Bayesian Matching Pursuit Based Estimation of Off-grid Channel for Millimeter Wave Massive MIMO System

You You, Chuan Zhang, *Senior Member, IEEE,* and Li Zhang, *Senior Member, IEEE*

***Abstract*—Millimeter wave (mmWave) frequency spectrum offers orders of magnitude greater spectrum to mitigate the severe spectrum shortage in conventional cellular bands. To overcome the high propagation loss in the mmWave band, massive multiple-input and multiple-output (MIMO) can be adopted at both transmitter and receiver to provide large beamforming gains. At the same time, hybrid architecture is applied to reduce the huge power consumption caused by devices operating at radio frequency (RF). However, because of the hybrid architecture and large number of antennas, it is hard to obtain the channel state information (CSI) which is crucial for obtaining desirable beamforming gains. Off-grid error and sparsity pattern (SP) estimation error are two main limiting factors of the performance of most existing compressive sensing (CS) based channel estimation (CE) algorithms. Off-grid error presents when the true angle does not lie on the discretized angle grid of mmWave channel in the spatial domain. In this paper, we first propose a fast Bayesian matching pursuit method with 'virtual sparsity' to improve the accuracy of SP estimation and name it as the improved Bayesian matching pursuit (IBMP). Then an enhanced algorithm, named off-grid IBMP (OG-IBMP), is developed to mitigate the off-grid problem, followed by a theoretical analysis of OG-IBMP. This method iteratively updates the selected grid points and updates the corresponding parameters based on the maximum a posteriori (MAP) criterion. Numerical simulations are performed to validate our theoretical analysis and evaluate the performance of the proposed method. Compared to other existing methods, the results show that our proposed OG-IBMP algorithm greatly reduces the off-grid error and significantly enhances the accuracy of the SP estimation with low computational complexity.**

***Index Terms*—Compressed sensing (CS), Channel estimation (CE), Bayes methods, Optimization methods.**

I. Introduction

MILLIMETER wave (mmWave) is a promising approach for the fifth Generation (5G) and beyond wireless networks because of its large amount of available bandwidths [1]. The critical challenge is the huge propagation loss in the mmWave frequency bands. Thanks to the mm-level wavelengths, massive multiple-input and multiple-output (MIMO) can be packed into a small area to provide desirable beamforming gain and compensates for the path loss. Additionally, a hybrid MIMO architecture consisting of an analog beamformer in the radio frequency domain cascaded with a digital MIMO processor in the baseband has been proposed for mmWave communication to reduce the high power consumption of the power hungry devices such as the analog to digital converters (ADC) in the radio frequency domain [2].

As in the conventional microwave systems, channel state information (CSI) is needed to design precoding and combining procedures at the transmitters and receivers in mmWave systems. Although massive MIMO and the hybrid architecture help overcome the path loss and power consumption issues, the large number of antennas and analog combining also make it hard to obtain the required CSI. Fortunately, the mmWave channels exhibit sparsity in the angular domain due to the large dimension of the channels [3]. Many previous studies leverage this channel sparsity and apply compressive sensing (CS) [4] based sparse recovery techniques to estimate the channel from a smaller set of measurements. Existing works can be divided into beam training based methods [5], [6] and open-loop channel estimation (CE) methods [7], [8]. The beam training-based methods apply various smart searching methods such as exploiting multi-resolution beamforming codebook [5] for transmit and receive beam pairs that maximize the received signal-to-noise ratio (SNR). However, the performance of the close-loop methods are limited by the design of codebooks and it is difficult to be applied to long-distance communications. Because the beam training-based methods begin with wide pilot beams based on the designed code-book that cover all of the angles of interest during the initial search stage, high transmit power is required to maintain SNR.

An alternative way is to apply the open-loop CE methods which can reduce the feedback overhead and use a fixed beam width, overcoming the limits of beam training strategies. Prior works on open-loop CE for mmWave communication can be divided into non-Bayesian based algorithms [7] and Bayesian based algorithms [8]–[11]. Orthogonal matching pursuit (OMP) [7] algorithm is a typical non-Bayesian based algorithm. It is an iterative algorithm finding the sub-optimal

This work was supported in part by Nanjing postdoctoral funded recruitment, in part by NSFC under Grants 62122020 and 61871115, in part by the Jiangsu Provincial NSF under Grant BK20211512, and in part by the Major Key Project of PCL under Grant CL2021A01-2. *(Corresponding author: Li Zhang.)*

Y. You was with the School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, U.K., and now with the LEADS of Southeast University, the National Mobile Communications Research Laboratory of Southeast University, and the Purple Mountain Laboratories, Nanjing 211189, China.

C. Zhang is with the LEADS of Southeast University, the National Mobile Communications Research Laboratory of Southeast University, and the Purple Mountain Laboratories, Nanjing 211189, China.

L. Zhang is with the School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, U.K. (L.X.Zhang@leeds.ac.uk).

solution by selecting at each iteration the column of the sensing matrix which is the most correlated with the current residuals. Recently, many low complexity OMP based CE methods have been proposed for mmWave communication. Most of these methods require the number of the significant elements (sparsity) [2], and even with known sparsity, sparsity patterns (SP) estimation can still be inaccurate. A SP specifies which elements of the vector are non-zero. And the SP estimation error is the difference between the estimated SP and the true SP. In the mmWave CE, non-Bayesian methods are easily deteriorated by large noise [2]. Sparse Bayesian learning (SBL) [9] and Bayesian compressive sensing (BCS) [10] are two typical Bayesian based algorithms. They assume that each element follows Gaussian distribution with unknown variance which is assigned the Gamma conjugate prior. SBL utilizes expectation maximization (EM) method to compute a maximum a posteriori (MAP) estimate, while BCS adopts a more efficient method by analysing the properties of the marginal likelihood function. Another Bayesian based iterative channel estimation algorithm using the least square estimation (LSE), EM and sparse message passing (SMP) is proposed to further improve the mmWave channel estimation accuracy [11]. These Bayesian based methods show better estimation accuracy than OMP but with hundreds of times greater complexity. In earlier works, the authors have proposed Bayesian matching pursuit (FBMP) [8] which makes appropriate assumptions according to the characteristics of the mmWave channel and selects a set of candidate SPs with high posterior probabilities to estimate CSI. FBMP shows superior performance than other Bayesian based algorithms. However, error floor occurs at high SNRs as shown by the simulation results in [8]. In [8], the authors also have to choose small virtual sparsity or large virtual sparsity based on the range of real sparsity. If there is a big difference between the virtual sparsity and real sparsity, the recovery performance of FBMP will deteriorate. This means, a rough priori information of sparsity is required to achieve accurate estimation.

Considering the wideband transmission in mmWave systems, the non-Bayesian based CS methods have been applied to the mmWave systems in the time domain [12], and in the frequency domain with simultaneous weighted-OMP (SW-OMP) [13]. Similarly, the Bayesian based CS methods such as sparse Bayesian learning (SBL) also has been extended to the hybrid wideband mmWave MIMO system for channel estimation with a single-carrier considering frequency-selective scenarios [14]. Moreover, in the case of using OFDM system in the frequency-selective fading mmWave channel estimation, angular sparsity is shared by multiple sub-carriers and is called common sparsity [15]. By utilizing the common sparsity, the channel estimation can be formulated as a CS multiple measurement vectors (MMV) problem to further improve the channel estimation performance [16].

All the above mentioned Bayesian based or non-Bayesian based algorithms take advantages of the channel sparsity in the angular domain using virtual channel representation [17] and the CS technology. As a result, these solutions are based on an assumption that the angles of arrival/departure (AoAs/AoDs) lie exactly on the grid. However, in practice, actual AoDs/AoAs are continuous and thus off-grid errors exist. Off-grid error is defined as the difference between the continuous angle and the nearest discrete grid point. It has been proved that the off-grid problem results in power leakage and degrades the CE accuracy significantly in mmWave communication [18]. The off-grid errors problem has been studied extensively in compressive sensing literatures such as [19], [20]. But most of them are not suitable for mmWave CE because of the unaffordable computational complexity. Recently, some off-grid mitigation methods have been proposed for mmWave CE. Optimization methods such as interior point and gradient descent were applied in [18] and [21] to mitigate the impact of the off-grid errors, respectively. And [22] proposed an OMP-based algorithms to exploit the implicit Dirichlet structure in the Fourier domain to combat the off-grid effects. However, [18], [21], [22] are non-Bayesian based methods and the SP estimation error deteriorates the performance. Some off-grid mitigation methods are proposed for Bayesian learning based methods to achieve super resolution mmWave CE. For example, [23] proposed an improved SBL based method utilizing Taylor expansion to find a more accurate angle set to reconstruct the CSI. However, Bayesian learning based algorithms such as [23] all have unbearable complexity due to the learning process.

In this paper, we first propose a matching pursuit method named as improved Bayesian matching pursuit (IBMP) for mmWave CE. After analysing the impact of the off-grid errors, we further propose an off-grid IBMP (OG-IBMP) method to solve the existing problems in IBMP. The contributions of this paper are described as follows:

1) We formulate the mmWave CE problem as a sparse signal recovery problem and propose the IBMP method to solve it using virtual sparsity instead of requiring the real sparsity. Among the methods without any off-grid error mitigation, IBMP achieves the best performance at lower SNRs. Moreover, it has evidently lower complexity compared with the Bayesian learning based methods and significantly better performance compared with the non-Bayesian based methods. However, we find that error floor occurs at high SNRs, and any significant difference between the virtual sparsity and real sparsity deteriorates estimation performance.
2) Theoretical analysis is presented to demonstrate the impact of the off-grid errors on the IBMP method. Then OG-IBMP method is proposed to overcome the disadvantages of IBMP. To the best of our knowledge, no previous studies have considered the off-grid problem in Bayesian matching pursuit based algorithms. And the proposed OG-IBMP is the first Bayesian matching CE algorithm with off-grid mitigation. It greatly reduces the computational complexity in comparison with other off-grid Bayesian learning based CE methods. Simulation results show that OG-IBMP achieves superior performance compared with the existing methods including IBMP for all SNRs without the need of sparsity information.

The remainder of this paper is organized as follows. In

Section II, we introduce the mmWave communication system model and formulate the CE as a sparse signal recovery problem. In Section III, we propose the IBMP algorithm for mmWave CE. In Section IV, we present theoretical analysis on IBMP for the performance deterioration at high SNRs and demonstrate the impact of off-grid errors. Based on the theoretical analysis, we propose a modified method, i.e. OG-IBMP based on sequential quadratic programming (SQP) in order to mitigate the off-grid problem. In Section V, simulation results are presented to demonstrate the superiority of OG-IBMP. In Section VI, we conclude the paper.

II. System Models

A. System Setup

We consider a single user massive MIMO system with fully connected hybrid architecture as shown in Fig. 1, where the transmitter employs N_T antennas and N_{RF} RF chains to communicate with a receiver with N_R antennas and N_{RF} RF chains ($N_{RF} \leq \min(N_T, N_R)$).

