Estimation of Broadband Multiuser Millimeter-Wave Massive MIMO-OFDM Channels by Exploiting Their Sparse Structure

Xincong Lin, Sheng Wu, Member, IEEE, Chunxiao Jiang, Senior Member, IEEE, Linling Kuang, Member, IEEE, Jian Yan and Lajos Hanzo, Fellow, IEEE

Abstract-In millimeter wave (mmWave) massive multipleinput multiple-output (MIMO) systems, acquiring accurate channel state information is essential for efficient beamforming (BF) and multiuser interference cancellation, which is a challenging task since a low signal-to-noise ratio is encountered before BF in large antenna arrays. The mmWave channel exhibits a 3-D clustered structure in the virtual angle of arrival (AOA), angle of departure (AOD) and delay domain that is imposed by the effect of power leakage, angular spread and cluster duration. We extend the approximate message passing (AMP) with nearest neighbor pattern learning algorithm for improving the attainable channel estimation performance, which adaptively learns and exploits the clustered structure in the 3-D virtual AOA-AOD-delay domain. The proposed method is capable of approaching the performance bound described by the state evolution based on vector AMP framework, and our simulation results verify its superiority in mmWave systems associated with a broad bandwidth.

Index Terms—Approximate message passing, broadband, channel estimation, mmWave, OFDM, sparse structure

I. INTRODUCTION

Communications at millimeter wave (mmWave) frequencies are regarded as a key enabling technique for 5G by exploiting their broad bandwidth. However, mmWave frequencies suffer from a high propagation loss. In order to mitigate the path loss, numerous antenna elements are packed for beamforming (BF). Conventional multiple-input multiple-output (MIMO) BF relies on digital processing which result in extremely highenergy consumption. Hybrid analog / digital precoding [1]–[4] is able to reduce the cost.

Accurate channel estimates are essential for designing analog and digital beamformers [4]. Experiments conducted in indoor [5] and outdoor [6] environments have shown that mmWave channels exhibit sparsity in the angle of arrival / departure (AOA / AOD) domain and delay domain due to their high path loss and sensitivity to blockage. It has been corroborated by experiments that the limited path components typically arrive in $1 \sim 4$ 'clusters' [5]–[7]. In the literature, by exploiting the distinct lack of scattering experienced by mmWave channels, several advanced channel estimation schemes have been proposed in [1], [8]–[11]. Codebook based BF methods have been conceived in [1], [8], [9], where the core idea is to search through the predefined BF-weight codebook in order to find the best BF-vector pair for transmission / reception. However, their contribution did not conceive explicit channel estimation schemes for multiuser interference cancellation. As a further development, it was shown that random compressive sensing (CS) using pseudo random phase shifters [10], [11] is more suitable for multiuser systems, since all users can simultaneously estimate their channels thanks to the random nature of the transmitted beams [12].

Beyond the above-mentioned sparsity, mmWave channels also exhibit additional subtle features [13], [14], which can be further exploited for improving the attainable channel estimation performance, especially for the low signal-to-noise ratios (SNRs) routinely encountered in mmWave communications before BF. To elaborate a little further, mmWave channels exhibit a clustered structure in the virtual AOA / AOD domain, which several large coefficients are grouped together due to the effect of power leakage (See Fig. 1 (a)) [13], [15]. By exploiting their clustered structure, the support detection (SD)-based channel estimation scheme proposed for narrowband flat-fading channels in [13] outperformed the orthogonal matching pursuit algorithm [16]. Furthermore, by exploiting the subtle changes between the adjacent channel elements, the algorithm of [17] outperformed the SD scheme of [13]. For estimating broadband frequency-selective fading channels, an efficient algorithm was proposed in [14] based on the assumption that the subchannels of orthogonal frequency division multiplexing (OFDM) systems have the same sparse common support (SCS) [18]. However, the angular spread exhibited in the AOA /AOD domain [7] and the cluster duration [19] exhibited in the delay domain were not considered in [13] and [14], which would enhance the clustered structure (See Fig. 1 (b), (c) and (d)). The work [20] exploited the sparsity in angular and delay domains that was designed for mmWave MIMO systems with few-bit analog-to-digital conversion. To the best of our knowledge, jointly exploiting the 3-D clustered structure of channels in the virtual AOA-AOD-delay domain for improving the channel estimation in mmWave systems with hybrid analog / digital precoding has not been proposed in the literature.

This work was supported by the National Nature Science Foundation of China with Grant Nos. 91438206 and 91638205. *Corresponding author: Linling Kuang.*

Xincong Lin is with the Department of Aerospace Engineering, Tsinghua University, Beijing 100084, China (e-mail: linxc15@mails.tsinghua.edu.cn).

Sheng Wu, Chunxiao Jiang, Linling Kuang and Jian Yan are with the Tsinghua Space Center, Tsinghua University, Beijing 100084, China (e-mail: {thuraya, jchx, kll, yanjian_ee}@tsinghua.edu.cn).

Lajos Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom (email: lh@ecs.soton.ac.uk).



(d) Clustered structure in the virtual AOA-delay domain (c) Clustered structure in the virtual AOA-AOD domain. Figure 1: (a) and (b) show 2-D clustered structure in the virtual AOA-AOD domain of channels in frequency domain. (c) and (d) show 3-D clustered structure in the virtual AOA-AOD-delay domain of channels in delay domain. The base station has 64 antennas and user equipment has 16 antennas. There are 4 physical clusters with one LOS path and three NLOS paths.

15

10

Virtual AOD Index

0.2

0 64 48

32

Virtual AOA Index

16

1

In our earlier work [21], the approximate message passing with nearest neighbor pattern learning (AMP-NNSPL) algorithm is proposed for learning the sparse clustered structure in the virtual angular domain of a conventional massive MIMO-OFDM system operating at the carrier frequency of 2 GHz. In this paper, to fully exploit the sparse 3-D clustered structure exhibited in the virtual angular and delay domain, which is deterministic but unknown for a specific propagation environment, we extend the AMP-NNSPL algorithm to adaptively learn the 3-D clustered structure for the sake of accurately estimating the channel of broadband mmWave massive MIMO-OFDM systems with multiuser hybrid precoding. Specifically, we develop a Delay-Domain (DD) algorithm based on the AMP framework, which is termed as AMP-NNSPL-DD, and then the state evolution (SE) of the AMP-NNSPL-DD algorithm is derived for our ensuring performance analysis. Since the vector AMP (VAMP) algorithm is more robust with respect to general measurement matrix and its SE eminently characterizes the attainable performance bound when the measurement matrix is large and right-rotationally invariant [22], we also developed a VAMP-NNSPL-DD algorithm and quantified its SE based on the VAMP framework. However, compared to the AMP-NNSPL-DD, the VAMP-

20

Virtual AOA Index

60

NNSPL-DD requires more storage space and imposes a higher complexity. Our simulation results demonstrate that jointly exploiting the sparse clustered structure in the angular and delay domain is capable of attaining considerable performance gains, over the algorithms of [13], [16], [17], [21], [23]-[26], which only exploit the sparse clustered structure in the angular domain. In particular, for obtaining a performance gain at a low complexity, switching back and forth between the delaydomain and frequency-domain is efficiently carried out by the fast Fourier transform (FFT). Furthermore, our VAMP-NNSPL-DD solution implies that passing messages from the frequency domain to the delay domain is equivalent to a multiple-measurement-vector (MMV) problem of [25]. However, the SE need make an idealized simplifying assumption which results in a generalized MMV (GMMV) problem [25]. As a result, the AMP-NNSPL-DD and its SE do not match well with short measuring time, but match better with long measuring time. By contrast, the SE of the VAMP-NNSPL-DD characterizes the performance bound of the (V)AMP-NNSPL-DD algorithm more accurately.

480

360

240 120 240 Delay Index

Notation: The transpose, complex conjugate and conjugate transpose operators are given by $(\cdot)^{T}$, $(\cdot)^{*}$ and $(\cdot)^{H}$, respectively. $\mathbb{E}[\cdot]$ denotes the statistical expectation. $||A||_{F}$ is the

3

Frobenius norm of matrix A. $||a||_2$ is the ℓ_2 norm of vector a. $\lfloor x \rfloor$ is the largest integer n that $n \leq x \cdot \lceil x \rceil$ is the smallest integer n that $n \geq x$. $\Phi(g, m)$ is the gth-row and mth-column element of the matrix Φ . $\ln(\cdot)$ and $\lg(\cdot)$ denotes the natural logarithm and logarithm to ten, respectively. **vec** (A) denotes vectorizing a matrix as a vector. \otimes denotes the Kronecker product. Finally, $\mathcal{N}_{\mathbb{C}}(h; \mu, \nu)$ denotes the Gaussian distribution function of h with mean μ and variance ν . And I_N is unit matrix with size N. $e_i \in \mathbb{R}^{N \times 1}$ or $e_i \in \mathbb{R}^{1 \times N}$ being the standard basis vector with the unique one in the *i*th entry, and with size being N. Tr (A) is the trace of a matrix. $\langle a \rangle = \frac{1}{N} \sum_n a_n$.

II. SYSTEM MODEL

A. Signal Transmission at the UEs

Let us consider the general family of hybrid analog-digital precoding and combining architectures invoked for mmWave communications [1], [12], [27]. The base station (BS) having N_{BS} antennas and $N_{BS}^{RF} = K$ RF chains serves K user equipment (UE) having N_{UE} antennas and $N_{UE}^{RF} = 1$ RF chain [14]. In this paper, we focus our attention on the uplink channels' estimation. The transmitter employs OFDM modulation, where P pilots are uniformly allocated across a total of N subcarriers and the set of pilot subcarriers is denoted by $\mathcal{P} = \{1\Delta, 2\Delta, \dots, P\Delta\}$, where we have $\Delta = \lfloor N/P \rfloor$. Furthermore, $\{s_{kgp} \mid k = 1, \dots, K, g = 1, \dots, G, p \in \mathcal{P}\}$ denotes the pilots associated with the *p*th subcarrier of the *g*th OFDM symbol of the *k*th user. After precoding by baseband transmit precoder f_{BBkgp} followed by an RF precoder $f_{RFkg} \in \mathbb{C}^{N_{UE} \times 1}$ can be written as

$$\boldsymbol{x}_{kgp} = \boldsymbol{f}_{kgp} \boldsymbol{s}_{kgp}, \tag{1}$$

where $f_{kgp} = f_{\mathsf{RF}kg} f_{\mathsf{BB}kgp} \in \mathbb{C}^{N_{\mathsf{UE}} \times 1}$ is the UE's combined transmit precoder matrix.

