for 5G: From fixed access to cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 168–178, Jan. 2015.

[3] P. V. Amadori and C. Masouros, "Low RF-complexity millimeter-wave beamspace-MIMO systems by beam selection," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2212–2223, Jun. 2015.

[4] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1259–1262, Jun. 2016.

[5] K. Venugopal, A. Alkhateeb, R. W. Heath, and N. G. Prelcic, "Time-domain channel estimation for wideband millimeter wave systems with hybrid architecture," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 6493–6497.

[6] J. Rodriguez-Fernandez, N. Gonzalez-Prelcic, K. Venugopal, and R. W. Heath, "Frequency-domain compressive channel estimation for frequency-selective hybrid mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2946–2960, May 2018.

[7] W. Ma and C. Qi, "Beamspace channel estimation for millimeter wave massive MIMO system with hybrid precoding and combining," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4839–4853, Sep. 2018.

[8] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[9] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016.

[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, C. Qi, and G. Y. Li, "Two-step codeword design for millimeter wave massive MIMO systems with quantized phase shifters," *IEEE Trans. Signal Process.*, vol. 68, pp. 170–180, Jan. 2020.

[12] 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.

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

[14] L. Zhao, D. W. K. Ng, and J. Yuan, "Multi-user precoding and channel estimation for hybrid millimeter wave systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1576–1590, Jul. 2017.

[15] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, Feb. 2018.

[16] Z. Qin, H. Ye, G. Y. Li, and B.-H.-F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[17] R. Liu, G. Yu, and G. Y. Li, "User association for ultra-dense mmWave networks with multi-connectivity: A multi-label classification approach," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1579–1582, Dec. 2019.

[18] Y. Wang, M. Narasimha, and R. W. Heath, "MmWave beam prediction with situational awareness: A machine learning approach," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2018, pp. 1–5.

[19] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmWave massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 852–855, Oct. 2018.

[20] T. Lin and Y. Zhu, "Beamforming design for large-scale antenna arrays using deep learning," 2019, *arXiv:1904.03657*. [Online]. Available: http://arxiv.org/abs/1904.03657

[21] Q. Wang and K. Feng, "PrecoderNet: Hybrid beamforming for millimeter wave systems using deep reinforcement learning," 2019, *arXiv:1907.13266*. [Online]. Available: http://arxiv.org/abs/1907.13266

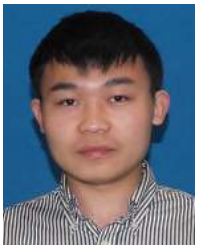
[22] J. Tao, C. Qi, and Y. Huang, "Regularized multipath matching pursuit for sparse channel estimation in millimeter wave massive MIMO system," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 169–172, Feb. 2019.

[23] 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.

[24] Z. Qin, J. Fan, Y. Liu, Y. Gao, and G. Y. Li, "Sparse representation for wireless communications: A compressive sensing approach," *IEEE Signal Process. Mag.*, vol. 35, no. 3, pp. 40–58, May 2018.

[25] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements," *IEEE Trans. Antennas Propag.*, vol. 61, no. 7, pp. 3814–3827, Jul. 2013.

[26] L. Rade and B. Westergren, *Mathematics Handbook for Science and Engineering*. Lund, Sweden: Studentlitteratur, 1998.



Wenyan Ma (Student Member, IEEE) received the B.S. degree (Hons.) in information engineering from the Chien-Shiung Wu College, Southeast University, Nanjing, China, in 2017, where he is currently pursuing the M.S. degree in signal and information processing. His research interests include signal processing for millimeter wave communications and massive multi-input multi-output (MIMO) systems. He received the Eleventh International Conference on Wireless Communications and Signal Processing (WCSP) Best Paper Award in 2019.

Chenhao Qi (Senior Member, IEEE) received the B.S. degree (Hons.) in information engineering from the Chien-Shiung Wu College, Southeast University, Nanjing, China, in 2004, and the Ph.D. degree in signal and information processing from Southeast University in 2010.

From 2008 to 2010, he visited the Department of Electrical Engineering, Columbia University, New York, NY, USA. Since 2010, he has been with the Faculty of the School of Information Science and Engineering, Southeast University, where he is currently an Associate Professor. His research interests include millimeter wave communications, massive multi-input multi-output (MIMO), satellite communications and intelligent signal processing. He received the IEEE Global Communications Conference (GLOBECOM) Best Paper Award and the Eleventh International Conference on Wireless Communications and Signal Processing (WCSP) Best Paper Award in 2019. He is an Exemplary Reviewer of IEEE COMMUNICATIONS LETTERS in 2017 and an Exemplary Editor of IEEE COMMUNICATIONS LETTERS in 2018. He is also an Outstanding Associate Editor of IEEE ACCESS in 2018. He serves as an Associate Editor for IEEE COMMUNICATIONS LETTERS, IEEE ACCESS, and the IEEE Open Journal of the Communications Society. He serves as the Symposium Co-Chair for international conferences, including the GLOBECOM, WCSP, and the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC).

Zaichen Zhang (Senior Member, IEEE) was born in Nanjing, China, in 1975. He received the B.S. and M.S. degrees in electrical and information engineering from Southeast University, Nanjing, China, in 1996 and 1999, respectively, and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong in 2002. From 2002 to 2004, he was a Post-Doctoral Fellow with the National Mobile Communications Research Laboratory, Southeast University. He joined the School of Information Science and Engineering, Southeast University, in 2004, where he is currently a Professor. He has published over 200 articles and issued 40 patents. His current research interests include 6G mobile communication systems, optical wireless communications, and quantum information processing.

Julian Cheng (Senior Member, IEEE) received the B.Eng. degree (Hons.) in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 1995, the M.Sc.(Eng.) degree in mathematics and engineering from Queen's University, Kingston, ON, Canada, in 1997, and the Ph.D. degree in electrical engineering from the University of Alberta, Edmonton, AB, Canada, in 2003. He was with Bell-Northern Research and Nortel Networks. He is currently a Full Professor with the Faculty of Applied Science, School of Engineering, The University of British Columbia, Kelowna, BC, Canada. His current research interests include digital communications over fading channels, statistical signal processing for wireless applications, optical wireless communications, and 5G wireless networks. He was the Co-Chair of the 12th Canadian Workshop on Information Theory in 2011, the 28th Biennial Symposium on Communications in 2016, and the 6th EAI International Conference on Game Theory for Networks (GameNets 216). He has served as a Guest Editor for a Special Issue of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS ON OPTICAL WIRELESS COMMUNICATIONS. He is also a Registered Professional Engineer with the Province of British Columbia, Canada. He serves as the President for the Canadian Society of Information Theory and the Secretary for the Radio Communications Technical Committee of the IEEE Communications Society. He was a past Associate Editor of IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and IEEE ACCESS. He serves as an Area Editor for IEEE TRANSACTIONS ON COMMUNICATIONS.