In the CE stage, transmitter applies N_T^B ($N_T^B \leq N_T$) different transmit beams denoted as $\{\mathbf{f}_m \in \mathbb{C}^{N_T \times 1} : m = 1, \dots, N_T^B\}$ to transmit pilots symbol x_p and receiver uses N_R^B ($N_R^B \leq N_R$) different receive beams denoted as $\{\mathbf{w}_n \in \mathbb{C}^{N_R \times 1} : n = 1, \dots, N_R^B\}$. We assume that the transmitter sends training beams $\mathbf{f}_m$ to receiver successively. Because the receiver has a limited number of RF chains, it only generates N_{RF} receive beams simultaneously. The receive signal in one time slot can be represented by $\mathbf{y}_q \in \mathbb{C}^{N_{RF} \times 1}, q \in \{1, \dots, N_R^b\}$ where q denotes the received block index and $N_R^b = \frac{N_R^B}{N_{RF}}$ is the number of received blocks. Note that, to simplify the mathematical expressions, we assume that the number of RF chains at transceiver are the same and the number of transceiver beams are multiples of the number of RF chains. The received vector for the q-th block and the m-th transmit beam is given by

$$\mathbf{y}_{q,m} = \mathbf{W}_q^H \mathbf{H} \mathbf{f}_m x_p + \mathbf{W}_q^H \mathbf{n}_{q,m}, \tag{1}$$

where $\mathbf{W}_q = [\mathbf{w}_{(q-1)N_{RF}+1}, \dots, \mathbf{w}_{qN_{RF}}] \in \mathbb{C}^{N_R \times N_{RF}}$ is the receive beam pattern matrix in the q-th time slot. $\mathbf{H} \in \mathbb{C}^{N_R \times N_T}$ represents the channel matrix, and $\mathbf{n} \in \mathbb{C}^{N_R \times 1}$ is the noise vector. Collecting $\mathbf{y}_{q,m}$ for $q \in \{1, \dots, N_R^b\}$, we get the complete received signal for the m-th transmit beam as

$$\begin{aligned} \mathbf{y}_m = \mathbf{W}^H \mathbf{H} \mathbf{f}_m x_p &+ \text{diag}(\mathbf{W}_1^H, \dots, \mathbf{W}_{N_R^b}^H) \\ &\times [\mathbf{n}_{1,m}^T, \dots, \mathbf{n}_{N_R^b,m}^T]^T, \end{aligned} \tag{2}$$

where $\mathbf{W} = [\mathbf{W}_1, \dots, \mathbf{W}_{N_R^b}] \in \mathbb{C}^{N_R \times N_R^B}$, $\mathbf{y}_m \in \mathbb{C}^{N_R^B \times 1}$. Collecting $\mathbf{y}_m$ for $m \in \{1, \dots, N_T^B\}$ to get the received signal for all N_T^B transmit beams as

$$\begin{aligned} \mathbf{Y} &= \mathbf{W}^H \mathbf{H} \mathbf{F} \mathbf{X} + \mathbf{N} \\ &= \sqrt{P} \mathbf{W}^H \mathbf{H} \mathbf{F} + \mathbf{N}, \end{aligned} \tag{3}$$

where $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_{N_T^B}] \in \mathbb{C}^{N_R^B \times N_T^B}$, $\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_{N_T^B}] \in \mathbb{C}^{N_T \times N_T^B}$ and $\mathbf{N} \in \mathbb{C}^{N_R^B \times N_T^B}$ is the noise matrix given by

$$\begin{aligned} \mathbf{N} = \text{diag}(\mathbf{W}_1^H, \dots, \mathbf{W}_{N_R^b}^H) \big[&[\mathbf{n}_{1,1}^T, \dots, \mathbf{n}_{N_R^b,1}^T]^T, \\ &\dots, [\mathbf{n}_{1,N_T^B}^T, \dots, \mathbf{n}_{N_R^b,N_T^B}^T]^T \big]. \end{aligned} \tag{4}$$

The matrix $\mathbf{X} \in \mathbb{C}^{N_T^B \times N_T^B}$ is a diagonal matrix with x_p on its diagonal. Throughout the paper, we assume identical pilot symbols so that $\mathbf{X} = \sqrt{P}\mathbf{I}_{N_T^B}$ where P is the pilot power. $\mathbf{F}$ and $\mathbf{W}$ are regarded as beamforming matrices. Because hybrid analog/digital architecture is employed in mmWave communication, they can be decomposed as $\mathbf{F} = \mathbf{F}_{RF}\mathbf{F}_{BB}$ and $\mathbf{W} = \mathbf{W}_{RF}\mathbf{W}_{BB}$, where $\mathbf{F}_{RF}$ and $\mathbf{W}_{RF}$ represent the RF beamforming matrices, $\mathbf{F}_{BB}$ and $\mathbf{W}_{BB}$ represent the baseband processing matrices. Specifically, we assume $\mathbf{F}_{RF} = [\mathbf{F}_{RF,1}, \dots, \mathbf{F}_{RF,N_T^b}] \in \mathbb{C}^{N_T \times N_T^B}$, where $\mathbf{F}_{RF,\bar{t}} \in \mathbb{C}^{N_T \times N_{RF}}$ represents the transmit RF beamforming matrix in the $\bar{t}$-th time slot. Therefore the matrix $\mathbf{F}_{RF}$, containing N_T^B/N_{RF} number of $\mathbf{F}_{RF,\bar{t}}$, has an dimension of $N_T \times N_T^B$. Similarly, we assume $\mathbf{W}_{RF} = [\mathbf{W}_{RF,1}, \dots, \mathbf{W}_{RF,N_R^b}] \in \mathbb{C}^{N_R \times N_R^B}$, where $\mathbf{W}_{RF,\bar{r}} \in \mathbb{C}^{N_R \times N_{RF}}$ represents the receive RF beamforming matrix in the $\bar{r}$-th time slot. Considering the baseband precoder, we assume $\mathbf{F}_{BB} = \text{diag}(\mathbf{F}_{BB,1}, \dots, \mathbf{F}_{BB,N_T^b}) \in \mathbb{C}^{N_T^B \times N_T^B}$, where $\mathbf{F}_{BB,\bar{t}} \in \mathbb{C}^{N_{RF} \times N_{RF}}$ represents the baseband precoder matrix in the $\bar{t}$ time slot. The columns and rows of $\mathbf{F}_{BB}$ are divided into N_T^b parts and each sub-matrix on the diagonal has dimension of $N_{RF} \times N_{RF}$. Therefore, the baseband precoder fully uses the RF chains in this case. Similarly, we assume $\mathbf{W}_{BB} = \text{diag}(\mathbf{W}_{BB,1}, \dots, \mathbf{W}_{BB,N_R^b}) \in \mathbb{C}^{N_R^B \times N_R^B}$, where $\mathbf{W}_{BB,\bar{r}} \in \mathbb{C}^{N_{RF} \times N_{RF}}$ represents the baseband precoder matrix in the $\bar{r}$ time slot. As a result, (3) can be formulated as

$$\mathbf{Y} = \sqrt{P}(\mathbf{W}_{RF}\mathbf{W}_{BB})^H \mathbf{H} (\mathbf{F}_{RF}\mathbf{F}_{BB}) + \mathbf{N}. \tag{5}$$

B. Channel Model

The mmWave channel is often represented in the frequency domain [2]. In general, it can be written as

$$\mathbf{H}(t, f) = \sum_{i=1}^{N_{cl}} \sum_{j=1}^{N_{ray}} \alpha_{ij} e^{j2\pi(\nu_{ij}t - \tau_{ij}f)} \mathbf{a}_r(\theta_{ij}^r) \mathbf{a}_t^H(\theta_{ij}^t), \tag{6}$$

where N_{cl} and N_{ray} represent the number of clusters and the number of rays in each cluster, respectively. The (i, j)th multipath component (the j-th ray in the i-th cluster) is described by 5 parameters including the AoD θ_{ij}^t, AoA θ_{ij}^r, delay τ_{ij}, complex gain α_{ij} and Doppler shift ν_{ij}. We assume that each scatterer contributes only one path of propagation between transmitter and receiver and L represents the number of paths. Suppose that the channel is slow time varying and the bandwidth of the channel is sufficiently small. In mmWave communications, the widely adopted narrowband spatial channel model [7] can be given by

$$\begin{aligned} \mathbf{H} &= \sum_{l=1}^{L} \alpha_l \mathbf{a}_r(\theta_l^r) \mathbf{a}_t^H(\theta_l^t) \\ &\approx \mathbf{A}_R \mathbf{H}_b \mathbf{A}_T^H, \end{aligned} \tag{7}$$

where α_l, θ_l^r and θ_l^t are the complex channel gain, AoA and AoD of the l-th path, respectively. $\mathbf{a}_t(\theta_l^t)$ and $\mathbf{a}_r(\theta_l^r)$ are the array response vector for the transmitter and receiver, respectively. Considering only the azimuth, and neglecting elevation, we assume that the transmitter and receiver only

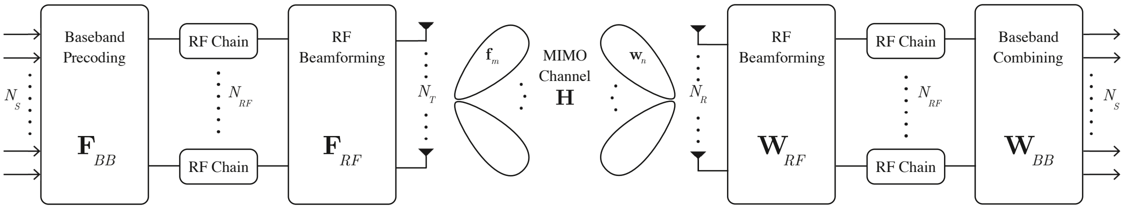

Fig. 1. Hybrid Massive MIMO system for mmWave communication [2].

implement horizontal (2-D) beamforming. Extensions to 3-D beamforming are possible [24]. While the algorithms and results developed in the paper can be applied to arbitrary antenna arrays, we consider the uniform linear arrays (ULAs) in the simulations of Section V so that the OMP [7], IP-OMP [18], and FBMP [8] can be adopted for performance comparison. If N_T and N_R ULA are assumed at transmitter and receiver, $\mathbf{a}_t(\theta_l^t)$ and $\mathbf{a}_r(\theta_l^r)$ can be written as

$$\begin{aligned}\mathbf{a}_t(\theta_l^t)=&\frac{1}{\sqrt{N_T}}[1,e^{-j\beta\cos\theta_l^t},\dots,e^{-j\beta\cos\theta_l^t(N_T-1)}]^T,\\ \mathbf{a}_r(\theta_l^r)=&\frac{1}{\sqrt{N_R}}[1,e^{-j\beta\cos\theta_l^r},\dots,e^{-j\beta\cos\theta_l^r(N_R-1)}]^T,\end{aligned}\tag{8}$$

where d and λ denote the normalized spacing antenna spacing and wavelength of operation, $\beta=-j2\pi\frac{d}{\lambda}$. In this paper, we consider $d=\frac{\lambda}{2}$.

In order to utilize the channel sparsity in angular domain, virtual channel representation [25] is widely employed in mmWave CE. Specifically, it assumes that all the angles fall onto a predefined set of discrete angles, namely, the 'grid'. In fact, the virtual channel representation is not exactly equal to the real channel matrix because the true continuous AoDs/AoAs do not fall onto the grid points precisely.