B. MmWave Channel Model

The mmWave channel can be modeled as a sum of N_{cl} scattered clusters, each of which contributes $L_{n_{cl}}$ propagation paths [27]. For a uniform linear array, the baseband frequency response of a quasi-static physical channel H(f) can be modeled as [28]–[30]

$$\boldsymbol{H}(f) = \sum_{i=1}^{L_{\text{path}}} \beta_i \boldsymbol{a}_{\mathsf{R}}(\theta_{\mathsf{R}i}) \, \boldsymbol{a}_{\mathsf{T}}^{\mathsf{H}}(\theta_{\mathsf{T}i}) \, e^{-j2\pi\tau_i f}, \qquad (2)$$

where $-B/2 \le f \le B/2$ with *B* is the two-sided bandwidth, $L_{\text{path}} = \sum_{n_{\text{cl}}=1}^{N_{\text{cl}}} L_{n_{\text{cl}}}$ is the total number of physical paths, β_i denotes the complex-valued path gain, τ_i is the path-delay, while

$$\boldsymbol{a}_{\mathsf{R}}(\theta) = \left[1, e^{-j2\pi\theta}, \cdots, e^{-j2\pi(N_{\mathsf{BS}}-1)\theta}\right]^{\mathsf{T}},\tag{3}$$

$$\boldsymbol{a}_{\mathsf{T}}(\theta) = \left[1, e^{-j2\pi\theta}, \cdots, e^{-j2\pi(N_{\mathsf{UE}}-1)\theta}\right]^{\mathsf{T}},\tag{4}$$

denotes the receive and transmit steering vector with θ being the normalized angle, respectively. The normalized angles $\theta_{Ri} \in (-1/2, 1/2)$ and $\theta_{Ti} \in (-1/2, 1/2)$ are related to the AOA $\phi_{Ri} \in (-\pi/2, \pi/2)$ by $\theta_{Ri} = \frac{1}{2} \sin(\phi_{Ri})$ and the AOD $\phi_{\mathsf{T}i} \in (-\pi/2, \pi/2)$ by $\theta_{\mathsf{T}i} = \frac{1}{2} \sin(\phi_{\mathsf{T}i})$, respectively. For notational simplicity, the user index k in the channel model (2) is omitted.

The physical channel H(f) is related to the channel W(f) in virtual AOA-AOD-frequency domain by [28]–[30]

$$\boldsymbol{H}\left(f\right) = \boldsymbol{A}_{\mathsf{R}}\boldsymbol{W}\left(f\right)\boldsymbol{A}_{\mathsf{T}}^{\mathsf{H}},\tag{5}$$

where

$$A_{\mathsf{R}} = \left[\boldsymbol{a}_{\mathsf{R}} \left(\boldsymbol{\theta}_{\mathsf{R}}^{1} \right), \boldsymbol{a}_{\mathsf{R}} \left(\boldsymbol{\theta}_{\mathsf{R}}^{2} \right), \cdots, \boldsymbol{a}_{\mathsf{R}} \left(\boldsymbol{\theta}_{\mathsf{R}}^{N_{\mathsf{B}}} \right) \right] / \sqrt{N_{\mathsf{B}}}, \quad (6)$$
$$A_{\mathsf{T}} = \left[\boldsymbol{a}_{\mathsf{T}} \left(\boldsymbol{\theta}_{\mathsf{T}}^{1} \right), \boldsymbol{a}_{\mathsf{T}} \left(\boldsymbol{\theta}_{\mathsf{T}}^{2} \right), \cdots, \boldsymbol{a}_{\mathsf{T}} \left(\boldsymbol{\theta}_{\mathsf{R}}^{N_{\mathsf{U}}} \right) \right] / \sqrt{N_{\mathsf{UE}}}, \quad (7)$$

is a version of discrete Fourier transform (DFT) matrix with phase shift, and

$$\theta_{\mathsf{R}}^{n_{\mathsf{R}}} = \frac{1}{N_{\mathsf{BS}}} \left[n_{\mathsf{R}} - \bar{N}_{\mathsf{BS}} \right], \quad \theta_{\mathsf{T}}^{n_{\mathsf{T}}} = \frac{1}{N_{\mathsf{UE}}} \left[n_{\mathsf{T}} - \bar{N}_{\mathsf{UE}} \right], \quad (8)$$

are the virtual AOA and AOD, respectively, with $\bar{N}_{BS} = (N_{BS} + 1)/2$, $\bar{N}_{UE} = (N_{UE} + 1)/2$, $n_R = 1, \dots, N_{BS}$ and $n_T = 1, \dots, N_{UE}$. n_R and n_T is referred to virtual AOA index and virtual AOD index, respectively, in Fig. 1. Furthermore, the channel W(f) is related to the channel H_l in virtual AOA-AOD-delay domain by [28]–[30]

$$\boldsymbol{W}_{p} = \boldsymbol{W}\left(f_{p}\right) = \frac{1}{\sqrt{L}} \sum_{l=1}^{\tilde{L}} \boldsymbol{H}_{l} e^{-j2\pi(l-1)\left(-\frac{1}{2} + \frac{P}{N}\right)}, \qquad (9)$$

where $f_p = -\frac{B}{2} + \frac{p}{N}B$ is *p*th subcarrier frequency, and $\bar{L} = L + 1$ with $L = \lceil B\tau_{max} \rceil$, where *B* and τ_{max} is the bandwidth and maximum path delay, respectively. The elements in H_l is [28]–[30]

$$H_{n_{\mathsf{R}}n_{\mathsf{T}}l} = \sqrt{\bar{L}} \sum_{i=1}^{L_{\mathsf{path}}} \beta_i f_{N_{\mathsf{BS}}} \left(\theta_{\mathsf{R}i} - \theta_{\mathsf{R}}^{n_{\mathsf{R}}} \right) f_{N_{\mathsf{UE}}}^* \left(\theta_{\mathsf{T}i} - \theta_{\mathsf{T}}^{n_{\mathsf{T}}} \right) \\ \times \operatorname{sinc} \left(B\tau_i - (l-1) \right), \tag{10}$$

where $f_N(\theta) = \frac{1}{\sqrt{N}} e^{-j\pi\theta(N-1)\frac{\sin(\pi N\theta)}{\sin(\pi\theta)}}$ and sinc $(x) = \frac{\sin(\pi x)}{\pi x}$.

Let us consider $N_{cl} = 4$ physical clusters associated with a line of sight (LOS) cluster and three non-LOS (NLOS) clusters, which is a common scenario in mmWave channels [5]-[7]. Since the power of LOS cluster may 20 dB higher than that of the NLOS cluster [5], [13], [14], the complex path gains are drawn from $\mathcal{N}_{\mathbb{C}}(\beta_i; 0, 10^{\mathsf{power}/10})$, with the path-power being 0 dBm for the LOS component and $-5 \sim -20$ dBm for the NLOS component. The maximum path delay is about $\tau_{\text{max}} = 600 \text{ ns}$ [6], therefore, the path delays τ_i are uniformly selected from $0 \sim 600$ ns. Other parameters are set as $f_0 = 28$ GHz, B = 800 MHz [6], $L = \lceil B\tau_{max} \rceil = 480$, $N_{BS} = 64$, $N_{\text{UE}} = 16, \ \left[\phi_{\text{R1}}, \phi_{\text{R2}}, \phi_{\text{R3}}, \phi_{\text{R4}}\right] = \ \left[70, 20, -20, -70\right]/180\pi$ and $[\phi_{T1}, \phi_{T2}, \phi_{T3}, \phi_{T4}] = [60, 30, -30, -60]/180\pi$. With these parameters, the channel H_l can be generated from (10), and the channel W_p can be generated from (9). The normalized amplitude of W_1 is shown in Fig. 1 (a), from which we can observe the so-called power leakage effect, namely that the signal power is not concentrated to a single rectangle, as demonstrated in [13] and [15]. Furthermore, each cluster exhibits an angular spread of about 15 degrees in the physical AOA / AOD [7], and also has cluster duration of about 9 ns in delay domain [19], which is shown in Fig. 1 (b). Fig. 1 (c) shows the channel H_1 , and Fig. 1 (d) shows the channel $[H_1e_1, \dots, H_{481}e_1] \in \mathbb{C}^{64\times481}$ with $e_1 \in \mathbb{R}^{16\times1}$. Generally speaking, due to the effect of power leakage, angular spread and cluster duration, the channels $[H_1, \dots, H_{\tilde{L}}]$ exhibit 3-D clustered structure, which constitutes the motivation for the proposed algorithm.

C. Signal Receiving at the BS

From (1) and (5), the BS receives the uplink signal $r_{gp} \in \mathbb{C}^{N_{BS} \times 1}$ at the *p*th subcarrier of the *g*th OFDM symbol from multiple users formulated as

$$\boldsymbol{r}_{gp} = \sum_{k=1}^{K} \boldsymbol{A}_{\mathsf{R}} \boldsymbol{W}_{kp} \boldsymbol{A}_{\mathsf{T}}^{\mathsf{H}} \boldsymbol{x}_{kgp} + \boldsymbol{w}_{gp}, \qquad (11)$$

where $\mathbf{w}_{gp} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{w}_{gp}; 0, \sigma)$ is the additive noise. The SNR is defined as SNR=1/ σ under the condition that $\mathbb{E}\left[s_{kgp}s_{kgp}^*\right] = 1$, $\left\|\mathbf{f}_{kgp}\right\|_{\mathrm{F}}^2 = 1$ and $\mathbb{E}\left[\left\|\mathbf{W}_{kp}\right\|_{\mathrm{F}}^2\right] = \mathcal{N}_{\mathrm{BS}}\mathcal{N}_{\mathrm{UE}}$ [27]. Then the received signal \mathbf{r}_{gp} is further processed by the combined matrix $\mathbf{Z}_{gp} = \mathbf{Z}_{RFg}\mathbf{Z}_{BBgp} \in \mathbb{C}^{\mathcal{N}_{\mathrm{BS}} \times K}$ as follows

$$\boldsymbol{y}_{gp} = \boldsymbol{Z}_{gp}^{H} \boldsymbol{r}_{gp} = \boldsymbol{Z}_{gp}^{H} \sum_{k=1}^{K} \boldsymbol{A}_{\mathsf{R}} \boldsymbol{W}_{kp} \boldsymbol{A}_{\mathsf{T}}^{\mathsf{H}} \boldsymbol{x}_{kgp} + \boldsymbol{Z}_{gp}^{\mathsf{H}} \boldsymbol{w}_{gp}.$$
(12)

By stacking K users' quantities from (12) with (1), we arrive at

$$\nu_{gp} = \mathbf{Z}_{gp}^{\mathsf{H}} \mathbf{A}_{\mathsf{R}} \bar{\mathbf{W}}_{p} \bar{\mathbf{A}}_{\mathsf{T}}^{\mathsf{H}} \bar{\mathbf{f}}_{gp} + \mathbf{n}_{gp}, \qquad (13)$$

where $\bar{W}_p = \begin{bmatrix} W_{1p}, W_{2p}, \cdots, W_{Kp} \end{bmatrix} \in \mathbb{C}^{N_{\text{BS}} \times KN_{\text{UE}}}, \ \bar{A}_{\text{T}} = \text{diag} \begin{bmatrix} A_{\text{T}}, A_{\text{T}}, \cdots, A_{\text{T}} \end{bmatrix} \in \mathbb{C}^{KN_{\text{UE}} \times KN_{\text{UE}}}, \ \bar{f}_{gp} = \begin{bmatrix} \left(f_{1gp} s_{1gp} \right)^{\text{T}}, \left(f_{2gp} s_{2gp} \right)^{\text{T}}, \cdots, \left(f_{Kgp} s_{Kgp} \right)^{\text{T}} \end{bmatrix}^{\text{T}} \in \mathbb{C}^{KN_{\text{UE}} \times 1} \ \text{and} \ \boldsymbol{n}_{gp} = \boldsymbol{Z}_{gp}^{\text{H}} \boldsymbol{w}_{gp} \in \mathbb{C}^{K \times 1}. \ \text{By vectorizing} \ \bar{W}_p, \ \text{we have} \ \begin{bmatrix} 1 \end{bmatrix}, \ \begin{bmatrix} 14 \end{bmatrix}$

$$\begin{aligned} \mathbf{y}_{gp} &= \operatorname{vec} \left(\mathbf{Z}_{gp}^{\mathsf{H}} \mathbf{A}_{\mathsf{R}} \bar{\mathbf{W}}_{p} \bar{\mathbf{A}}_{\mathsf{T}}^{\mathsf{H}} \bar{f}_{gp} \right) + \mathbf{n}_{gp} \\ &= \left(\left(\bar{\mathbf{A}}_{\mathsf{T}}^{\mathsf{H}} \bar{f}_{gp} \right)^{\mathsf{T}} \otimes \mathbf{Z}_{gp}^{\mathsf{H}} \mathbf{A}_{\mathsf{R}} \right) \operatorname{vec} \left(\bar{\mathbf{W}}_{p} \right) + \mathbf{n}_{gp} \\ &= \mathbf{\Phi}_{gp} \mathbf{w}_{p} + \mathbf{n}_{gp}. \end{aligned}$$
(14)