To complete the problem formulation without losing generality and simplicity, we assume that the AoAs, and AoDs are taken from the uniform grids of G_r and G_t grid points, respectively. Specifically, the uniform grid points for the AoAs are selected from $[0,\frac{\pi}{G_r-1},\frac{2\pi}{G_r-1},\dots,\frac{\pi(G_r-1)}{G_r-1}]$, and $G_r\gg L$ for the desired resolution as in [5] and [7]. The G_t uniform grid points for the AoDs are selected in a similar way. Note that other well-designed grids such as those in [25] and [7] can be employed to improve the accuracy of mmWave CE. The impact of the different grid designs will be discussed and shown in Section V. In this section, we only consider the uniformly quantized AoAs/AoDs. Based on the discrete angle grid, the channel matrix $\mathbf{H}$ in (7) can be approximated as $\mathbf{A}_R\mathbf{H}_b\mathbf{A}_T^H$ where $\mathbf{A}_R=[\mathbf{a}_r(0),\dots,\mathbf{a}_r(\frac{\pi}{G_r-1}),\dots,\mathbf{a}_r(\frac{\pi(G_r-1)}{G_r-1})]\in\mathbb{C}^{N_R\times G_r}$, $\mathbf{A}_T=[\mathbf{a}_t(0),\dots,\mathbf{a}_t(\frac{\pi}{G_t-1}),\dots,\mathbf{a}_t(\frac{\pi(G_t-1)}{G_t-1})]\in\mathbb{C}^{N_T\times G_t}$ and $\mathbf{H}_b\in\mathbb{C}^{G_r\times G_t}$ is a L-sparse channel gain matrix. In this paper, we assume that $G_t=G_r=G$ for simplicity. The difference between $\mathbf{H}$ and $\mathbf{A}_R\mathbf{H}_b\mathbf{A}_T^H$ is caused by the off-grid error as a result of quantification.

C. Problem Formulation

Considering the system model in (3) and the large dimension of channel matirx $\mathbf{H}$, conventional CE such as the least square (LS) requires very large training overhead in the mmWave systems. In order to reduce the training overhead and computational load by utilizing the channel sparsity in angular domain, the problem can be formulated as a sparse signal recovery problem by vectorizing $\mathbf{Y}$ in (3). Using the property of Kronecker product [7] $\text{vec}(\mathbf{ABC})=(\mathbf{C}^T\otimes\mathbf{A})\cdot\text{vec}(\mathbf{B})$ for $\mathbf{Y}$ and $\mathbf{H}$, we can get

$$\begin{aligned}\mathbf{y}_v\approx&\sqrt{P}(\mathbf{F}^T\otimes\mathbf{W}^H)\cdot\text{vec}(\mathbf{H})+\text{vec}(\mathbf{N})\\ =&\sqrt{P}(\mathbf{F}^T\otimes\mathbf{W}^H)\text{vec}(\mathbf{A}_R\mathbf{H}_b\mathbf{A}_T^H)+\mathbf{n}_Q\\ =&\sqrt{P}(\mathbf{F}^T\otimes\mathbf{W}^H)(\mathbf{A}_T^*\otimes\mathbf{A}_R)\text{vec}(\mathbf{H}_b)+\mathbf{n}_Q\\ =&\sqrt{P}(\mathbf{F}^T\otimes\mathbf{W}^H)\mathbf{A}_D\mathbf{h}+\mathbf{n}_Q\\ =&\mathbf{Q}\cdot(\mathbf{h})+\mathbf{n}_Q,\end{aligned}\tag{9}$$

where $\mathbf{y}_v\in\mathbb{C}^{M\times 1}$ is the vectorized received signal and $M=N_T^BN_R^B$ is the measurement dimension. $\mathbf{A}_D=\mathbf{A}_T^*\otimes\mathbf{A}_R$ is an $N_TN_R\times G^2$ dictionary matrix that consists of the G^2 column vectors, and the $\big(G(u-1)+v\big)$th column is calculated from $\mathbf{a}_t^*(\theta_u)\otimes\mathbf{a}_r(\theta_v)$, with θ_u and θ_v, the uth and vth points of the angle grid. $\mathbf{h}=\text{vec}(\mathbf{H}_b)=(h_1,h_2,\dots,h_N)^T$ is the vectorized channel gain of the corresponding quantized directions where $N=G^2$ and $\{h_n\}_{n=0}^N$ are the elements. $\mathbf{Q}=\sqrt{P}(\mathbf{F}^T\otimes\mathbf{W}^H)\mathbf{A}_D\in\mathbb{C}^{M\times N}$ is the sensing matrix. According to (9), sparse vectorized channel path gain $\mathbf{h}$ can be recovered from the noisy received signal $\mathbf{y}_v$ with known sensing matrix $\mathbf{Q}$ by the CS methods.

III. Proposed Improved Bayesian Matching Pursuit Algorithm

CS based algorithms including OMP, SBL, BCS and FBMP have been applied in mmWave CE. Among them, FBMP has significantly better performance compared with other methods especially at low SNRs [8]. In the FBMP method, appropriate assumptions are made according to the characteristics of mmWave channel and a set of candidate SPs with high posterior probabilities are selected for a minimum mean square error (MMSE) estimator to improve the CE performance. However, it has many disadvantages such as relatively high complexity, and degradation due to off-grid errors. Thus, in this section, we simplify the FBMP based CE method from a MMSE estimator to a MAP estimator with appropriate assumption to reduce the computational load.

In order to apply the Bayesian matching pursuit idea, we need to choose our signal model and priors according to the characteristics of the mmWave channel. In this paper, we assume that the path amplitudes α_l are Rayleigh distributed, i.e., $\{\alpha_l\}_{l=1}^L \sim \mathcal{CN}(0, \sigma^2)$ with $\sigma^2 = 1$ the average power gain and the noise $\mathbf{n}_Q$ in (9) is assumed to be white circular Gaussian noise as $\mathbf{n}_Q \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_M)$.

In this application, $\{h_n\}_{n=0}^N$ are assumed to be drawn from two specific Gaussian distributions indexed by $s_n = t \in \{0, 1\}$. $s_n = 0$ indexes the distribution with $(\mu_0, \sigma_0^2) = (0, 0)$ which implies $h_n = 0$ and $s_n = 1$ indexes the distribution with (μ_1, σ_1^2) which allows $h_n \neq 0$. Without prior information, we choose $(\mu_1, \sigma_1^2) = (0, J)$ where J can be any positive number. We choose $J = 1$ in this section and different values of J will be discussed according to the specific application in Section V. $\{s_n\}_{n=0}^{N-1}$ are treated as i.i.d random variables as $\Pr\{s_n = t\} = \lambda_t$ $(0 < \lambda_t \leq 1)$. λ_t is the probability that h_n follows Gaussian distribution indexed by $s_n = t$. We make $\lambda_1 \ll 1$ to ensure that $\mathbf{h}$ is sparse. Considering $\mathbf{h} = [h_0, \ldots, h_{N-1}]^T$ and $\mathbf{s} = [s_0, \ldots, s_{N-1}]^T$, the conditional probability of $\mathbf{h}$ given that $\mathbf{s}$ occurs can be written as $\mathbf{h} \mid \mathbf{s}$ where

$$\mathbf{h} \mid \mathbf{s} \sim \mathcal{CN}(\boldsymbol{\mu}(\mathbf{s}), \mathbf{R}(\mathbf{s})), \tag{10}$$

$[\boldsymbol{\mu}(\mathbf{s})]_n = \mu_t$ and $\mathbf{R}(\mathbf{s})$ has diagonal $[\mathbf{R}(\mathbf{s})]_{n,n} = \sigma_t^2$. Considering (9), the channel vector $\mathbf{h}$ and the received signal $\mathbf{y}_v$ are joint Gaussian conditioned on the mixture parameters $\mathbf{s}$ as

$$\begin{bmatrix} \mathbf{y}_v \\ \mathbf{h} \end{bmatrix} \Bigg| \mathbf{s} \sim \mathcal{CN}\left(\begin{bmatrix} \mathbf{Q}\boldsymbol{\mu}(\mathbf{s}) \\ \boldsymbol{\mu}(\mathbf{s}) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Phi}(\mathbf{s}) & \mathbf{Q}\mathbf{R}(\mathbf{s}) \\ \mathbf{R}(\mathbf{s})\mathbf{Q}^H & \mathbf{R}(\mathbf{s}) \end{bmatrix} \right), \tag{11}$$

where

$$\boldsymbol{\Phi}(\mathbf{s}) \triangleq \mathbf{Q}\mathbf{R}(\mathbf{s})\mathbf{Q}^H + \sigma_n^2 \mathbf{I}_M. \tag{12}$$

To estimate the CSI, we store the set of all possible SPs as $\mathbf{S}$ and seek to find the MAP estimate of $\mathbf{h}$ from $\mathbf{y}_v$ as

$$\hat{\mathbf{h}}_{\text{map}} \triangleq \mathrm{E}\{\mathbf{h} | \mathbf{y}_v, \mathbf{s}_{\text{map}}\}, \tag{13}$$

where $\mathbf{s}_{\text{map}}$ is the sparsity pattern which has the largest posterior probability $p(\mathbf{s}_{\text{map}} | \mathbf{y}_v)$ among all possible 2^N $p(\mathbf{s}|\mathbf{y}_v)_{\mathbf{s} \in \mathbf{S}}$. From (11) it is straightforward [26] to obtain

$$\begin{aligned} \mathrm{E}\{\mathbf{h}|\mathbf{y}_v, \mathbf{s}_{\text{map}}\} = \boldsymbol{\mu}(\mathbf{s}_{\text{map}}) + \mathbf{R}(\mathbf{s}_{\text{map}})\mathbf{Q}^H \boldsymbol{\Phi}(\mathbf{s}_{\text{map}})^{-1} \\ \big(\mathbf{y}_v - \mathbf{Q}\boldsymbol{\mu}(\mathbf{s}_{\text{map}})\big). \end{aligned} \tag{14}$$

We note that the primary challenge in the computation of (14) is to find out $\mathbf{s}_{\text{map}}$ and calculate $\boldsymbol{\Phi}(\mathbf{s}_{\text{map}})^{-1}$. So, we first apply a fast method to search for $\mathbf{s}_{\text{map}}$.

A. Search for the Most Likely SP

We search for $\mathbf{s}_{\text{map}}$ by selecting $\mathbf{s} \in \mathbf{S}$ with the largest posterior probability $p(\mathbf{s}|\mathbf{y}_v)$. According to the Bayesian rule, the posterior probability can be written as

$$p(\mathbf{s}|\mathbf{y}_v) = \frac{p(\mathbf{y}_v|\mathbf{s})p(\mathbf{s})}{p(\mathbf{y}_v)}, \tag{15}$$

where $p(\mathbf{s}|\mathbf{y}_v)$ are equal to $p(\mathbf{y}_v|\mathbf{s})p(\mathbf{s})$ up to a scale. For convenience, we work in logarithm domain and define $\alpha(\mathbf{s}, \mathbf{y}_v)$ as SP selection metric:

$$\begin{aligned} \alpha(\mathbf{s}, \mathbf{y}_v) &\triangleq \ln p(\mathbf{y}_v|\mathbf{s})p(\mathbf{s}) \\ &= \ln p(\mathbf{y}_v|\mathbf{s}) + \sum_{n=0}^{N-1} \ln p(s_n) \\ &= -\big(\mathbf{y}_v - \mathbf{Q}\boldsymbol{\mu}(\mathbf{s})\big)^H \boldsymbol{\Phi}(\mathbf{s})^{-1} \big(\mathbf{y}_v - \mathbf{Q}\boldsymbol{\mu}(\mathbf{s})\big) \\ &\quad - \ln\det\big(\boldsymbol{\Phi}(\mathbf{s})\big) - M\ln\pi + \sum_{n=0}^{N-1} \ln \lambda_{s_n}. \end{aligned} \tag{16}$$

The largest $p(\mathbf{s}|\mathbf{y}_v)$ corresponds to the largest value of $\alpha(\mathbf{s}, \mathbf{y}_v)$. So we search $\mathbf{s}_{\text{map}}$ based on metric $\alpha(\mathbf{s}, \mathbf{y}_v)$ using a non-exhaustive search tree method.