By stacking G successive received signals y_{gp} , we get the system model in the virtual AOA-AOD-frequency domain as

$$\mathbf{y}_p = \mathbf{\Phi}_p \mathbf{w}_p + \mathbf{n}_p, \tag{15}$$

where we have $\boldsymbol{y}_p = \begin{bmatrix} \boldsymbol{y}_{1p}^{\mathsf{T}}, \boldsymbol{y}_{2p}^{\mathsf{T}}, \cdots, \boldsymbol{y}_{Gp}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \in \mathbb{C}^{KG \times 1},$ $\boldsymbol{\Phi}_p = \begin{bmatrix} \boldsymbol{\Phi}_{1p}^{\mathsf{T}}, \boldsymbol{\Phi}_{2p}^{\mathsf{T}}, \cdots, \boldsymbol{\Phi}_{Gp}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \in \mathbb{C}^{KG \times KN_{\mathsf{BS}}N_{\mathsf{UE}}}, \text{ and } \boldsymbol{n}_p = \begin{bmatrix} \boldsymbol{n}_{1p}^{\mathsf{T}}, \boldsymbol{n}_{2p}^{\mathsf{T}}, \cdots, \boldsymbol{n}_{Gp}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \in \mathbb{C}^{KG \times 1}.$ Denote

$$\boldsymbol{H} = \left[\operatorname{vec}\left(\boldsymbol{\bar{H}}_{1} \right), \cdots, \operatorname{vec}\left(\boldsymbol{\bar{H}}_{\bar{L}} \right) \right] \in \mathbb{C}^{M \times \bar{L}}, \qquad (16)$$

$$\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_P] \in \mathbb{C}^{M \times P},\tag{17}$$

where $\bar{H}_l = [H_{1l}, H_{2l}, \cdots, H_{Kl}] \in \mathbb{C}^{N_{BS} \times KN_{UE}}$, and $M = KN_{BS}N_{UE}$, then we have

$$\boldsymbol{W} = \boldsymbol{H}\boldsymbol{\phi}^T, \tag{18}$$

where $\phi \in \mathbb{C}^{P \times \overline{L}}$ with elements being $\phi_{pl} = \frac{1}{\sqrt{\overline{L}}} e^{-j2\pi(l-1)\left(-\frac{1}{2} + \frac{P}{N}\right)}$. From (15), (18), and by vectorizing H, we have

$$y_{p} = \operatorname{vec}\left(\Phi_{p}H\phi^{T}e_{p}\right) + n_{p}$$
$$= \left(\left(\phi^{T}e_{p}\right)^{T}\otimes\Phi_{p}\right)\operatorname{vec}\left(H\right) + n_{p}$$
$$= \Psi_{p}\bar{h} + n_{p}, \tag{19}$$

where $\Psi_p \in \mathbb{C}^{KG \times KN_{BS}N_{UE}\bar{L}}$, $\bar{h} \in \mathbb{C}^{KN_{BS}N_{UE}\bar{L} \times 1}$, and $e_p \in \mathbb{R}^{P \times 1}$. By stacking *P* subchannel received signals y_p , we get the system model in the virtual AOA-AOD-delay domain as

$$y = \Psi \bar{h} + n, \tag{20}$$

where $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1^\mathsf{T}, \mathbf{y}_2^\mathsf{T}, \cdots, \mathbf{y}_P^\mathsf{T} \end{bmatrix}^\mathsf{T} \in \mathbb{C}^{KGP \times 1},$ $\mathbf{\Psi} = \begin{bmatrix} \mathbf{\Psi}_1^\mathsf{T}, \mathbf{\Psi}_2^\mathsf{T}, \cdots, \mathbf{\Psi}_P^\mathsf{T} \end{bmatrix}^\mathsf{T} \in \mathbb{C}^{KGP \times KN_{\mathsf{BS}}N_{\mathsf{UE}}\bar{L}}$ and $\mathbf{n} = \begin{bmatrix} \mathbf{n}_1^\mathsf{T}, \mathbf{n}_2^\mathsf{T}, \cdots, \mathbf{n}_P^\mathsf{T} \end{bmatrix}^\mathsf{T} \in \mathbb{C}^{KGP \times 1}.$

The AMP-NNSPL proposed in [21] can be readily applied on model (15), which is termed as AMP-NNSPL-FD (frequency domain) in the following. The complexity of AMP-NNSPL-FD is dominated by matrix-vector multiplies with $\Phi_p \in \mathbb{C}^{KG \times KN_{BS}N_{UE}}$, i.e., the scale of $O(K^2GPN_{BS}N_{UE})$. However, directly applying the AMP-NNSPL to the model (20) would lead to high complexity of $O(K^2GPN_{BS}N_{UE}\bar{L})$, as the term \bar{L} is usually huge in broadband mmWave systems. Given the model of (15) and (18), we extend our lowcomplexity AMP-NNSPL solution based on the (vector) AMP framework of [22], [31]–[33] to estimate the channel in the virtual AOA-AOD-delay domain in the next section.

III. PROPOSED MMWAVE CHANNEL ESTIMATION ALGORITHM

A. The AMP-NNSPL-DD Algorithm

Our goal is to infer the channel H in the virtual AOA-AOD-delay domain from the measurements y under the model (15) and the constraint (18). In particular, the *aposteriori* probability can be computed according to Bayesian rule as

$$p(\boldsymbol{H} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{W}) p(\boldsymbol{W} \mid \boldsymbol{H}) p(\boldsymbol{H})}{p(\boldsymbol{y})}, \qquad (21)$$

where $p(\mathbf{y}) = \int \int p(\mathbf{y} | \mathbf{W}) p(\mathbf{W} | \mathbf{H}) p(\mathbf{H}) d\mathbf{W} d\mathbf{H}$. The numerator in (21) can be factored into

$$p(\boldsymbol{H}) p(\boldsymbol{W} | \boldsymbol{H}) p(\boldsymbol{y} | \boldsymbol{W}) = \prod_{m=1}^{M} p(\boldsymbol{h}_{m}) p(\boldsymbol{w}_{m} | \boldsymbol{h}_{m})$$
$$\times \prod_{p \in \mathcal{P}} f_{p}(\boldsymbol{y}_{p} | \boldsymbol{w}_{p}). \quad (22)$$

where $h_m \in \mathbb{C}^{\bar{L}\times 1}$ is the *m*th row of channels $H \in \mathbb{C}^{M\times \bar{L}}$, $w_m \in \mathbb{C}^{P\times 1}$ and $w_{\cdot p} \in \mathbb{C}^{M\times 1}$ are the *m*th row and *p*th column of channels $W \in \mathbb{C}^{M\times P}$, respectively. To exploit the 3-D clustered structure in the virtual AOA-AOD-delay domain as shown in Fig. 1 (c) and (d), we apply a flexible spike and slab priori model to the channels $H \in \mathbb{C}^{M\times \bar{L}}$,

$$p(\boldsymbol{H};\boldsymbol{\xi}) = \prod_{m=1}^{M} p(\boldsymbol{h}_m)$$



Figure 2: The factor graph representation of broadband mmWave massive MIMO-OFDM Systems.

$$= \prod_{m=1}^{M} \prod_{l=1}^{\tilde{L}} p(h_{ml})$$

=
$$\prod_{m=1}^{M} \prod_{l=1}^{\tilde{L}} \left[(1 - \rho_{ml}) \,\delta(h_{ml}) + \rho_{ml} \mathcal{N}_{\mathbb{C}}(h_{ml}; 0, \eta_{ml}) \right],$$

(23)

where $\boldsymbol{\xi} \triangleq \{\rho_{ml}, \eta_{ml}, \sigma \mid m = 1, 2, \dots, M, l = 1, 2, \dots, \bar{L}\}$ represents the system parameters. Specifically, $\rho_{ml} \in (0, 1)$ denotes the sparsity ratio, while η_{ml} and σ are the *apriori* variance of channel coefficient and noise, respectively, and $\delta(h_{ml})$ is the Dirac delta function. The conditional probability $p(\boldsymbol{W} \mid \boldsymbol{H})$ reads

$$p(\boldsymbol{W} \mid \boldsymbol{H}) = \prod_{m=1}^{M} p(\boldsymbol{w}_{m} \mid \boldsymbol{h}_{m})$$
$$= \prod_{m=1}^{M} \delta(\boldsymbol{w}_{m} - \boldsymbol{\phi}\boldsymbol{h}_{m})$$
$$= \prod_{m=1}^{M} \prod_{p \in \mathcal{P}} \delta\left(\boldsymbol{w}_{mp} - \sum_{l=1}^{\tilde{L}} \boldsymbol{\phi}_{pl} h_{ml}\right), \quad (24)$$

and the channel transition function $p(\mathbf{y} \mid \mathbf{W})$ is factorized into

$$p(\mathbf{y} \mid \mathbf{W}) = \prod_{p \in \mathcal{P}} f_p(\mathbf{y}_p \mid \mathbf{w}_{\cdot p}) = \prod_{p \in \mathcal{P}} \prod_{g=1}^{KG} f_{gp}(\mathbf{y}_{gp} \mid \mathbf{w}_{\cdot p}),$$
(25)

where

$$f_p\left(\mathbf{y}_p \mid \mathbf{w}_{\cdot p}\right) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{y}_p; \mathbf{\Phi}_p \mathbf{w}_{\cdot p}, \sigma\right), \tag{26}$$

$$f_{gp}\left(y_{gp} \mid \boldsymbol{w}_{\cdot p}\right) = \mathcal{N}_{\mathbb{C}}\left(y_{gp}; \boldsymbol{e}_{g}\boldsymbol{\Phi}_{p}\boldsymbol{w}_{\cdot p}, \boldsymbol{\sigma}\right).$$
(27)

The factorization described by (22)-(25) can be represented by the factor graph of Fig. 2, where $\delta \left(w_{mp} - \sum_{l=1}^{\tilde{L}} \phi_{pl} h_{ml} \right)$ appears as a function node g_{mp} . Fig. 2 includes two parts. In the left part, the message passing is carried by the approximate message passing (AMP) algorithm [31]–[33], while the 1) The AMP-NNSPL-DD Algorithm: The channels H, W and the parameters ξ can be iteratively estimated by message passing and by minimizing the Bethe free energy under a neighborhood constraint [21]. By fixing the parameters ξ to the values estimated at the previous iteration, the terms κ_{gp}^t , z_{gp}^t , ϑ_{mp}^t and ω_{mp}^t defined at each node, i.e., f_{gp} and w_{mp} , are calculated by AMP as shown in lines 6 and 7 of table I. In the right part of factor graph, denote

$$\varepsilon_{ml}^{t} = \sum_{p \in \mathcal{P}} \frac{\left(\phi_{pl}\right)^{-1} \hat{z}_{g_{mp} \to h_{ml}}^{t}}{v_{g_{mp}}^{t}},$$
(28)

$$C_{mp}^{t} = \sum_{l=1}^{\bar{L}} \frac{v_{h_{ml}}^{t} \hat{z}_{g_{mp}}^{t} \rightarrow h_{ml}}{v_{g_{mp}}^{t}},$$
 (29)

where $v_{g_{mp}}^t$ and $v_{h_{ml}}^t$ are the variance defined at the function node g_{mp} and the variable node h_{ml} , respectively, and

$$\hat{z}_{g_{mk} \to h_{ml}}^{t} = \hat{w}_{w_{mp} \to g_{mp}}^{t} - \sum_{l' \neq l}^{L} \phi_{pl'} \hat{h}_{h_{ml'} \to g_{mp}}^{t-1}, \qquad (30)$$

with $\hat{w}_{w_{mp} \to g_{mp}}^{t}$ being the mean of the message $u_{w_{mp} \to g_{mp}}^{t} \left(w_{mp} \right) = \mathcal{N}_{\mathbb{C}} \left(w_{mp}; \hat{w}_{w_{mp} \to g_{mp}}^{t}, v_{w_{mp} \to g_{mp}}^{t} \right)$ passed from the variable node w_{mp} to the factor node g_{mp} , and $\hat{h}_{h_{ml'} \to g_{mp}}^{t-1}$ being the mean of the message passed from the variable node $h_{ml'}$ to the factor node g_{mp} . ε_{ml}^{t} and ζ_{mp}^{t} are calculated recursively by GMP as shown in lines 10 and 17 of table I, respectively.

By the sum-product message passing rule [36], the message $u_{w_{mp} \rightarrow g_{mp}}^{t} (w_{mp})$ is calculated by

$$u_{w_{mp} \to g_{mp}}^{t} \left(w_{mp} \right) = \prod_{g=1}^{KG} u_{f_{gp} \to w_{mp}}^{t} \left(w_{mp} \right)$$
$$\stackrel{(a)}{\approx} \mathcal{N}_{\mathbb{C}} \left(w_{mp}; \omega_{mp}^{t}, \vartheta_{mp}^{t} \right), \qquad (31)$$

where step $\stackrel{(a)}{\approx}$ is shown in [37]. Hence, we have

$$\hat{w}^t_{w_{mp} \to g_{mp}} = \omega^t_{mp}. \tag{32}$$

Then the variance $v_{g_{mp}}^t$ and mean $\hat{z}_{g_{mp}}^t$ defined at the factor node g_{mp} , and the variance ϵ_m^t and mean μ_{ml}^t defined at the variable node h_{ml} are calculated by GMP as shown in lines 9 and 11 of table I. The *aposteriori* distributions of h_{ml} are obtained as follows

$$p\left(h_{ml} \mid \mathbf{y}; \boldsymbol{\xi}^{t-1}\right) = \frac{1}{\int p\left(h_{ml}\right) \mathcal{N}_{\mathbb{C}}\left(h_{ml}; \boldsymbol{\mu}_{ml}^{t}, \boldsymbol{\epsilon}_{m}^{t}\right) dh_{ml}} \times p\left(h_{ml}\right) \mathcal{N}_{\mathbb{C}}\left(h_{ml}; \boldsymbol{\mu}_{ml}^{t}, \boldsymbol{\epsilon}_{m}^{t}\right)}$$

$$= \left(1 - \pi_{ml}^{t}\right) \delta\left(h_{ml}\right) + \pi_{ml}^{t} \mathcal{N}_{\mathbb{C}}\left(h_{ml}; \tilde{h}_{ml}^{t}, \tilde{v}_{ml}^{t}\right),$$
(33)

where π_{ml}^t , \tilde{v}_{ml}^t and \tilde{h}_{ml}^t are shown in lines 13 and 14 of table I. The *aposteriori* mean and variance of h_{ml} are calculated as

$$g_{\text{mean}}\left(\mu_{ml}^{t}, \epsilon_{m}^{t}; \boldsymbol{\xi}^{t-1}\right) = \int h_{ml} p\left(h_{ml} \mid \boldsymbol{y}; \boldsymbol{\xi}^{t-1}\right) dh_{ml}, \quad (34)$$
$$g_{\text{var}}\left(\mu_{ml}^{t}, \epsilon_{m}^{t}; \boldsymbol{\xi}^{t-1}\right) = \int |h_{ml}|^{2} p\left(h_{ml} \mid \boldsymbol{y}; \boldsymbol{\xi}^{t-1}\right) dh_{ml}$$
$$- \left|g_{\text{mean}}\left(\mu_{ml}^{t}, \epsilon_{m}^{t}; \boldsymbol{\xi}^{t-1}\right)\right|^{2}, \quad (35)$$

and they are shown in line 15 of table I. According to the sumproduct message passing rule, the *aposteriori* distributions of w_{mp} are obtained by

$$p\left(w_{mp} \mid \mathbf{y}_{p}\right)$$

$$= u_{w_{mp} \rightarrow g_{mp}}^{t}\left(w_{mp}\right)u_{g_{mp} \rightarrow w_{mp}}^{t}\left(w_{mp}\right)$$

$$= \mathcal{N}_{\mathbb{C}}\left(w_{mp}; \omega_{mp}^{t}, \vartheta_{mp}^{t}\right)\mathcal{N}_{\mathbb{C}}\left(w_{mp}; \hat{w}_{g_{mp} \rightarrow w_{mp}}^{t}, v_{g_{mp} \rightarrow w_{mp}}^{t}\right)$$

$$\propto \mathcal{N}_{\mathbb{C}}\left(w_{mp}; \hat{w}_{mp}^{t}, v_{w_{mp}}^{t}\right),$$
(36)

where $\hat{w}_{g_{mp} \to w_{mp}}^t$ and $v_{g_{mp} \to w_{mp}}^t$ are calculated by GMP as shown in line 18 of table I, $v_{w_{mp}}^t$ and \hat{w}_{mp}^t are calculated as shown in line 19 of table I.

As shown in Fig. 1 (c) and (d), the channel H of (16) exhibits 3-D clustered structure in virtual AOA-AOD-delay domain. For notational convenience, let $m = (k-1)N_{\text{BS}}N_{\text{UE}}+(i-1)N_{\text{BS}}+j$, where $m = 1, \dots, M$, $k = 1, \dots, K$, $i = 1, \dots, N_{\text{BS}}$ and $j = 1, \dots, N_{\text{UE}}$. Then we denote $h_{kijl} \triangleq h_{ml}$, where i, j and l being the index in the 3-D virtual AOA-AOD-delay domain. In the following, the user index k is dropped for notational simplicity, and we reuse the user index k when we summarize the proposed algorithm in the following table I. It is observed from Fig. 1 (c) and (d) that h_{ijl} and its neighbors $\{h_{qrs \in N_{ijl}}\}$ tend to be either simultaneously small value or large value, where N_{nij} denotes the set of neighbor indices of element h_{ijl} , and is defined as follows¹

$$\mathcal{N}_{ijl} = \{ (i - 1, j, l), (i + 1, j, l), (i, j - 1, l), \\ (i, j + 1, l), (i, j, l - 1), (i, j, l + 1) \}.$$
(37)

Hence, the sparsity ratio ρ_{ijl} and the *apriori* variance η_{ijl} of h_{ijl} should be close to $\{\rho_{qrs\in\mathcal{N}_{nij}}\}$ and $\{\eta_{qrs\in\mathcal{N}_{nij}}\}$, which can be described by minimizing $\sum_{qrs\in\mathcal{N}_{nij}} (\rho_{ijl} - \rho_{qrs})^2$ and $\sum_{qrs\in\mathcal{N}_{nij}} (\eta_{ijl} - \eta_{qrs})^2$, respectively. By fixing the *aposteriori* distribution of h_{ml} , the parameters $\boldsymbol{\xi}$ are updated by minimizing the Bethe free energy under a neighborhood constraint [21] as follows,

$$\boldsymbol{\xi}^{t} = \arg\min_{\boldsymbol{\xi}} Q(\boldsymbol{\xi}), \tag{38}$$

$$Q(\boldsymbol{\xi}) = B(\boldsymbol{\xi}) + w \sum_{i} \sum_{j} \sum_{l} \sum_{qrs \in N_{ijl}} \left[\left(\rho_{ijl} - \rho_{qrs} \right)^2 + \left(\eta_{ijl} - \eta_{qrs} \right)^2 \right],$$
(39)

 1 In 3-D domain, each rectangle (i, j, l) has 6 neighbors, i.e., in the location of top, bottom, left, right, front and back.

$$B(\boldsymbol{\xi}) = -\sum_{p \in \mathcal{P}} \sum_{g=1}^{KG} \int_{\boldsymbol{w}_p} p\left(\boldsymbol{e}_g \boldsymbol{\varPhi}_p \boldsymbol{w}_{\cdot p} | \boldsymbol{y}_p; \boldsymbol{\xi}^{t-1}\right) \ln p\left(y_{gp} | \boldsymbol{w}_{\cdot p}; \boldsymbol{\xi}\right) \\ -\int_{\boldsymbol{H}} p\left(\boldsymbol{H} \mid \boldsymbol{y}; \boldsymbol{\xi}^{t-1}\right) \ln p\left(\boldsymbol{H}; \boldsymbol{\xi}\right) + \text{Const,}$$
(40)

where $p(y_{gp}|\mathbf{w}_{\cdot p}; \boldsymbol{\xi}) = \mathcal{N}_{\mathbb{C}}(y_{gp}; \boldsymbol{e}_{g}\boldsymbol{\Phi}_{p}\mathbf{w}_{\cdot p}, \sigma)$, the aposteriori distribution $p(\boldsymbol{e}_{g}\boldsymbol{\Phi}_{p}\mathbf{w}_{\cdot p}|\mathbf{y}_{p}; \boldsymbol{\xi}^{t-1})$ and $p(\boldsymbol{H}|\mathbf{y}; \boldsymbol{\xi}^{t-1})$ are obtained by AMP as $p(\boldsymbol{e}_{g}\boldsymbol{\Phi}_{p}\mathbf{w}_{\cdot p}|\mathbf{y}_{p}; \boldsymbol{\xi}^{t-1}) = \mathcal{N}_{\mathbb{C}}(\boldsymbol{e}_{g}\boldsymbol{\Phi}_{p}\mathbf{w}_{\cdot p}; m_{gp}^{t}, v_{gp}^{t})$ and $p(\boldsymbol{H}|\mathbf{y}; \boldsymbol{\xi}^{t-1}) = \prod_{m} \prod_{l} p(h_{ml} | \mathbf{y}; \boldsymbol{\xi}^{t-1})$, respectively, with $m_{gp}^{t} = (z_{gp}^{t}\sigma^{t-1} + y_{gp}\kappa_{gp}^{t})/(\sigma^{t-1} + \kappa_{gp}^{t})$ and $v_{gp}^{t} = \sigma^{t-1}\kappa_{gp}^{t}/(\sigma^{t-1} + \kappa_{gp}^{t})$. Other items that are independent of the parameters $\boldsymbol{\xi}$ are absorbed into Const. By taking a derivative of $Q(\boldsymbol{\xi})$ with respect to ρ_{ijl}, η_{ijl} and σ , respectively, and setting the three derivatives to zero with $w \to \infty$, ρ_{ijl}, η_{ijl} and σ are updated as in lines 23 and 26, respectively [21].