The search starts with $\mathbf{s} = \mathbf{0}$ at Layer 0. We change only one element from 0 to 1 in $\mathbf{s}$ which leads to N different 'one non-zero element' SPs in Layer 1. We calculate the metric $\alpha(\mathbf{s})$ for all SPs at Layer 1 and store the SP with the largest metric as $\mathbf{S}_1$. For Layer 2, we activate one more element from $\mathbf{S}_1$ so that we have $N - 1$ possible 'two-element active' SPs. Again, we calculate the metrics for SPs at Layer 2 and store the SP with the largest metric as $\mathbf{S}_2$. This procedure is repeated K times to get the 'K-element active' SP with the largest posterior possibility as the $\mathbf{s}_{map}$. An example of the non-exhaustive search tree method is shown in Fig. 2, where $N = 5$ and $K = 3$.

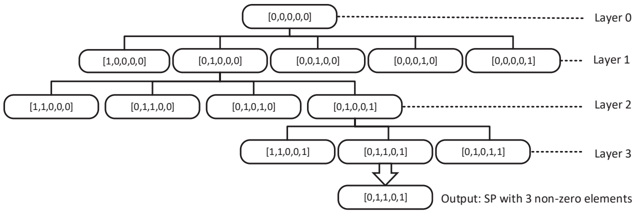

Fig. 2. Example of the non-exhaustive search tree ($K = 3$, $N = 5$).

However, we don't know the real sparsity of mmWave channel so that we can not determine the proper value of K. K should be a little larger than the real sparsity to ensure that we have enough nonzero elements in SPs. In fact, we can stop the searching after enough layers or set a reasonable threshold to stop. It means that we need a rough priori information of sparsity. For mmWave communication, we know that the real sparsity for mmWave channel is generally less than 10 based on real world measurements [1]. So it is reasonable to set a fixed number of layers to stop and we only need to search for the most likely SP instead of the absolutely correct SP.

In this case, we introduce a virtual sparsity L' according to specific applications. For mmWave CE, we choose $L' = 5$ considering that the real sparsity is generally less than 10 . Then we calculate the non-zero probability λ_1 with virtual sparsity which is $L^{'}/N$. Because $L^{'}$ follows Binomial (N, λ_1) distribution, it is common to use the approximation $L^{'} \sim \mathcal{N}(N\lambda_1, N\lambda_1(1 - \lambda_1))$, in which case $Pr(L^{'} > K) = \frac{1}{2}erfc(\frac{K - N\lambda_1}{\sqrt{2N\lambda_1(1-\lambda_1)}})$. Through choosing a very small target value of $Pr(L^{'} > K)$ as $K_0 = 0.01$, we can find the proper value of K as $\lceil \text{erfc}^{-1}(2K_0)\sqrt{2N\lambda_1(1-\lambda_1)} + N\lambda_1 \rceil$. The use of pre-determined virtual sparsity L' provides superior performance with low complexity without the need to know real sparsity. Note that, it may induce degraded performance when the real sparsity is much bigger than the virtual sparsity. We will solve this problem in Section IV.

B. Fast Metric Update

Another primary challenge in the computation of (13) is the high computational load for calculating $\mathbf{\Phi}(\mathbf{s}_{\text{M}})^{-1}$. We adopt the fast metric update method that we proposed in [8] to reduce the computational complexity.

In our search tree, the search begins from the root node ($\mathbf{S}_0 = \mathbf{0}$) which has the following metric

$$\alpha(\mathbf{0}, \mathbf{y}_v) = -\frac{1}{\sigma_n^2}\|\mathbf{y}_v\|_2^2 - M\ln\sigma_n^2 - M\ln\pi + N\ln\lambda_0. \quad (17)$$

We notice that the candidate SPs in the kth layer have only one additional non-zero element compared with the chosen SP $\mathbf{S}_K$ in the $(k-1)$th layer. This characteristic can be used to derive fast metric update. We use $[\mathbf{s}]_n$ to represent the value of the nth element in $\mathbf{s}$. For the case that $[\mathbf{s}]_n = 0$ and $[\mathbf{s}']_n = 1$, where $\mathbf{s}$ and $\mathbf{s}'$ are identical except for the nth coefficient, we describe an efficient method to compute $\Delta_{n,\delta}(\mathbf{s}, \mathbf{y}_v) \triangleq \alpha(\mathbf{s}', \mathbf{y}_v) - \alpha(\mathbf{s}, \mathbf{y}_v)$. For brevity, we define $\mu_\delta \triangleq \mu_1 - \mu_0$ and $\sigma_\delta^2 \triangleq \sigma_1^2 - \sigma_0^2$. To derive the fast metric update, starting with property

$$\mathbf{\Phi}(\mathbf{s}') = \mathbf{\Phi}(\mathbf{s}) + \sigma_\delta^2\mathbf{q}_n\mathbf{q}_n^H, \quad (18)$$

where $\mathbf{q}_n$ is the nth column of $\mathbf{Q}$. The matrix inversion lemma implies

$$\mathbf{\Phi}(\mathbf{s}')^{-1} = \mathbf{\Phi}(\mathbf{s})^{-1} - \beta_n\mathbf{c}_n\mathbf{c}_n^H, \quad (19)$$

$$\mathbf{c}_n \triangleq \mathbf{\Phi}(\mathbf{s})^{-1}\mathbf{q}_n, \quad (20)$$

$$\beta_n \triangleq \sigma_\delta^2(1 + \sigma_\delta^2\mathbf{q}_n^H\mathbf{c}_n)^{-1}. \quad (21)$$

According to [8], (18)-(21) imply

$$\begin{aligned}\Delta_{n,\delta}(\mathbf{s}, \mathbf{y}_v) = \beta_n \big| \mathbf{c}_n^H\big(\mathbf{y}_v - \mathbf{Q}\boldsymbol{\mu}(\mathbf{s})\big) + \mu_\delta/\sigma_\delta^2 \big| \\ - |\mu_\delta|^2/\sigma_\delta^2 + \ln(\beta_n/\sigma_\delta^2) \\ + \ln(\lambda_1/\lambda_0),\end{aligned} \quad (22)$$

where $\Delta_{n,\delta}(\mathbf{s}, \mathbf{y}_v)$ quantifies the change to $\alpha(\mathbf{s}, \mathbf{y}_v)$ corresponding to the change of the nth index in $\mathbf{s}$ from 0 to 1. And then we can work out the metric for $\mathbf{s}'$ as $\alpha(\mathbf{s}, \mathbf{y}_v) + \Delta_{n,\delta}(\mathbf{s}, \mathbf{y}_v)$.

Even though, the complexity still remains high. The main reason is that $\mathbf{c_n}$ needs $O(M^2)$ operations using standard matrix multiplication according to (20). We further reduce this complexity to $O(M)$ by making use of the structure of $\mathbf{\Phi}(\mathbf{s})^{-1}$.

Assuming that $\mathbf{s}$ is the SP which is identical with $\mathbf{s_{pre}}$ but with one more active element at the ${n_{\text{pre}}}^{th}$ coefficient, if we have computed and stored the corresponding parameters for $\mathbf{s}_{\text{pre}}$ as $\beta_{n_{\text{pre}}}$ and $\mathbf{c}_{n_{\text{pre}}}$, (20)-(21) imply that

$$\begin{aligned}\mathbf{c}_n &= [\mathbf{\Phi}(\mathbf{s})_{\text{pre}}^{-1} - \beta_{n_{\text{pre}}}\mathbf{c}_{n_{\text{pre}}}\mathbf{c}_{n_{\text{pre}}}^H]\mathbf{q}_n \\ &= \mathbf{c}_{n_{\text{pre}}} - \beta_{n_{\text{pre}}}\mathbf{c}_{n_{\text{pre}}}\mathbf{c}_{n_{\text{pre}}}^H\mathbf{q}_n.\end{aligned} \quad (23)$$

Comparing (23) and (20), we can successfully reduce complexity by M times via making use of the stored $\mathbf{c}_{n_{\text{pre}}}$. Accordingly, $\mathbf{z}(\mathbf{s}) \triangleq \mathbf{y} - \mathbf{Q}\mu(\mathbf{s})$ can be recursively updated as

$$\mathbf{z}(\mathbf{s}) = \mathbf{y} - \mathbf{Q}\mu(\mathbf{s}_{\text{pre}}) - \mathbf{q}_{n_{\text{pre}}}\mu_\delta. \quad (24)$$

If we define $\mathbf{C} \triangleq [\mathbf{c}_0, \ldots, \mathbf{c}_{N-1}]$, and have computed $\{\mathbf{c}_n\}_{n=0}^{N-1}$ and $\{\beta_n\}_{n=0}^{N-1}$, (14) can be represented as

$$\text{E}\{\mathbf{h}|\mathbf{y}_v, \mathbf{s}_{\text{map}}\} = \boldsymbol{\mu}(\mathbf{s}_{\text{map}}) + \mathbf{R}(\mathbf{s}_{\text{map}})\mathbf{C}^H\mathbf{z}(\mathbf{s}_{\text{map}}), \quad (25)$$

because $\mathbf{C} = \mathbf{\Phi}(\mathbf{s}_{\text{map}})^{-\mathbf{1}}\mathbf{Q}$ and $\mathbf{\Phi}(\mathbf{s}_{\text{map}})$ is Hermitian.

C. Computational Complexity Analysis

In summary, we speed up the algorithm by reducing FBMP [8] from a MMSE estimator with multiple candidate SPs to a MAP based single SP estimation with some appropriate assumptions. The proposed algorithm is shown in Algorithm 1 and named as IBMP. When the search ends, the algorithm returns the estimation of $\mathbf{h}$ based on (14). In contrast to IBMP which only selects one candidate SP $\mathbf{s}_{map}$, multiple (D) 'K-element' candidate SPs $\mathbf{s}_{mmse}$ are selected in FBMP for the MMSE estimator and thus the computational complexity is D times higher than IBMP. As a result, it is straightforward to find that the number of multiplications required by IBMP and FBMP are $\mathcal{O}(NMK)$ and $\mathcal{O}(NMKD)$, respectively, where $D \geq 5$ [27]. Comparisons with other algorithms will presented in Section V.

Algorithm 1 Matching Pursuit Based on MAP

```
$\alpha^{\text{root}} = -\frac{1}{\sigma_n^2}\|\mathbf{y}_v\|_2^2 - M\ln\sigma_n^2 - M\ln\pi + N\ln\lambda_0$,
for $n = 0 : N-1$ do
    $\mathbf{c}_n^{(0)} = \frac{1}{\sigma_n^2}\mathbf{q}_n$, $\beta_n^{(0)} = \sigma_1^2(1 + \sigma_1^2\mathbf{q}_n^H\mathbf{c}_n^{(0)})^{-1}$,
    $\alpha_n^{(1)} = \alpha^{\text{root}} + \ln\frac{\beta_n^{(0)}}{\sigma_1^2} + \beta_n^{(0)}|(\mathbf{c}_n^{(0)})^H\mathbf{y}_v|^2 + \ln\frac{\lambda_1}{\lambda_0}$,
end for
n=[], $\hat{\mathbf{s}}^{(0)} = \mathbf{0}$, $\mathbf{z} = \mathbf{y}_v$,
for $k = 1 : K$ do
    $n_* = n$ indexing the largest element in $\{\alpha_n^{(k)}\}_{n=0:N-1}$ which leads to an unexplored node,
    $\alpha^{(k)} = \alpha_{n_*}^{(k)}$, $\hat{\mathbf{s}}^{(k)} = \hat{\mathbf{s}}^{(k-1)} + \boldsymbol{\delta}_{[n_*]}$, $\mathbf{n} = [\mathbf{n}, n_*]$,
    while $k < K$ do
        for $n = 0 : N-1$ do
            $\mathbf{c}_n^{(k)} = \mathbf{c}_n^{(k-1)} - \beta_{n_*}^{(k-1)}\mathbf{c}_{n_*}^{(k-1)}(\mathbf{c}_{n_*}^{(k-1)})^H\mathbf{q}_n$,
            $\beta_n^{(k)} = \sigma_1^2(1 + \sigma_1^2\mathbf{q}_n^H\mathbf{c}_n^{(k)})^{-1}$,
            $\alpha_n^{(k+1)} = \alpha^{(k)} + \ln\frac{\beta_n^{(k)}}{\sigma_1^2} + \beta_n^{(k)}|(\mathbf{c}_n^{(k)})^H\mathbf{z}|^2 + \ln\frac{\lambda_1}{\lambda_0}$,
        end for
    end while
end for
$\mathbf{h} = \sum_{k=1}^{K}[\sigma_1^2\mathbf{c}_{[\mathbf{n}]_k}^H\mathbf{z}]$.
```

D. Disadvantages

Although IBMP is able to reduce the complexity of FBMP, it has some similar disadvantages as the FBMP. For example, both IBMP and FBMP have error floor at high SNRs. In addition, both IBMP and FBMP are not reliable when the selected 'virtual sparsity' is too different with the real sparsity.

In Section IV, the causes of the problems will be analysed and an off-grid error mitigation method will be proposed to overcome the problems including the need of sparsity information and the performance degradation at high SNRs.

IV. Proposed Off-grid Improved Bayesian Matching Pursuit Algorithm

A. Theoretical Analysis of the Performance Deterioration

In order to overcome the disadvantages of IBMP, we first focus on the analysis of the performance deterioration at high SNRs.

According to model (9), we define $\mathbf{h}_s$ which is the sub-vectors of $\mathbf{h}$ containing only the non-zeros elements. And $\mathbf{Q}_s$ is defined as the sub-matrix consisting of columns of the matrix $\mathbf{Q}$ corresponding to $\mathbf{h}_s$. (9) can be rewritten as

$$\mathbf{y}_v = \mathbf{Q}_s \mathbf{h}_s + \mathbf{E}. \tag{26}$$

When the off-grid error is ignored in the IBMP, $\mathbf{E}$ is the same as $\mathbf{n}_Q$ in (9). The metric in (16) can be represented as

$$\begin{aligned}\alpha(\mathbf{s}, \mathbf{y}_v) = & -{\mathbf{y}_v}^H \mathbf{\Phi}(\mathbf{s})^{-1} \mathbf{y}_v - \ln\det\big(\mathbf{\Phi}(\mathbf{s})\big) \\ & - M \ln \pi + \sum_{n=0}^{N-1} \ln \lambda_{s_n},\end{aligned} \tag{27}$$

where $\mathbf{y}_v$, M and N are known and unchanged. And it can be found that $\mathbf{\Phi}(\mathbf{s})$ is essential for metric comparison at each layer. According to (12), $\mathbf{\Phi}(\mathbf{s}) \triangleq \sigma_1^2 \mathbf{Q}_s \mathbf{Q}_s^H + \sigma_n^2 \mathbf{I}_M$. At high SNRs, $\ln\det\big(\mathbf{\Phi}(\mathbf{s})\big)$ can be approximated as $\ln\det\big(\sigma_1^2 \mathbf{Q}_s \mathbf{Q}_s^H\big)$ because σ_n^2 is extremely small. Considering that the column vectors of matrix $\mathbf{Q}_s \mathbf{Q}_s^H$ are linearly related, $\det\big(\sigma_1^2 \mathbf{Q}_s \mathbf{Q}_s^H\big) = 0$. As a result, $-\ln\det\big(\mathbf{\Phi}(\mathbf{s})\big)$ tends to infinity. And the metrics $\alpha(\mathbf{s}, \mathbf{y}_v)$ in (27) turns to be infinity with respect to any different supports. Hence the metrics based support selection does not function properly and this causes performance deterioration at high SNRs.

Another reason limiting the performance is the ignorance of the off-grid errors. Considering that the true continuous AoDs/AoAs may lie off the grid, (26) can be written as

$$\mathbf{y}_v = \mathbf{Q}_s \mathbf{h}_s + \mathbf{n}_Q + \mathbf{n}_e, \tag{28}$$

where $\mathbf{n}_e$ represents the off-grid error. Because the continuous AoDs/AoAs are independent with noise and follow uniform distribution, after uniform quantization, the off-grid error should also follow uniform distribution with fixed variance σ_e^2 which only depends on the grid size G. It means that $\mathbf{n}_e$ is always a non-zero value and doesn't decrease as $\mathbf{n}_Q$ does at high SNRs. As a result, at high SNR where $\mathbf{n}_Q$ decreases and $\mathbf{n}_e$ dominants, ignoring the off-grid error causes problem.

B. Proposed Solution

Because the off-grid error can not be reduced to a certain extent by increasing SNR or increasing resolution of the grid [18], we modify the existing grid points according to the off-grid error as

$$\mathbf{y}_v = \mathbf{Q}_s \mathbf{h}_s + \mathbf{n}_Q + \mathbf{n}_e = \hat{\mathbf{Q}}_s \mathbf{h}_s + \mathbf{n}_Q, \tag{29}$$

where $\hat{\mathbf{Q}}_s$ is the new sensing matrix with 'modified grid points'. If we take into account the quantification error, it is possible to find out a more accurate estimated AoD/AoA pair AoD'/AoA' around the initial estimated AoD/AoA by maximizing metric. In this paper, we choose to employ the sequential quadratic programming (SQP) method. In this way, off-grid errors can be mitigated. Incorporating this method into IBMP, we propose the OG-IBMP algorithm to solve the off-grid problem, as summarised in Algorithm 2 and explained below.

In the initial stage, root metric α^{root}, $\mathbf{c}_n^{(0)}$ and $\beta_n^{(0)}$ for layer 0 are calculated so that we can obtain candidate metrics for layer 1 as $\alpha_n^{(1)}$. Iteration begins for layer $k = 1 : K$. In each iteration, we choose the largest metric from candidate metrics $\alpha_n^{(k)}$ for layer k and store the index as n_*. Then the 4-step off-grid mitigation begins.

In the first step, we estimate the initial value of AoD/AoA using the column index n_*. Specifically, $\mathbf{A}_D = \bar{\mathbf{A}}_T^* \otimes \bar{\mathbf{A}}_R$ is an $N_T N_R \times G^2$ dictionary matrix that consists of G^2 column vectors. And the $\big(G(u-1)+v\big)$th column is calculated using $\mathbf{a}_t^*(\theta_u) \otimes \mathbf{a}_r(\theta_v)$, where θ_u and θ_v are the uth and vth discrete points of the uniform angle grid, respectively. As a result, the estimated initial AoD/AoA are $AoD_k = 0 + \text{ceil}(\frac{n_*}{G})\frac{\pi}{G-1}$ and $AoA_k = 0 + (\text{mod}(n_* - 1, G) + 1)\frac{\pi}{G-1}$, where $u = \text{ceil}(\frac{j}{G})$, $v = \text{mod}(n_* - 1, G) + 1$.

In step 2, we set $x_k = (AoD_k, AoA_k)$ as the original point corresponding to the ${n_*}^{th}$ column in $\mathbf{Q}$. And we define the objective function for optimization as f with

$$\mathbf{q}_{n_*} = \big(\mathbf{F}^T \otimes \mathbf{W}^H\big)\left(\mathbf{a}^*\left(AoD_k\right) \otimes \mathbf{a}\left(AoA_k\right)\right), \tag{30}$$

$$\mathbf{c}_{n_*}^{(k-1)} = \mathbf{\Phi}(\hat{\mathbf{s}}^{(k-1)})^{-1} \mathbf{q}_{n_*}, \tag{31}$$

$$\beta_{n_*}^{(k-1)} = \sigma_1^2 (1 + \sigma_1^2 \mathbf{q}_{n_*}^H \mathbf{c}_{n_*}^{(k-1)})^{-1}, \tag{32}$$

$$\alpha_{n_*}^{(k)} = \ln \frac{\beta_{n_*}^{(k-1)}}{\sigma_1^2} + \beta_{n_*}^{(k-1)} |(\mathbf{c}_{n_*}^{(k-1)})^H \mathbf{z}|^2 + \ln \frac{\lambda_1}{\lambda_0}, \tag{33}$$

$$f = -\alpha_{n_*}^{(k)}. \tag{34}$$

Through minimizing the objective function f between the adjacent grid points, we can obtain new angle pair $x'_k = (AoD'_k, AoA'_k)$ which results in the largest metric $\alpha_{n_*}^{(k)}$. This optimization problem based on (IV-B)-(34) is formulated as

$$\begin{aligned} & \min_{AoD'_k, AoA'_k} \quad f(AoD'_k, AoA'_k), \\ & s.t. \quad \begin{cases} |AoD'_k - AoD_k| < \frac{\pi}{2(G-1)}, \\ |AoA'_k - AoA_k| < \frac{\pi}{2(G-1)}. \end{cases} \end{aligned} \tag{35}$$

SQP method is adopted because it is proved to be highly effective for solving constrained optimization problems with smooth nonlinear objective function and constraints [28]. In step 3, after obtaining x'_k using SQP method, we refine the grid point by adjusting the corresponding dictionary vector $\mathbf{A}_D$, so that the column indexed by n_* is updated as (IV-B) as $\mathbf{q}_{n_*} = \big(\mathbf{F}^T \otimes \mathbf{W}^H\big)\left(\mathbf{a}^*\left(AoD'_k\right) \otimes \mathbf{a}\left(AoA'_k\right)\right)$. When the grid points are adjusted towards the continuous true angle point, off-grid impact is reduced.