The proposed algorithm is summarized in table I. The initialization of the parameters $\boldsymbol{\xi}$ is consistent with [21] and are shown in line 2 of table I. Note that \hat{z}_{gmp}^t , ζ_{mp}^t , $\hat{w}_{gmp \to wmp}^t$ and ε_{ml}^t in lines 9, 17, 18 and 10 of table I, respectively, can be efficiently calculated by FFT and inverse FFT. The complexity of proposed algorithm is $O(K^2GPN_{\text{BS}}N_{\text{UE}} + KN_{\text{BS}}N_{\text{UE}}P\log_2 P)$, while the orthogonal matching pursuit (OMP) algorithm [16], the support detection (SD)-based channel estimation scheme [13], the distributed sparsity adaptive matching pursuit (DSAMP) algorithm [25] and the expectation-maximization Bernoulli-Gaussian AMP (EM-BG-AMP) algorithm [26] have a complexity order of $O(K^2GPN_{\text{BS}}N_{\text{UE}})$.

2) State Evolution of the AMP-NNSPL-DD Algorithm : We can use the SE to characterize the normalized mean square error (NMSE) performance of the proposed algorithm. The NMSE and average variance of W and H are defined as

$$e_{w}^{t} = \frac{1}{MP} \sum_{m=1}^{M} \sum_{p \in \mathcal{P}} \left(\hat{w}_{mp}^{t} - w_{mp} \right)^{2}, v_{w}^{t} = \frac{1}{MP} \sum_{m=1}^{M} \sum_{p \in \mathcal{P}} v_{w_{mp}}^{t},$$
(41)
$$t = \frac{1}{MP} \sum_{m=1}^{M} \sum_{p \in \mathcal{P}} \left(\hat{v}_{mp}^{t} - w_{mp} \right)^{2} = t = \frac{1}{MP} \sum_{m=1}^{M} \sum_{p \in \mathcal{P}} v_{w_{mp}}^{t},$$

$$e_{h}^{t} = \frac{1}{M\bar{L}} \sum_{m=1}^{\infty} \sum_{l=1}^{\infty} \left(\hat{h}_{ml}^{t} - h_{ml}\right)^{2}, \quad v_{h}^{t} = \frac{1}{M\bar{L}} \sum_{m=1}^{\infty} \sum_{l=1}^{\infty} v_{h_{ml}}^{t},$$
(42)

respectively. It is shown in [31] that in the large M limit and when the elements of the measuring matrix $\boldsymbol{\Phi}_p$ are drawn from $\mathcal{N}_{\mathbb{C}}(x; 0, 1/M)$, ϑ_{mp}^t and ω_{mp}^t can be expressed as

$$\vartheta^{t} = \frac{\sigma^{t-1} + v_{w}^{t-1}}{KG/M}, \quad \omega_{mp}^{t} = w_{mp} + \sqrt{\frac{\sigma_{0} + e_{w}^{t-1}}{KG/M}}z, \quad (43)$$

respectively, where $\sigma_0 = 1/\text{SNR}$ with SNR being the signalto-noise ratio and z admits the distribution $\mathcal{N}_{\mathbb{C}}(z; 0, 1)$. From (42), (43) and line 9 of table I, $v_{g_{mp}}^t$ is calculated as

$$v_{g_{mp}}^{t} = v^{t} \stackrel{\triangle}{=} L\left(\vartheta^{t} + v_{h}^{t}\right). \tag{44}$$

1: Input:
$$y_{p}, \Phi_{p}$$
.
2: Initialization: $\forall m, l: \rho_{ml}^{0} = 0.5, \sigma^{0} = \frac{\sum p \in p}{PKG}, \eta_{ml}^{0} = \frac{1}{p\rho_{ml}^{0}} \sum_{p \in P} \frac{\left(\|y_{p}\|_{c}^{2} - KG\sigma^{0} \right)}{\|\Phi_{p}\|_{c}^{2}}$
3: $\forall m, l: \hat{h}_{ml}^{0} = 0, v_{pl}^{0} = 1, \forall g, p: k_{gp}^{0} = 1, z_{gp}^{0} = y_{gp}$.
4: for $t = 1, \dots T$, where T is the number of iterations
5: $d' | AMP$: passing message $w_{mp} \to fgp \to w_{mp}$
6: $\forall g, p: \kappa_{gp}^{t} = \sum_{m}^{m} |\Phi_{p}(g,m)| \tilde{v}_{mn}^{t-1}, \qquad z_{gp}^{t} = \frac{\kappa_{gp}^{t}}{p} (g,m) \tilde{w}_{mp}^{t-1} - \frac{\kappa_{gp}^{t}}{q^{t-1} + \kappa_{gp}^{t}}) (y_{gp} - z_{gp}^{t-1})$
7: $\forall m, p: \vartheta_{mp}^{t} = \left(\sum_{g} \frac{|\Phi_{p}(g,m)|}{q^{t-1} + \kappa_{gp}^{t}} \right)^{-1}, \qquad \omega_{mp}^{t} = \tilde{w}_{mp}^{t-1} + \vartheta_{mp}^{t} \sum_{g} \frac{(\Phi_{p}(g,m))(y_{gp} - z_{gp}^{t-1})}{(g^{t} - \kappa_{gp}^{t})}$
8: $d' (GMP: passing message $w_{mp} \to g_{mp} \to h_{ml}$
9: $\forall m, p: v_{m,p}^{t} = L\vartheta_{mp}^{t} + \sum_{l=1}^{l} v_{l-1}^{t-1}$, $\qquad z_{gmp}^{t} = \omega_{mp}^{t} - \sum_{l=1}^{L} \vartheta_{pl} \tilde{h}_{ml}^{t-1} + \zeta_{mp}^{t-1}$
10: $\forall m, l: \epsilon_{ml}^{t} = \sum_{l=p} (\phi_{pl})^{-1} \frac{\tilde{v}_{mm}^{t}}{\tilde{v}_{mmp}^{t}} + h_{ml}^{t-1} \sum_{p} \frac{1}{v_{mm}^{t-1}} - \frac{p_{ml}^{t-1}}{\sum_{p} v_{mm}^{t}}$
11: $\forall m, l: \epsilon_{ml}^{t} = \ln \frac{e_{m}^{t}}{e_{m}^{t-1} + |\mu_{ml}^{t-1}|^{t}} - \frac{|\mu_{ml}^{t-1}|^{t}}{e_{m}^{t} + \eta_{ml}^{t-1}}, \qquad \mu_{ml}^{t} = \epsilon_{m}^{t} \epsilon_{ml}^{t}$
12: d'' The aposterior: passing message $h_{ml} \to p(h_{ml}) \to h_{ml}$
13: $\forall m, l: \tilde{v}_{ml} = \frac{q_{ml}^{t-1}}{e_{m}^{t-1} + |\mu_{ml}^{t-1}|^{t}} - \frac{|\mu_{ml}^{t-1}|^{t}}{e_{m}^{t} + \eta_{ml}^{t-1}}, \qquad \pi_{ml}^{t} = \frac{q_{ml}^{t-1}}{\mu_{ml}^{t-1} + (1-\nu_{ml}^{t-1}) \exp(\kappa_{ml})}$
14: $\forall m, l: \tilde{v}_{ml} = \pi_{ml}^{t} (\tilde{h}_{ml}^{t})^{2} (1 - \pi_{ml}^{t}) + \tilde{v}_{ml}^{t}), \qquad h_{ml}^{t} = \pi_{ml}^{t} h_{ml}^{t}$
15: $\forall m, l: v_{ml} = \frac{q_{ml}^{t-1} + |\mu_{ml}^{t} + \sum v_{ml}^{t} \psi_{mp} \psi_{mp} - \psi_{mp}} + \frac{q_{ml}^{t}}{(m_{m}^{t} + \eta_{ml}^{t})^{2}} (1 - \pi_{ml}^{t}) + \tilde{v}_{mp}^{t})$
16: $d'' (GMP: passing message h_{ml} \to g_{mp} \to w_{mp}$
17: $\forall m, p: \zeta_{mp}^{t} = \frac{1}{2} \sum_{k=1}^{t} v_{ml}^{t} + \sum v_{ml}^{t} \psi_{ml} \psi_{ml} \psi_{ml} \psi_{m$$

From (30), (32), (43) and (18), we have

$$\hat{z}_{g_{mk} \to h_{ml}}^{t} = \phi_{pl} h_{ml} + \sqrt{\frac{\sigma_0 + e_w^{t}}{KG/M}} z + \sum_{l' \neq l}^{\bar{L}} \phi_{pl'} \left(h_{ml} - \hat{h}_{h_{ml'} \to g_{mp}}^{t-1} \right)$$
(45)

$$\stackrel{(a)}{\approx} \phi_{pl} h_{ml} + \sqrt{\frac{\sigma_0 + e_w^t}{KG/M}} z + \sqrt{e_h^{t-1}} r, \qquad (46)$$

where step $\stackrel{(a)}{\approx}$ is by the assumption that $\phi_{pl'}$ are drawn from $\mathcal{N}_{\mathbb{C}}\left(\phi_{pl'}; 0, 1/\bar{L}\right)$, then in the large \bar{L} limit, the third term of (45) admits a Gaussian random with zero mean and variance e_h^{t-1} according to the central limit theorem, and the variable r in (46) admits the distribution $\mathcal{N}_{\mathbb{C}}(r; 0, 1)$. From (28) and

(46), we have

$$\varepsilon_{ml}^{t} = \frac{1}{v^{t}} \left(Ph_{ml} + \sum_{p \in \mathcal{P}} \left(\phi_{pl} \right)^{-1} \left(\sqrt{\frac{\sigma_{0} + e_{w}^{t}}{KG/M}} z + \sqrt{e_{h}^{t-1}} r \right) \right). \tag{47}$$

 $-\zeta_{mp}^t$

Similarly, from (29), (46) and (18), we have

$$\zeta_{mp}^{t} = \frac{v_h^t}{v^t} \left(w_{mp} + \bar{L} \left(\sqrt{\frac{\sigma_0 + e_w^t}{KG/M}} z + \sqrt{e_h^{t-1}} r \right) \right).$$
(48)

With (43), (44), (47) and (48), the mean \hat{h}_{ml}^t and variance $v_{h_{ml}}^t$ of h_{ml} , and the mean \hat{w}_{mp}^t and variance $v_{w_{mp}}^t$ of w_{mp} are calculate by the AMP-NNSPL-DD algorithm as shown in lines 12 and 17 of table II, respectively. Then the NMSE and average variance of W and H are updated as in lines 13 and 18, where $Dw_{mp} = p(w_{mp}) dw_{mp}$, $Dh_{ml} = p(h_{ml}) dh_{ml}$, $Dz = \mathcal{N}_{\mathbb{C}}(z;0,1) dz$ and $Dr = \mathcal{N}_{\mathbb{C}}(r;0,1) dr$. Dw_{mp} and

Table II: The State Evolution of AMP-NNSPL-DD algorithm.