In step 4, we update $\mathbf{c}_{n_*}^{(k-1)}$, $\beta_{n_*}^{(k-1)}$, $\alpha_{n_*}^{(k)}$ as (31)-(33) based on the updated $\bar{\mathbf{q}}_{n_*}$ from step3. $\alpha_{n_*}^{(k)}$ is the optimized largest

Algorithm 2 Off-Grid Improved Bayesian Matching Pursuit

```
$\alpha^{\text{root}} = -\frac{1}{\sigma_n^2}\|\mathbf{y}_v\|_2^2 - M\ln\sigma_n^2 - M\ln\pi + N\ln\lambda_0$,
for $n = 0 : N-1$ do
    $\boldsymbol{\Phi}(\hat{\mathbf{s}}^{(0)})^{-1} = (\sigma_n^2\mathbf{I_M})^{-1}$,
    $\mathbf{c}_n^{(0)} = \frac{1}{\sigma_n^2}\mathbf{q}_n$, $\beta_n^{(0)} = \sigma_1^2(1+\sigma_1^2\mathbf{q}_n^H\mathbf{c}_n^{(0)})^{-1}$,
    $\alpha_n^{(1)} = \alpha^{\text{root}} + \ln\frac{\beta_n^{(0)}}{\sigma_1^2} + \beta_n^{(0)}|(\mathbf{c}_n^{(0)})^H\mathbf{y}_v|^2 + \ln\frac{\lambda_1}{\lambda_0}$,
end for
n=[], $\hat{\mathbf{s}}^{(0)} = \mathbf{0}$, $\mathbf{z} = \mathbf{y}_v$,
for $k = 1 : K$ do
    $n_* = n$ indexing the largest element in $\{\alpha_n^{(k)}\}_{n=0:N-1}$
    which leads to an unexplored node,

    *Off-grid mitigation begins
    1: $AoD_k = 0 + \text{ceil}(\frac{n_*}{G})\frac{\pi}{G-1}$
       $AoA_k = 0 + (\text{mod}(n_*-1, G)+1)\frac{\pi}{G-1}$
       $x_k = (AoD_k, AoA_k)$
    2: $\min\limits_{AoD'_k, AoA'_k} f(AoD'_k, AoA'_k)$
       output: $x'_k = (AoD'_k, AoA'_k)$, $F = f(x_k^{'})$
    3: $\bar{\mathbf{q}}_{n_*} = \left(\mathbf{F}^T \otimes \mathbf{W}^H\right)\left(\mathbf{a}^*\left(AoD'_k\right) \otimes \mathbf{a}\left(AoA'_k\right)\right)$
    4: update $\mathbf{c}_{n_*}^{(k-1)}$, $\beta_{n_*}^{(k-1)}$ and optimized $\alpha_{n_*}^{(k)}$
    *Off-grid mitigation ends

    $\alpha^{(k)} = \alpha_{n_*}^{(k)}$, $\hat{\mathbf{s}}^{(k)} = \hat{\mathbf{s}}^{(k-1)} + \boldsymbol{\delta}_{[n_*]}$, $\mathbf{n} = [\mathbf{n}, n_*]$
    $\boldsymbol{\Phi}(\hat{\mathbf{s}}^{(k-1)})^{-1} = \boldsymbol{\Phi}(\hat{\mathbf{s}}^{(k-1)})^{-1} - \beta_{n_*}^{(k-1)}\mathbf{c}_{n_*}^{(k-1)}(\mathbf{c}_{n_*}^{(k-1)})^H$
    while $k < K$ do
        for $n = 0 : N-1$ do
            $\mathbf{c}_n^{(k)} = \mathbf{c}_n^{(k-1)} - \beta_{n_*}^{(k-1)}\mathbf{c}_{n_*}^{(k-1)}(\mathbf{c}_{n_*}^{(k-1)})^H\mathbf{q}_n$,
            $\beta_n^{(k)} = \sigma_1^2(1+\sigma_1^2\mathbf{q}_n^H\mathbf{c}_n^{(k)})^{-1}$,
            $\alpha_n^{(k+1)} = \alpha^{(k)} + \ln\frac{\beta_n^{(k)}}{\sigma_1^2} + \beta_n^{(k)}|(\mathbf{c}_n^{(k)})^H\mathbf{z}|^2 + \ln\frac{\lambda_1}{\lambda_0}$
        end for
    end while
end for
$\hat{\mathbf{h}} = \sum_{k=1}^K[\sigma_1^2\mathbf{c}_{[\mathbf{n}]_k}^H\mathbf{z}]$
```

metric with updated grid points for $\hat{\mathbf{s}}^{(k-1)}$. Till now, off-grid mitigation is completed and it is ready to calculate $\alpha_n^{(k+1)}$ for the next layer.

We continue the iteration as IBMP does for all the K layers and find the $\hat{\mathbf{s}}^{(K)}$ with the largest metric as the estimated SP. Finally, the estimated channel matrix $\hat{\mathbf{h}}$ can be obtained using (25). It is worth noting that the complexity of the optimization process can be reduced by reuse the calculated parameters. For example, in the third step in Algorithm 2, the n_* column in the sensing matrix $\mathbf{Q}$ is updated as , where $\mathbf{F}^T \otimes \mathbf{W}^H$ remains the same in each iteration and can be obtained in the calculation of the sensing matrix $\mathbf{Q}$ according to (9).

C. Convergence Analysis

Considering that the proposed OG-IBMP algorithm is an iterative optimization based method, the convergence analysis of the OG-IBMP algorithm is described in this subsection. Fig. 3 first investigates the convergence of the optimization method for each layer in the proposed algorithm. Then, the impact of the varying iteration numbers on the channel estimation at different SNR is presented in Fig. 4.

We assume that each estimated SP has K non-zero elements which are selected in K layers. For each element selection, SQP method is used to optimize the estimated angles iteratively. The changes of the objective function of the k-th layer in the i-th iteration is measured by $\mathbf{v}_k(i)$ as

$$\mathbf{v}_k(i) = |\mathbf{f}_k(i) - \mathbf{f}_k(i-1)|, \tag{36}$$

where $\mathbf{f}_k(i)$ denotes the objective function of the k-th layer in the i-th iteration. In Fig. 3, we show the $\mathbf{v}_k(i)$ from layer $k=1$ to layer $k=7$ ($K=7$) with the increasing number i of the iterations when SNR $= 0\,$dB. The iteration process stops when the $\mathbf{v}_k(i)$ is less than 10^{-6}. The parameters are summarized in Table I with $\sigma_1^2 = 100, \sigma_n^2 = P_r/10$ and the simulation results are averaged over 500 channel realizations.

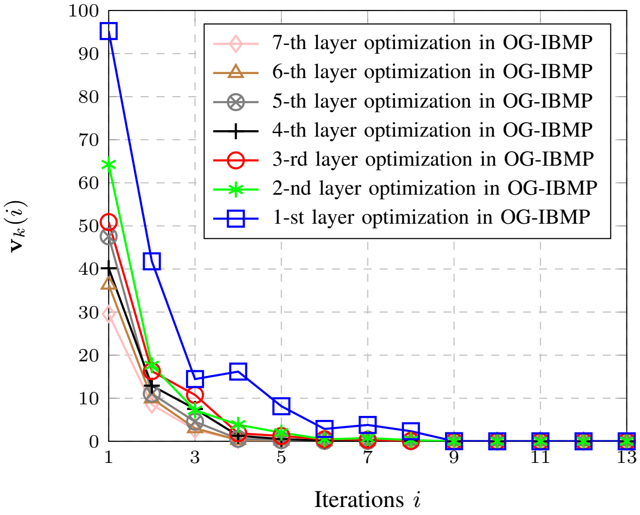

Fig. 3. Metric changes of different layers with increasing iterations.

As shown in Fig. 3, the iterations of all the layers stop ($\mathbf{v}_k(i) < 10^{-6}$) after 13 iterations. It can be found that the later layer converges more quickly and less iterations are required for a desirable optimization, because the later layer is based on the result of the previous layer, which has been improved by the optimizations. Specifically, in the 1-st layer, 9 iterations are needed to achieve $\mathbf{v}_k(i) < 1$. Compared with the 1-st layer, $\mathbf{v}_k(i)$ can be reduced to 0.301953 with only 4 iterations in the 7-th layer. Fig. 3 proves that the proposed algorithm converges quickly, even at the first layer with large noise.

The impact of the number of iterations on the channel estimation at different SNR is evaluated via simulations in terms of the normalized mean square error (NMSE) which is defined as $10\log_{10}\left(\text{E}(\|\mathbf{H}-\mathbf{H}^{\text{estimate}}\|_F^2/\|\mathbf{H}\|_F^2)\right)$. The proposed algorithms are named as OG-S-IBMP and OG-L-IBMP with the virtual sparsity $L_1 = 5$ and $L_2 = 10$, respectively. As expected, the estimation accuracy tends to be the same when the number of iterations increases. In Fig. 4, OG-S-IBMP and OG-L-IBMP with more than 20 iterations achieve almost the same NMSE performance at all SNRs.

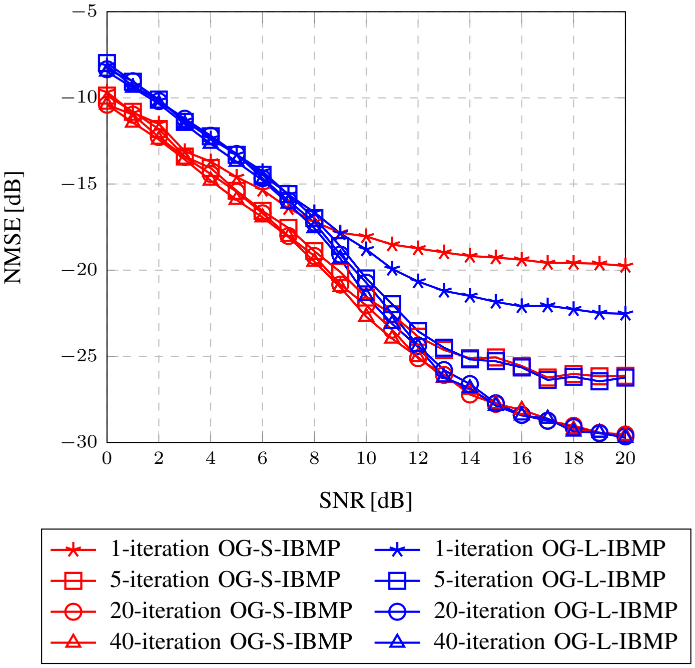

Fig. 4. NMSE vs SNR with different iterations.

In summary, based on the above simulations, it can be found that the proposed OG-IBMP algorithm is able to converge in limited number of iterations. In addition, with enough iterations, the proposed algorithm can achieve almost the same best estimation accuracy.

V. Simulation Results

The performance of the proposed methods IBMP and OG-IBMP are evaluated via computer simulations in terms of NMSE. ULAs are assumed at both transmitter and receiver. All simulation results are averaged over 500 channel realizations with carrier frequency at 60 GHz. At each channel realization, the system parameters are listed in Table I.

TABLE I
System parameters in the simulations.

Parameters	Values
Number of antennas: (N_T, N_R)	(32, 32)
Number of beams: (N_T^B, N_R^B)	(32, 32)
Number of scatterers: L	7
Virtual Sparsity: (L_1, L_2)	(5, 10)
Channel gains: $\{\alpha_\ell\}_{\ell=1}^L$	$\mathcal{CN}(0, 1)$
AoD/AoA: $\{\theta_l^t\}_{\ell=1}^L, \{\theta_l^r\}_{\ell=1}^L$	$\mathcal{U}(0, \pi)$
Grid size: G	64

* In this case, the off-grid error follow uniform distribution $\mathcal{U}(0, \frac{\pi}{2(G-1)})$ with $\sigma_e^2 = (\frac{\pi}{2(G-1)})^2/12$.