1: Initialization : $\forall m, l$: $\rho_{ml}^0 = 0.5, \sigma^0 = 1, \eta_{ml}^0 = 1.$ 2: for $t = 1, \cdots T$ // AMP: passing message $w_{mp} \rightarrow f_{gp} \rightarrow w_{mp}$ 3: $\forall m, p: \ \omega_{mp}^t = w_{mp} + \sqrt{\frac{\sigma_0 + e_w^{t-1}}{KG/M}} z$ $\vartheta^t = \frac{\sigma^{t-1} + v_w^{t-1}}{KG/M},$ 4: // GMP: passing message $w_{mp} \rightarrow g_{mp} \rightarrow h_{ml}$ $v^t = L\left(\vartheta^t + v_h^{t-1}\right),$ 5: 6: $\forall m, l: \varepsilon_{ml}^{t} = \frac{Ph_{ml} + \sum_{p \in \mathcal{P}} (\phi_{pl})^{-1} \left(\sqrt{\frac{\sigma_0 + e_w^{t}}{KG/M}} z + \sqrt{e_h^{t-1}} r \right)}{v^{t}}$ $\epsilon^{t} = \frac{v^{t}}{P},$ // The aposteriori: passing message $h_{ml} \rightarrow p(h_{ml}) \rightarrow h_{ml}$ $\forall m, l: \kappa_{ml} = \ln \frac{\epsilon^{t}}{\epsilon^{t} + \eta^{t-1}} + \frac{|\mu_{ml}^{t}|^2}{\epsilon^{t}} - \frac{|\mu_{ml}^{t}|^2}{\epsilon^{t} + \eta_{ml}^{t-1}},$ 7: $\forall m, l: \ \mu_{ml}^t = \epsilon^t \varepsilon_{ml}^t$ 8: 9. $\pi_{ml}^{t} = \frac{\rho_{ml}^{t-1}}{\rho_{ml}^{t-1} + (1 - \rho_{ml}^{t-1}) \exp(-\kappa_{ml})}$ $\forall m, l: \quad \tilde{h}_{ml}^{t} = \frac{\eta_{ml}^{t-1} \mu_{ml}^{t}}{\epsilon^{t} + \eta_{ml}^{t-1}}$ 10: $\tilde{v}_{ml}^t = \frac{\eta_{ml}^{t-1} \epsilon^t}{\epsilon^t + \eta_{ml}^{t-1}},$ 11: $\forall m, l: v_{h_{ml}}^{t} = \pi_{ml}^{t} \left(\left(\tilde{h}_{ml}^{t} \right)^{2} \left(1 - \pi_{ml}^{t} \right) + \tilde{v}_{ml}^{t} \right),$ $\hat{h}_{ml}^t = \pi_{ml}^t \tilde{h}_{ml}^t$ 12: $\begin{array}{l} v_{h}^{t} = \int Dh_{ml} \int Dz \int Dr v_{h_{ml}}^{t}, \\ // \text{ GMP: passing message } h_{ml} \to g_{mp} \to w_{mp} \\ \forall m, p: \zeta_{mp}^{t} = \frac{v_{h}^{t}}{v^{t}} \left(w_{mp} + \bar{L} \left(\sqrt{\frac{\sigma_{0} + e_{w}^{t}}{KG/M}} z + \sqrt{e_{h}^{t-1}} r \right) \right) \end{array}$ $e_{h}^{t} = \int Dh_{ml} \int Dz \int Dr \left(\hat{h}_{ml}^{t} - h_{ml}\right)^{2}$ 13: 14: 15: $\forall m, p: v_{g_{mp} \to w_{mp}}^t = v_h^t,$ $\hat{w}_{g_{mp} \to w_{mp}}^{t} = \sum_{l=1}^{L} \phi_{pl} \hat{h}_{ml}^{t} - \zeta_{mp}^{t}, \\ \hat{w}_{mp}^{t} = v_{w_{mp}}^{t} \left(\frac{\omega_{mp}^{t}}{\vartheta^{t}} + \frac{\hat{w}_{g_{mp} \to w_{mp}}^{t}}{v_{g_{mp} \to w_{mp}}^{t}} \right)$ 16: $\forall m, p: \ v_{w_{mp}}^t = \frac{1}{\frac{1}{\vartheta^t} + \frac{1}{v_{g_{mp} \to w_{mp}}^t}},$ 17: $v_w^t = \int Dw_{mp} \int Dz \int Dr v_{w_{mp}}^t$, //Simultaneously upadte ρ_{ml} and η_{ml} as $e_w^t = \int Dw_{mp} \int Dz \int Dr \left(\hat{w}_{mp}^t - w_{mp} \right)^2$ 18: 19: 20: Denote: $m \stackrel{\triangle}{=} (k-1)N_{\text{BS}}N_{\text{UF}} + (i-1)N_{\text{BS}} + j$. $\begin{aligned} \forall k, i, j, l : & \pi_{kijl}^{t} = \pi_{ml}^{t}, \\ \forall k, i, j, l : & \rho_{kijl}^{t} (z, r) = \frac{1}{6} \sum_{kqrs, qrs \in N_{ijl}} \pi_{kqrs}^{t}, \\ \forall k, i, j, l : & \rho_{kijl}^{t} = \int Dz \int Dr \rho_{kijl}^{t} (z, r), \end{aligned}$ $\varpi_{kijl} = \left| \tilde{h}_{ml}^t \right|^2 + \tilde{v}_{ml}^t$ 21:
$$\begin{split} \eta_{kijl}^{t}(z,r) &= \frac{1}{6} \sum_{kqrs,qrs \in \mathcal{N}_{ijl}} \varpi_{kijl} \\ \eta_{kijl}^{t} &= \int Dz \int Dr \eta_{kijl}^{t}(z,r) \end{split}$$
22: 23: $\begin{aligned} \forall m,l: \ \rho_{ml}^{t} &= \rho_{kijl}^{t}, \\ \text{Update } \sigma \text{ as } \\ \sigma^{t} &= \frac{\sigma_{0} + e_{w}^{t}}{\left(1 + v_{w}^{t} / \sigma^{t-1}\right)^{2}} + \frac{\sigma^{t-1} v_{w}^{t}}{\sigma^{t-1} + v_{w}^{t}} \end{aligned}$ 24: $\eta_{ml}^t = \eta_{kijl}^t$ 25: 26: 27: end 28: Output : e_h^t .

 Dh_{ml} are implemented by Monte Carlo method with w_{mp} and h_{ml} generated from (18) and (10), respectively. The parameters $\rho_{ijl}^t(z,r)$ and $\eta_{ml}^t(z,r)$ are calculated by the AMP-NNSPL-DD algorithm as shown in line 22 of table II. Then ρ_{ijl}^t and η_{ml}^t are updated as in line 23. The noise variance is updated as in line 26 [21]. The SE of the AMP-NNSPL-DD algorithm is summarized in Table II.

B. The VAMP-NNSPL-DD algorithm

The factorization described by (22) can be represented by the vector-valued factor graph of Fig. 3, where $\delta_1(\cdot) = \delta_2(\cdot) = \delta(\cdot)$ is the Dirac delta function and the node with "=" represents $\mathbf{w}_{\cdot p} = \begin{bmatrix} w_{1p}, w_{2p}, \cdots, w_{Mp} \end{bmatrix}^T$ or $\mathbf{w}_{m.} = \begin{bmatrix} w_{m1}, w_{m2}, \cdots, w_{mp} \end{bmatrix}^T$. The channel $\mathbf{W} \in \mathbb{C}^{M \times P}$ in the virtual AOA-AOD-frequency domain is represented by a matrix; the factor of $\prod_{p \in \mathcal{P}} f_p(\mathbf{y}_p | \mathbf{w}_{\cdot p})$ represents P independent measurements along the column of the matrix $\mathbf{W} \in \mathbb{C}^{M \times P}$, and the factor of $\prod_{m=1}^{M} p(\mathbf{w}_m | \mathbf{h}_m)$ describes the M relationships among the channels in the delay-domain and frequencydomain along the row of the matrix. Passing messages between the factor node f_p and the variable node $w_{.p}$ of Fig. 3 obeys the vector AMP (VAMP) algorithm of [22]. Passing messages between w_m . and h_m is also based on the VAMP framework. We term these procedures of message passing as the VAMP-NNSPL-DD algorithm and the corresponding its SE, as follows.

1) The VAMP-NNSPL-DD algorithm: The VAMP-NNSPL-DD algorithm is formally stated in Table III. Please refer to the Appendix for a detailed derivation of the algorithm. As presented in Table III, lines 6-13 describe three similar modules. Consider the first module for example, where lines 6-7 perform the LMMSE estimation at the factor node f_p , where the function $g(\cdot)$ is the LMMSE estimator defined in the Appendix. Specifically, providing the *apriori* information of $w_{.p}$ by $\mathcal{N}_{\mathbb{C}}\left(w_{.p}; \omega_{.p}^{t}, \gamma^{t} I_{M}\right)$ (intput information), and measurements of $\mathbf{y}_{p} = \mathbf{\Phi}_{p} w_{.p} + \mathbf{n}_{p}$ with $\mathbf{n}_{p} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{n}_{p}; 0, \sigma^{t-1} I_{KG}\right)$,

1: Input :
$$\mathbf{y}_{p}, \mathbf{\Phi}_{p}$$
.
2: Initialization : $\forall m, l : \rho_{ml}^{0} = 0.5, \sigma^{0} = \frac{\sum_{p \in \mathcal{P}} \|\mathbf{y}_{p}\|_{2}^{2}}{PKG}, \eta_{ml}^{0} = \frac{1}{P\rho_{ml}^{0}} \sum_{p \in \mathcal{P}} \frac{\left(\|\mathbf{y}_{p}\|_{2}^{2} - KG\sigma^{0}\right)}{\|\mathbf{\Phi}_{p}\|_{F}^{2}}$
3: $\forall p : \omega_{p}^{1} = 0, \forall m : \mathbf{x}_{m}^{0} = 0, \gamma^{1} = 100, \mathbf{x}^{0} = 100.$
4: for $t = 1, \dots, T$, where T is the number of iterations
5: $//LMMSE$: transform message $w_{p} \to f_{p} \to w_{p}$
6: $\forall p : v_{w,p}^{t} = \frac{1}{M} \operatorname{Tr} \left[\gamma^{t} g' \left(\omega_{p}^{t}, \gamma^{t}, \mathbf{y}_{p}, \mathbf{\Phi}_{p}, \sigma^{t-1} \right) \right], \qquad \mathbf{x}_{p}^{t} = g \left(\omega_{p}^{t}, \gamma^{t}, \mathbf{y}_{p}, \mathbf{\Phi}_{p}, \sigma^{t-1} \right)$
7: $\forall p : \zeta_{p}^{t} = (1/v_{w,p}^{t} - 1/\gamma^{t})^{-1}, \qquad r^{t,p} = \zeta^{t} \left(\hat{u}_{w,p}^{t} / v_{w,p}^{t} - \omega_{p}^{t} / \gamma^{t} \right)$
8: $//LMMSE$: transform message $w_{m} \to \delta_{2} \to \tilde{h}_{m} \to h_{m}$
9: $v_{h}^{t} = \frac{1}{L} \operatorname{Tr} \left[\kappa_{m}^{t-1} g' \left(\mathbf{x}_{m}^{t-1}, \mathbf{x}_{m-1}^{t-1}, \mathbf{x}_{m}^{t-1}, \mathbf{x}_{p}^{t} + \mathbf{y}_{p} \mathbf{y}_{p}^{t} \right], \qquad \forall m : \hat{u}_{h_{m}}^{t} = g \left(\mathbf{x}_{m}^{t-1}, \mathbf{x}_{m-1}^{t-1}, \mathbf{x}_{p}^{t-1} + \mathbf{y}_{p} \mathbf{y}_{p} \mathbf{y}_{p} \right)$
10: $\kappa^{t} = \left(1/v_{h}^{t} - 1/\tilde{\kappa}^{t-1} \right)^{-1}, \qquad \forall m : s_{m}^{t} = \kappa^{t} \left(\hat{u}_{h_{m}}^{t} / v_{h}^{t} - \mathbf{x}_{m-1}^{t-1} / \tilde{\kappa}_{p}^{t-1} \right)$
11: $//\text{The aposteriori: transform message $h_{m} \to p(h_{m}) \to h_{m} \to \tilde{h}_{m}$
12: $v_{h}^{t} = \frac{1}{M} \sum_{m} \langle gvar(s_{m}^{t}, \kappa^{t}; \xi^{t-1}) \rangle, \qquad \forall m : \hat{u}_{h_{m}}^{t} = gmean(s_{m}^{t}, \kappa^{t}; \xi^{t-1})$
13: $\tilde{\kappa}^{t} = (1/v_{h}^{t} - 1/\tilde{\kappa}^{t})^{-1}, \qquad \forall m : \tilde{u}_{h_{m}}^{t} = \tilde{\kappa}^{t} \left(\hat{u}_{h_{m}}^{t} / v_{h}^{t} - s_{m}^{t} / \kappa^{t} \right)$
14: $//\text{transform message} \tilde{h}_{m} \to w_{m}.$
15: $\gamma^{t+1} = \tilde{\kappa}^{t}, \qquad \forall m : \omega_{m}^{t+1} = \phi \tilde{s}_{m}^{t}$
16: Updating ρ_{ml} and η_{ml} similar to lines 21-24 of Table I, respectively.
17: $//\text{Updating } \sigma$ as
18: $\sigma^{t} = \frac{1}{P} \sum_{p} \left[\left(\left| \mathbf{y}_{p} - \mathbf{\Phi}_{p} \hat{u}_{w,p}^{t} \right|^{2} \right) + \frac{1}{KG} \operatorname{Tr} \left(\mathbf{\Phi}_{p} \mathbf{\Phi}_{p}^{H} \right) \right]$$





Figure 3: Vector-valued factor graph representation of the broadband mmWave massive MIMO-OFDM Systems.

the refined information $\mathcal{N}_{\mathbb{C}}\left(\boldsymbol{w}_{\cdot p}; \hat{\boldsymbol{u}}_{\boldsymbol{w},p}^{t}, \boldsymbol{v}_{\boldsymbol{w},p}^{t}\boldsymbol{I}_{M}\right)$ (ouput information) is obtained for $\boldsymbol{w}_{\cdot p}$ by LMMSE estimation. Next, the extrinsic information² of $\boldsymbol{w}_{\cdot p}$ is given by $\mathcal{N}_{\mathbb{C}}\left(\boldsymbol{w}_{\cdot p}; \boldsymbol{r}_{\cdot p}^{t}, \zeta_{p}^{t}\boldsymbol{I}_{M}\right)$ (line 7), which is fed into the next module as the *apriori* information. The first module (lines 6-7) corresponds to pass-

ing the messages gleaned from measurements to channels in the virtual AOA-AOD-frequency domain. The second module of Table III (line 9-10) corresponds to passing the messages from channels in the virtual AOA-AOD-frequency domain to those in the virtual AOA-AOD-delay domain. The third module of Table III (lines 12-13) corresponds to refining the messages of channels in the virtual AOA-AOD-delay domain by taking the *apriori* distribution into consideration. Line 15 passes the messages backward from the virtual AOA-AODdelay domain to the virtual AOA-AOD-frequency domain under the constraint $\delta (\mathbf{w}_m - \boldsymbol{\phi} \boldsymbol{\check{h}}_m)$, which updates the apriori information in the first module (lines 6-7). Lines 16-18 update the parameter $\boldsymbol{\xi}$ by minimizing the Bethe free energy under neighborhood constraint, which is similar to (38). Line 9 implies that passing the message $\mathcal{N}_{\mathbb{C}}\left(\boldsymbol{w}_{m}; \boldsymbol{r}_{m}^{t}, \frac{1}{P}\sum_{p}\zeta_{p}^{t}I_{P}\right)$ of w_m to \check{h}_m corresponds to applying the LMMSE estimation on the following model

$$\boldsymbol{r}_{m}^{t} = \boldsymbol{\phi} \boldsymbol{\check{h}}_{m} + \boldsymbol{n}_{m}, \tag{49}$$

where $\mathbf{n}_m \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{n}_m; 0, \frac{1}{P}\sum_p \zeta_p^t I_P\right)$. For $m = 1, \dots, M$, (49) is a multiple-measurement-vector (MMV) problem [25], which represents simultaneously recovering multiple vectors $\{\check{\mathbf{h}}_m\}$ from multiple measurements $\{\mathbf{r}_m^t\}$ with the aid of a common measurement matrix $\boldsymbol{\phi}$. The MMV is helpful for interpreting the mismatch between the AMP-NNSPL-DD and its SE, as we will discuss in Section IV.

Matrix inversion in the LMMSE estimator can be avoided by invoking the singular value decomposition (SVD) of $\Phi_p = U_p \operatorname{diag}(s_p) V_p^H$ and $\phi = U_\phi \operatorname{diag}(s_\phi) V_\phi^H$. For more details, we refer the motivated reader to [22]. The SVD is precomputed off-line and corresponding results are saved. The

²Suppose $N_{\mathbb{C}}(x; a, b)$ and $N_{\mathbb{C}}(x; c, d)$ is the intput information and ouput information of a estimator, respectively. Then the extrinsic information of the estimator is defined as $N_{\mathbb{C}}(x; e, f) \propto N_{\mathbb{C}}(x; c, d) / N_{\mathbb{C}}(x; a, b)$. Generally, the extrinsic information is fed into the other module as a intput information [38].

1: Input : y_p, Φ_p . 2: Initialization : the same as that of lines 2-3 of Table III. 3: for $t = 1, \dots T$, where T is the number of iterations //LMMSE: transform message $w_p \rightarrow f_p \rightarrow w_p$ 4: $\begin{aligned} \forall p: \ \boldsymbol{d}_{p} &= \frac{\frac{1}{\sigma^{t-1}} \boldsymbol{s}_{p}^{2}}{\frac{1}{\gamma^{t}} + \frac{1}{\sigma^{t-1}} \boldsymbol{s}_{p}^{2}}, \qquad \boldsymbol{\zeta}_{p}^{t} &= \gamma^{t} \frac{M - R_{p} \langle \boldsymbol{d}_{p} \rangle}{R_{p} \langle \boldsymbol{d}_{p} \rangle}, \qquad \boldsymbol{r}_{\cdot p}^{t} &= \boldsymbol{w}_{\cdot p} + \sqrt{\boldsymbol{\zeta}_{p}^{t}} \boldsymbol{z}_{1} \\ \forall p: \ \boldsymbol{v}_{\boldsymbol{w},p}^{t} &= \gamma^{t} \left[1 - \frac{1}{M} R_{p} \left\langle \boldsymbol{d}_{p} \right\rangle \right], \qquad \boldsymbol{\hat{u}}_{\boldsymbol{w},p}^{t} &= \int D \boldsymbol{z}_{1} \boldsymbol{v}_{\boldsymbol{w},p}^{t} \end{aligned}$ 5: $\hat{\boldsymbol{u}}_{\boldsymbol{w},p}^{t} = \int D\boldsymbol{z}_{1} \boldsymbol{v}_{\boldsymbol{w},p}^{t} \left(\frac{\boldsymbol{r}_{\cdot p}^{t}}{\boldsymbol{\zeta}_{p}^{t}} + \frac{\boldsymbol{\omega}_{\cdot p}^{t}}{\boldsymbol{\gamma}^{t}} \right),$ 6: $\begin{aligned} & \forall p \cdot v_{\mathbf{w},p} \leq t \quad \text{i. In the second states of the second sta$ 7: 8: 9: 10: 11: 12: $e_{H}^{t} = \int DH \int Dz_{2} \frac{1}{M\bar{L}} \left(\left[\hat{\boldsymbol{u}}_{\boldsymbol{h}_{1}}^{t}, \hat{\boldsymbol{u}}_{\boldsymbol{h}_{2}}^{t}, \cdots, \hat{\boldsymbol{u}}_{\boldsymbol{h}_{M}}^{t} \right]^{\mathsf{T}} - H \right)^{2}$ 13: //transform message $\check{h}_m \to w_m$. 14: $\gamma^{t+1} = \check{\kappa}^t, \qquad \forall m : \omega_{m}^{t+1} = \phi \check{s}_m^t$ Updating ρ_{ml} and η_{ml} similar to lines 20-24 of Table II, respectively. 15: 16: //Updating σ as 17: $\sigma^{t} = \frac{1}{P} \sum_{p} \left[\left\langle \left| \boldsymbol{y}_{p} - \boldsymbol{\Phi}_{p} \hat{\boldsymbol{u}}_{\boldsymbol{w},p}^{t} \right|^{2} \right\rangle + \frac{1}{KG} v_{\boldsymbol{w},p}^{t} R_{p} \left\langle s_{p}^{2} \right\rangle \right]$ 18: 19: end 20: Output : e_H^t

complexity of the proposed algorithm is on the order of $O(K^2 R_p P N_{\text{BS}} N_{\text{UE}} + K N_{\text{BS}} N_{\text{UE}} R_{\phi} \bar{L})$, where R_p and R_{ϕ} are the rank of Φ_p and ϕ , respectively. Hence, compared with the AMP-NNSPL-DD algorithm, the VAMP-NNSPL-DD requires more storage space and has a higher complexity.

2) State Evolution of the VAMP-NNSPL-DD Algorithm : It is shown in [22] that when Φ_p is large and rightrotationally invariant, r_{p}^{t} in line 7 of Table III appears as the true component w_{p} corrupted by Gaussian noise with a variance of ζ_p^t , which is shown in line 5 of Table IV. In line 5 of Table IV, ζ_p^t is calculated by substituting the SVD of $\Phi_p = U_p \operatorname{diag}(s_p) V_p^H$ into lines 6-7 of Table III, where the equation for calculating d_p refers to the element-wise division between two vectors. And z_1 admits the distribution of $\mathcal{N}_{\mathbb{C}}(z_1; 0, I_M)$. Similarly, line 9 of table IV is obtained from lines 9-10 of Table III, where z_2 obeys the distribution of $\mathcal{N}_{\mathbb{C}}(z_2; 0, I_{\bar{L}})$. Next, other quantities are calculated by the VAMP-NNSPL-DD and are shown in Table IV. The operations $\int Dz_1$ and $\int Dz_2$ in lines 6 and 12 refer to $\int \mathcal{N}_{\mathbb{C}}(z_1; 0, I_M) dz_1$ and $\int \mathcal{N}_{\mathbb{C}}(z_2; 0, I_{\bar{L}}) dz_2$, respectively. Furthermore, $\int DH = \int p(H) dH$ in line 13 is implemented by the classic Monte Carlo method with H generated from (10). The quantities of $v_{\boldsymbol{w},p}^t$, $\hat{\boldsymbol{u}}_{\boldsymbol{w},p}^t$ (line 6 of Table IV), $\check{\boldsymbol{s}}_m^t$, and ω_m^{t+1} (right part of lines 12 and 15, respectively) are used for updating the noise variance σ^t (line 18). But if noise variance σ is known, these quantities could be removed. In such a case, the SE is not dependent on y_p , Φ_p and ϕ , but dependent on s_p and s_{ϕ} (lines 5 and 10, respectively). In other words, the SE of VAMP-NNSPL-DD takes the specific character of the measurement matrix into considered. By contrast, the SE based on AMP assumes that the elements of the measurement matrix obey the independent and identically distributed (i.i.d.) Gaussian distribution. Hence, it is expected that the SE based on VAMP is capable of characterizing the performance bound better than that based on AMP.