The design of hybrid precoding and combining matrices have been extensively investigated, so we adopt the precoder and combiner presented in [29]. $\mathbf{F} = (\mathbf{\Lambda}_F^{-1/2}\mathbf{U}_F^H)^T$ where $\mathbf{U}_F$ and $\mathbf{\Lambda}_F$ are the matrices of the eigenvectors and eigenvalues of $\mathbf{A}_T^*(\mathbf{A}_T^*)^H$. $\mathbf{W} = (\mathbf{\Lambda}_W^{-1/2}\mathbf{U}_W^H)^H$ where $\mathbf{U}_W\mathbf{\Lambda}_W\mathbf{U}_W^H = \mathbf{A}_R(\mathbf{A}_R)^H$. For our proposed IBMP method, two different 'virtual sparsity' ($L_1 = 10$ and $L_2 = 5$) are considered as L-IBMP and S-IBMP respectively. The proposed IBMP algorithm with off-grid mitigation method is named as OG-IBMP. BCS is included for comparison because of its state-of-the-art performance. Note that the true noise power is given to BCS to decrease the huge complexity to comparable level with other algorithms.

In order to reduce the complexity in our application, we first investigate the impact of σ_1^2 (variance of the Gaussian distributions assumption indexed by $s_n = 1$) and σ_n^2 (variance of the noise) so that we can choose the value of parameters accurately and achieve desirable performance. Simulation results for IBMP with varying σ_1^2 and σ_n^2 are presented in Fig. 5 and Fig. 6, respectively. Note that, σ_1^2 and σ_n^2 can be estimated by EM algorithm to improve the performance at the cost of complexity.

Specifically, In Fig. 5, we compare the performance with different σ_1^2 and known σ_n^2. σ_1^2 is chosen as 0.1,1 and 100. In fact, we can choose any positive value for σ_1^2. But it is straightforward to find that, with small σ_1^2 and large σ_n^2, value of $\mathbf{\Phi}(\mathbf{s})$ is dominated by the noise and so is the value of the metric. As a result, the support estimation evidently deteriorates at low SNRs. As shown in Fig. 5, IBMP performance is even worse than OMP when σ_1^2 is 0.1 and 1 at low SNRs. On the contrary, choosing larger $\sigma_1^2 = 100$ significantly improves the estimation accuracy. Our simulation based analysis shows that variance larger than 100 would not improve performance further in mmWave CE. Thus, we choose $\sigma_1^2 = 100$ for IBMP and OG-IBMP in our application.

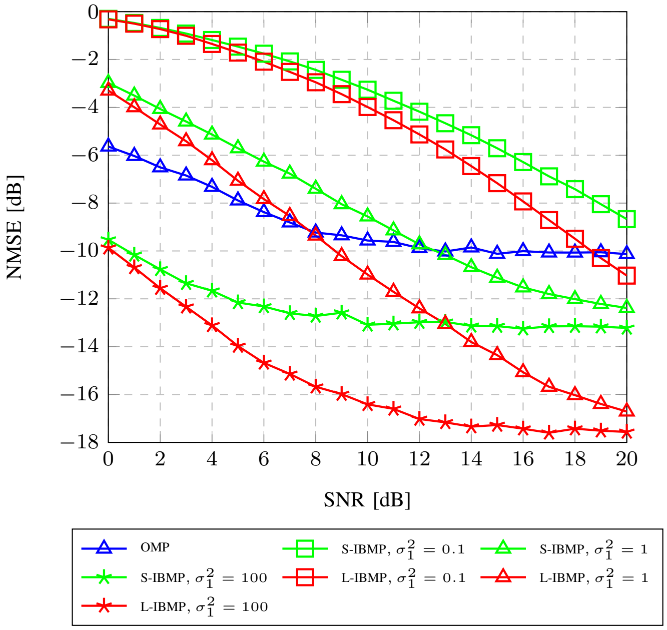

Fig. 5. NMSE [dB] comparison of IBMP at different SNRs [dB] with different σ_1^2 and known σ_n^2.

In Fig. 6, we fix $\sigma_1^2 = 100$ and compare the performance with different σ_n^2. We select σ_n^2 as $P_r/100$, $P_r/10$, and true noise where P_r is the received signal power. Although all IBMP results are better than that of OMP, IBMP with known

σ_n^2 achieves the best performance among them. Without the knowledge of σ_n^2, different assumptions of σ_n^2 do not affect much estimation accuracy. We found that S-IBMP achieves better performance at low SNRs (SNR < 2 dB) and is more noise resistant, but bigger virtual sparsity is required for higher SNRs. This is because the accuracy of CE is affected by both the noise and off-grid errors. At higher SNRs, where off-grid error dominates, additional active elements can help mitigating off-grid error impact and improving the estimation performance. On the contrary, noise dominates at lower SNRs. In such case, it is very difficult to choose extra correct locations of active elements. The increasing number of wrong active locations will lead to even worse performance. Compared the performance of L-IBMP with and without known σ_n^2, there is a significant performance gap at low SNRs. The theoretical analysis also proves that large σ_n^2 degrades the performance of L-IBMP which is more sensitive to noise than S-IBMP. In our application, σ_n^2 is usually unknown so that L-IBMP can not perform well at low SNRs. As a result, for mmWave CE, we adopt S-IBMP with $\sigma_1^2 = 100, \sigma_n^2 = P_r/10$ at low SNRs and L-IBMP with the same σ_1^2 and σ_n^2 at high SNR. In this case, S-IBMP achieves nearly 4 dB improvement at low SNRs compared with OMP, and L-IBMP achieves nearly 7 dB improvement at high SNRs. However, both S-IBMP and L-IBMP have error floors at high SNRs as discussed in Section III, and it is difficult to choose the virtual sparsity without the knowledge of the sparsity and SNRs.

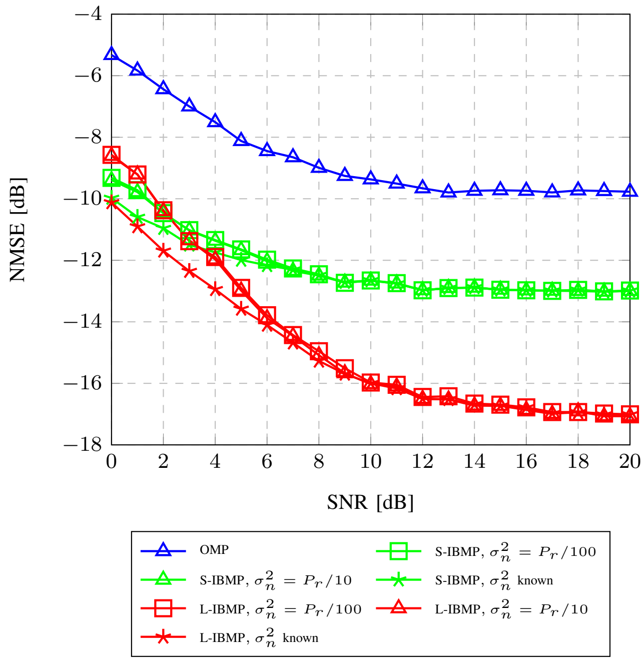

Fig. 6. NMSE [dB] comparison of IBMP at different SNRs [dB] with $\sigma_1^2 = 100$ and different σ_n^2.

Fig. 7 investigates the impact of the off-grid error. The performance of OMP, S-IBMP and L-IBMP with and without the off-grid error are presented. We assume that σ_n^2 is known and $\sigma_1^2 = 100$. For performances without off-grid error, we only generate the AoDs/AoAs at the predetermined grid. And for with off-grid error, we generate the AoDs/AoAs randomly from 0 to π. Clearly, S-IBMP and L-IBMP without off-grid error are able to continuously improve the NMSE performance with the increase of SNR. As proved by the theoretical analysis that mitigating off-grid errors is the way to remove the error floor of the IBMP at high SNRs especially when small virtual sparsity is used.

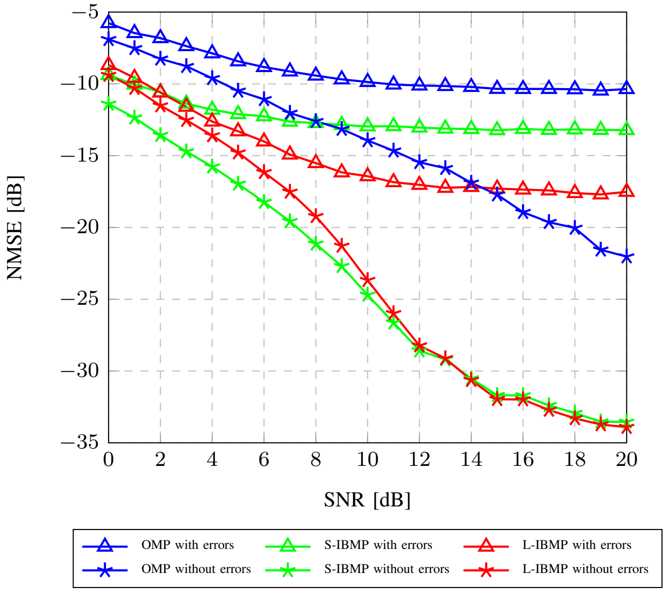

Fig. 7. NMSE [dB] comparison of IBMP at different SNRs [dB] with and without off-grid error.

In Fig. 8, we assume that $\sigma_1^2 = 100, \sigma_n^2 = P_r/10$ and apply the proposed off-grid mitigation methods for S-IBMP and L-IBMP, namely OG-S-IBMP and OG-L-IBMP. OMP [7], BCS [10], S-FBMP [8] and IP-OMP [18] are included for comparison as the representatives of non-Bayesian based method, Bayesian based method and non-Bayesian method with off-grid mitigation. Note that Bayesian learning based method with off-grid mitigation such as improved SBL [23] are not included for comparison considering the several orders of magnitude higher computational complexity compared with other algorithms. Fig. 8 shows that IP-OMP, OG-S-IBMP and OG-L-IBMP all have a better performance because of the integrated off-grid error mitigation, and S-FBMP has almost the same performance as S-IBMP. The improvement of OG-S-IBMP compared to S-IBMP is much bigger than OG-L-IBMP compared to L-IBMP. It is because that the S-IBMP is affected more seriously by the off-grid error due to the smaller number of non-zero elements and hence the off-grid mitigation is more effective. Specifically, OG-L-IBMP is 1-2 dB worse than OG-S-IBMP when SNR is less than 12 dB. When the noise is very small (SNR>12 dB) and off-grid error dominates, OG-L-IBMP can achieve almost the same performance as OG-S-IBMP. We can then conclude that small virtual sparsity should be chosen in any scenarios without the need of a prior information of the noise and sparsity. Compared with the state of art algorithm such as BCS, OG-S-IBMP achieves more than 5 dB improvement performance at all SNRs.

The complexity comparisons are listed in Table II. Specif-

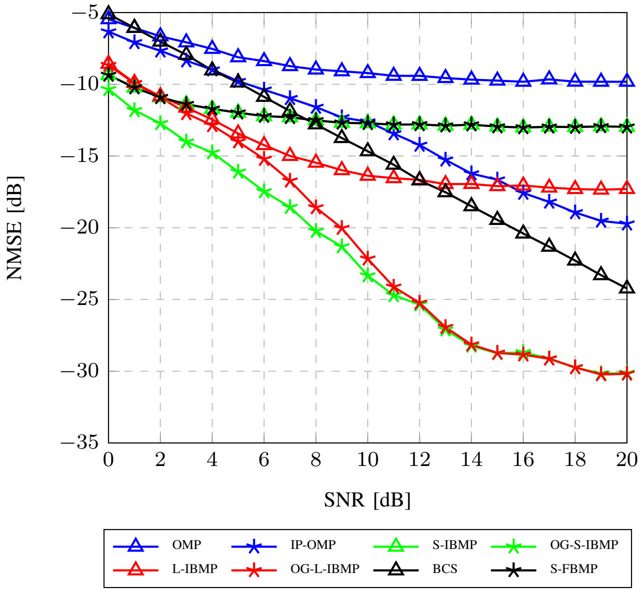

Fig. 8. NMSE [dB] comparison of OG-IBMP at different SNRs [dB] with $\sigma_1^2 = 100$ and $\sigma_n^2 = P_r/10$.

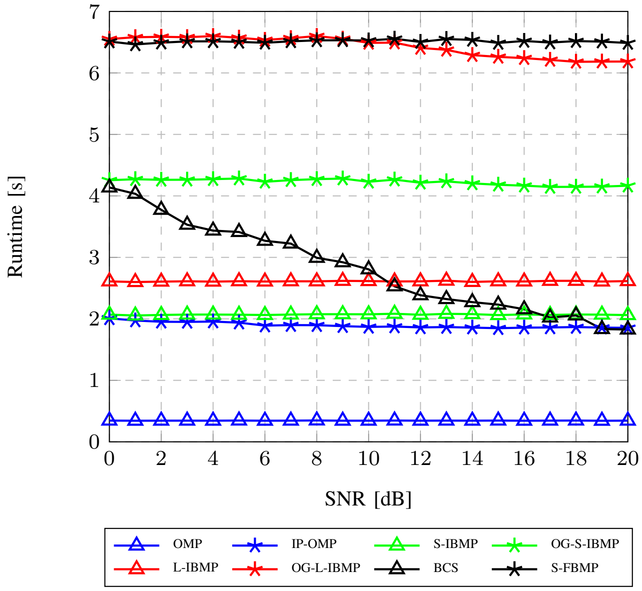

Fig. 9. Runtime [s] of proposed algorithms at different SNRs (dB) with $\sigma_1^2 = 100$ and $\sigma_n^2 = P_r/10$.

ically, the computational complexity of OMP, S-IBMP, L-IBMP, and S-FBMP are $\mathcal{O}(LNM)$, $\mathcal{O}(K_2NM)$, $\mathcal{O}(K_1NM)$ and $\mathcal{O}(NMK_1D)$ [8], respectively, where K_1 and K_2 are calculated based on L_1 and L_2. In the mmWave CE, $L < K_2 < K_1$. Considering BCS is a learning based method and IP-OMP, OG-S-IBMP, OG-L-IBMP are optimization based methods, it is difficult to evaluate the complexity. Fortunately, due to the constrains of the optimization, we find that only a few iterations are required for the optimization based methods. As a result, the computational complexity is mainly determined by the complexity of the original algorithms (i.e. OMP, S-IBMP, L-IBMP) and approximately several times higher [18]. According to [10], BCS algorithm is about an order of magnitude slower than OMP, even when the Adaptive Compressive Sensing option is turned off (i.e. not to include the learning procedure for noise estimation). Therefore, we can anticipate that the learning based algorithms with off-grid mitigation methods are not practical due to the unacceptable computational complexity.

TABLE II
COMPUTATIONAL COMPLEXITY COMPARISON OF DIFFERENT ALGORITHMS.

Method	Computational complexity
OMP	$\mathcal{O}(LNM)$
IP-OMP	several times higher than $\mathcal{O}(LNM)$
S-IBMP	$\mathcal{O}(K_2NM)$
OG-S-IBMP	several times higher than $\mathcal{O}(K_2NM)$
L-IBMP	$\mathcal{O}(K_1NM)$
OG-L-IBMP	several times higher than $\mathcal{O}(K_1NM)$
BCS	an order higher than OMP
S-FBMP	$\mathcal{O}(K_2NMD)$

* $\lambda_1 = L_1/N, \lambda_2 = L_2/N, K_1 = \lceil \mathrm{erfc}^{-1}(2K_0)\sqrt{2N\lambda_1(1-\lambda_1)} + N\lambda_1 \rceil$, $K_2 = \lceil \mathrm{erfc}^{-1}(2K_0)\sqrt{2N\lambda_2(1-\lambda_2)} + N\lambda_2 \rceil$, $K_0 = 0.01$, $D = 5$.

As shown in Fig. 9, we use average runtime of the algorithms to verify the above analysis on the computational complexity. Note that, S-FBMP chooses $D = 5$ candidate SPs and uses the same virtual sparsity as S-IBMP. It is assumed that the noise power is known in BCS so that the complexity can be decreased to a comparable level with other algorithms. Fig. 9 validates our analysis. Specifically, S-IBMP is the fastest Bayesian based algorithm and it is multiple times faster than S-FBMP. OG-S-IBMP is significantly faster than OG-L-IBMP and S-FBMP. OMP is the fastest among all these algorithms. OG-S-IBMP, BCS, L-IBMP, S-IBMP and IP-OMP have comparable runtimes.

In summary, the proposed OG-S-IBMP algorithm can overcome the disadvantages of the S-IBMP algorithm including the error floor at high SNRs and the requirement of a prior information of the sparsity and noise power. OG-S-IBMP algorithm achieves the best performance among the existing algorithms with acceptable complexity.

VI. Conclusion

In this paper, we proposed a fast MAP based method named IBMP for the channel estimation of the mmWave massive MIMO system. We then proposed the OG-S-IBMP to overcome the disadvantages of the IBMP by integrating off-grid mitigation. Specifically, the proposed OG-S-IBMP algorithm starts from the proposed on-grid algorithm IBMP, and iteratively modifies the grid using the SQP method to mitigate the impact of the off-grid angles. Simulation results confirmed that our proposed OG-S-IBMP algorithm outperforms the state-of-the-art mmWave CE methods with low computational complexity. In future works, the proposed schemes can be extended to more complex emerging systems, i.e. reconfigurable intelligent surfaces (RIS) aided system [30], and considering the frequency-selectivity of the mmWave channels.

References

[1] T. S. Rappaport, S. Sun, R. Mayzus *et al.*, “Millimeter wave mobile communications for 5G cellular: It will work!” *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[2] R. W. Heath, N. Gonzalez-Prelcic, Rangan *et al.*, “An overview of signal processing techniques for millimeter wave MIMO systems,” *IEEE J. Sel. Top. Signal Process.*, vol. 10, no. 3, pp. 436–453, Feb. 2016.

[3] P. F. Smulders and L. Correia, “Characterisation of propagation in 60 GHz radio channels,” *Electron. Commun. Eng. J.*, vol. 9, no. 2, pp. 73–80, Apr. 1997.

[4] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[5] A. Alkhateeb, O. El Ayach, G. Leus *et al.*, “Channel estimation and hybrid precoding for millimeter wave cellular systems,” *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 5, pp. 831–846, Jul. 2014.

[6] Z. Xiao, P. Xia, and X.-G. Xia, “Codebook design for millimeter-wave channel estimation with hybrid precoding structure,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 141–153, Oct. 2016.

[7] J. Lee, G.-T. Gil, and Y. H. Lee, “Channel estimation via orthogonal matching pursuit for hybrid MIMO systems in millimeter wave communications,” *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2370–2386, Jun. 2016.

[8] Y. You and L. Zhang, “Bayesian matching pursuit-based channel estimation for millimeter wave communication,” *IEEE Commun. Lett.*, vol. 24, no. 2, pp. 344–348, Nov. 2019.

[9] A. Mishra, A. Rajoriya, A. K. Jagannatham *et al.*, “Sparse bayesian learning-based channel estimation in millimeter wave hybrid MIMO systems,” in *Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun.*, Sapporo, Japan, Jul. 2017, pp. 1–5.

[10] S. Ji, Y. Xue, and L. Carin, “Bayesian compressive sensing,” *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, May 2008.

[11] C. Huang, L. Liu, C. Yuen *et al.*, “Iterative channel estimation using LSE and sparse message passing for mmwave MIMO systems,” *IEEE Trans. Signal Process.*, vol. 67, no. 1, pp. 245–259, Jan. 2019.

[12] K. Venugopal, A. Alkhateeb, R. W. Heath *et al.*, “Time-domain channel estimation for wideband millimeter wave systems with hybrid architecture,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 6493–6497.

[13] J. Rodríguez-Fernández, N. González-Prelcic *et al.*, “Frequency-domain compressive channel estimation for frequency-selective hybrid millimeter wave MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2946–2960, May 2018.

[14] S. Srivastava, A. Mishra, A. Rajoriya *et al.*, “Quasi-static and time-selective channel estimation for block-sparse millimeter wave hybrid MIMO systems: Sparse bayesian learning (SBL) based approaches,” *IEEE Trans. Signal Process.*, vol. 67, no. 5, pp. 1251–1266, Mar. 2019.

[15] Z. Gao, L. Dai, Z. Wang *et al.*, “Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO,” *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169–6183, Dec. 2015.

[16] Z. Gao, L. Dai, S. Han *et al.*, “Compressive sensing techniques for next-generation wireless communications,” *IEEE Trans. Wireless Commun.*, vol. 25, no. 3, pp. 144–153, Jun. 2018.

[17] M. Rossi, A. M. Haimovich, and Y. C. Eldar, “Spatial compressive sensing for MIMO radar,” *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 419–430, Nov. 2013.

[18] Y. You, L. Zhang, and M. Liu, “IP aided OMP based channel estimation for millimeter wave massive MIMO communication,” in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Marrakesh, Morocco, Apr. 2019, pp. 1–6.

[19] G. Tang, B. N. Bhaskar, P. Shah *et al.*, “Compressed sensing off the grid,” *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7465–7490, Aug. 2013.

[20] W. Si, X. Qu, Z. Qu *et al.*, “Off-grid DOA estimation via real-valued sparse bayesian method in compressed sensing,” *Circuits Syst. Signal Process.*, vol. 35, no. 10, pp. 3793–3809, Jan. 2016.

[21] A. C. Gurbuz, Y. Yapici, and I. Guvenc, “Sparse channel estimation in millimeter-wave communications via parameter perturbed OMP,” in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[22] C. K. Anjinappa *et al.*, “Off-grid aware channel and covariance estimation in mmwave networks,” *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3908–3921, Jun. 2020.

[23] H. Tang, J. Wang, and L. He, “Off-grid sparse bayesian learning-based channel estimation for mmwave massive MIMO uplink,” *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 45–48, Jun. 2018.

[24] O. El Ayach, S. Rajagopal, S. Abu-Surra *et al.*, “Spatially sparse precoding in millimeter wave MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Jan. 2014.

[25] A. M. Sayeed, “Deconstructing multiantenna fading channels,” *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Nov. 2002.

[26] H. V. Poor, *An introduction to signal detection and estimation*. New York, NY, USA: Springer-Verlag, 1994.

[27] P. Schniter, L. C. Potter, and J. Ziniel, “Fast bayesian matching pursuit,” in *Proc. Inf. Theory Appl. Workshop*, San Diego, CA, USA, Jan. 2008, pp. 326–333.

[28] P. T. Boggs and J. W. Tolle, “Sequential quadratic programming,” *Acta Numer.*, vol. 4, pp. 1–51, Jan. 1995.

[29] L. Zelnik-Manor, K. Rosenblum, and Y. C. Eldar, “Sensing matrix optimization for block-sparse decoding,” *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4300–4312, Jun. 2011.

[30] C. Huang, Z. Yang, G. C. Alexandropoulos *et al.*, “Multi-hop RIS-empowered terahertz communications: A DRL-based hybrid beamforming design,” *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1663–1677, Jun. 2021.