IV. EXPERIMENTAL RESULTS

Let us consider a broadband mmWave system, where the downlink transmissions are organized in N = 8192 OFDM symbols in a bandwidth of B = 800 MHz at the carrier frequency of $f_0 = 28$ GHz. We compare the NMSE performance of as many as ten CS algorithms, i.e. of the Basis Pursuit (BP) [23], of the group LASSO³ [24], of the OMP algorithm [16], of the SD-based channel estimation scheme [13], of the DSAMP algorithm⁴ [25], of the EM-BG-AMP algorithm [26], of the sparse non-informative parameter estimator-based cosparse analysis AMP for imaging (SCAMPI) algorithm [17], of the AMP-NNSPL-FD algorithm , and finally of the (V)AMP-NNSPL-DD algorithm, with various measuring time durations *G*, SNRs, angular spreads and cluster durations. The NMSE is defined as

$$\mathsf{NMSE}[\mathsf{dB}] = 10\mathsf{lg}\left[\mathsf{E}\left(\frac{\left\|\mathsf{vec}\left(\hat{\boldsymbol{H}} - \boldsymbol{H}\right)\right\|_{2}^{2}}{\left\|\mathsf{vec}\left(\boldsymbol{H}\right)\right\|_{2}^{2}}\right)\right], \quad (50)$$

³A version of LASSO that can learn clustered structure of signals.

⁴Another version of the algorithm [14] that can exploit the SCS. In the simulation, only the DSAMP is compared with other CS algorithms, since the algorithm [14] was mainly designed for the mmWave system without angular spread.



(a) Comparing the proposed algorithms with other CS algorithms.



(b) Comparing the proposed algorithms with the oracle LS.

Figure 4: NMSE performance versus measuring time duration *G* for 4 clusters with angular spread 15 degrees and cluster duration 9 ns. Simulation setting: $N_{\text{BS}} = 64$, $N_{\text{UE}} = 16$, P = 512, SNR= 10 dB, K = 4 and B = 800 MHz.

where \hat{H} is the output of channel estimation, or it is transformed according to (18) when the channel of the virtual AOA-AOD-frequency domain is estimated.

In the simulations, we use a relatively large number of pilots of P = 512 given the broadband of mmWave systems. Assuming that the number of users is K = 4, the number of clusters is 4, each cluster has $1 \sim 10$ physical paths, where the complex path gains are drawn from $N_{\mathbb{C}} (\beta_i; 0, 10^{\text{power/10}})$, with the path-power being 0 dBm for the LOS component and $-5 \sim -20$ dBm for the NLOS component. The number of antennas at BS and user is set as $N_{\text{BS}} = 64$ and $N_{\text{UE}} = 16$, respectively. The elements of the precoders $(f_{\text{RF}tk}, f_{\text{BB}ntk})$ and combiners $(W_{\text{RF}}, W_{\text{BB}})$ are of the form of $e^{j\phi}$, where ϕ is randomly and uniformly selected from the set of quantized angles $\{45, 135, 225, 315\}/180 \times \pi$ [11], [13], [14], [17].

Consider the scenario of having 4 clusters, an angular spread of 15 degrees [7] and a cluster duration of about 9 ns [19]. Fig. 4 (a) compares the NMSE performance of these CS algorithms versus measuring time durations recorded at SNR=10 dB. It is clearly seen that the group LASSO performs slightly better

than the BP, since it partially exploits the clustered structure. By contrast, the SD and the DSAMP algorithm outperform the OMP algorithm with short measuring time. This is because that the SD and the DSAMP algorithm exploit the clustered structure in the virtual angular and subchannel dimension, respectively. The SD and the DSAMP algorithm perform no better than the OMP algorithm with long measuring time. This is because both the SD and the DSAMP algorithm only recover those channel coefficients that are in the vicinity of the cluster center, while neglecting those small coefficients which are far away from the cluster center. The SCAMPI algorithm outperforms many other algorithms, since it exploits the subtle changes between the adjacent channel elements. By exploiting the clustered structure in the virtual AOA-AODfrequency domain, the AMP-NNSPL-FD performs slightly better than the SCAMPI. The SCAMPI is derived in a rigorous way in terms of exploiting the subtle changes. Although our proposed algorithm is indeed heuristic in terms of exploiting the clustered structure under the neighborhood constraint, we characterize the performance bound of the AMP-NNSPL-FD by the SE derived in [21]. For improving the attainable channel estimation performance, the (V)AMP-NNSPL-DD jointly exploit the 3-D clustered structure in virtual AOA-AOD-delay domain, and obtain a 7.6 dB NMSE performance gain compared with the AMP-NNSPL-FD at G = 550.

Let us analyze the performance of the AMP-NNSPL-DD by SE. The performance of the AMP-NNSPL-DD is quite accurately predicted by its SE with long measuring time, but not so well with short measuring time. We interpret these phenomena as follows. Model (49) implies that passing message from channels in the virtual AOA-AOD-frequency domain to those in the virtual AOA-AOD-delay domain by the proposed algorithm is equivalent to a MMV problem. However, the derivation of SE need make an idealized simplifying assumption that elements of measurement matrix are i.i.d. random variables, which results in a generalized MMV (GMMV) problem [25] of $\mathbf{r}_{m}^{t} = \mathbf{A}_{m} \mathbf{\check{h}}_{m} + \mathbf{n}_{m}$, where the *M* measurement matrices $\{A_m\}$ are not expected to be the same, since they are random matrices. The GMMV represents simultaneously recovering multiple vectors from multiple measurements with the aid of different measurement matrices, where the diversity nature of the different measurement matrices provides performance gains⁵ [25]. Experimented results of [25] show that the MMV recovers multiple vectors less reliably than the GMMV with short measuring time, but achieves the same performance as that of GMMV with long measuring time. Therefore, it is reasonable to expect that the AMP-NNSPL-DD and its SE do not match well with short measuring time, but match better with long measuring time. On the other hand, it is observed from Fig. 4 (a) that the SE of AMP-NNSPL-DD performs the same as the SE [21] applied to the model (20). This means

⁵This can be readily verified by simulations with synthetic data. Specifically, Fig. 3 is helpful for describing the generation of the synthetic data. One can generate each row of the matrix $W \in \mathbb{C}^{M \times P}$ by different partial DFT matrices from h_m , then obtain *P* measurements along the column of $W \in \mathbb{C}^{M \times P}$. Next, h_m can be recover by the AMP-NNSPL-DD with those measurements. In such synthetic case, the performance of the AMP-NNSPL-DD by ould approach the performance bound characterized by its SE according to our simulation, although it is not presented here.

that both of them characterize the performance bound when elements of Ψ obey i.i.d. Gaussian distribution. However, the MMV problem results in elements in the equivalent measurement matrix Ψ of (20) that do not obey the i.i.d. Gaussian distribution. These also explains the mismatch between the AMP-NNSPL-DD and the corresponding SE. Comparing with the SE of AMP-NNSPL-FD, the SE of AMP-NNSPL-DD provide an insight that if elements of measurement matrix obey i.i.d. Gaussian distribution, estimating channels in angulardelay domain would also attain a considerable performance gain with short measuring time. Hence, it is beneficial to design a measurement matrix, whose elements obey the i.i.d. Gaussian distribution, for obtaining the potential performance gain with short measuring time. Additionally, it is observed from Fig. 4 (a) that the SE of VAMP-NNSPL-DD characterizes the performance bound more accurately.

The oracle LS associated with known support⁶ is usually considered as the performance bound. However, mmWave channels in the virtual domain are approximate sparse, i.e., none of elements are expected to be exactly zero. Therefore, we define a threshold $\alpha \in (0, 1)^7$, and assume that the support only includes the location of those elements whose amplitudes are higher than the threshold α , but not the other elements whose amplitudes are lower than the threshold α . In this case, given different thresholds, one can obtain different supports, which corresponds to different location information of clusters. The performance of the oracle LS with different thresholds is shown in Fig. 4 (b). When $\alpha = 0$, i.e., none cluster location information, the performance of LS is poor as it respects to an under-determined problem (G < M). When $\alpha > 0$, the oracle LS significantly outperforms the LS which implies that the location information of clusters is critical for achieving accurate channel estimation. Furthermore, it is observed in Fig. 4 (b) that with different measuring time durations, LS requires different cluster location information to acquire attainable channel estimation performance. However, the cluster location information is cite-specific and typically unknown. By contrast, the proposed AMP-NNSPL-FD algorithm is capable of adaptively learning the cluster location information under a neighborhood constraint, which makes the proposed algorithm well estimating the channel.

Fig. 5 (a) characterizes the convergence of the proposed algorithms. The SE of AMP-NNSPL-FD accurately characterizes the corresponding performance bound, and the AMP-NNSPL-FD converges within 20 iterations. The SE of AMP-NNSPL-DD also converges within 20 iterations, which characterizes the performance bound when elements of measurement matrix obey i.i.d. Gaussian distribution. However, the AMP-NNSPL-DD converges in about 60 iterations, while the SE of the VAMP-NNSPL-DD matches the corresponding algorithm more accurately.

The NMSE performance of these CS algorithms versus the SNR recorded for a measuring time duration of G = 500 is shown in Fig. 5 (b), where we can see that the Group Lasso performs slightly better than the BP. By contrast, the



(a) NMSE performance versus iteration times.



Figure 5: NMSE performance for 4 clusters with angular spread 15 degrees and cluster duration 9 ns. Simulation setting: $N_{\text{BS}} = 64$, $N_{\text{UE}} = 16$, P = 512, G = 500 and B = 800 MHz.

SD and the DSAMP algorithm are capable of outperforming the OMP algorithm for SNRs below 5 dB. The SCAMPI and the AMP-NNSPL-FD outperform the EM-BG-AMP, and the performance of the AMP-NNSPL-FD is well predicted by the SE [21]. The proposed (V)AMP-NNSPL-DD algorithm outperforms other CS algorithms.

Fig. 6 (a) compares the NMSE versus angular spread performance of all these algorithms at SNR=10 dB with a measuring time duration of G = 500. It is shown that the angular spread obeys a near-exponential distribution with a mean of 15 degrees [7]. When the angular spread appears to be at its maximum value of about 40 degrees [7], each CS algorithm still works well, comparing to that of at an angular spread of 10 degrees.

Fig. 6 (b) compares the NMSE versus cluster duration performance of all these algorithms at SNR=10 dB for a measuring time duration of G = 500. It is shown that the cluster duration of a cluster obeys a near-exponential distribution with a mean of 9 ns [19]. It is observed that the performance of each CS algorithm remains similar, when the cluster duration is within 45 ns.

⁶The location of non-zero values in a sparse vector.

⁷Without loss of generality, supposing that the largest amplitude of the channel elements is normalized to 1.



(a) NMSE performance versus angular spread for 4 clusters with cluster duration 9 ns.



(b) NMSE performance versus cluster duration for 4 clusters with angular spread 15 degrees.

Figure 6: Simulation setting: $N_{BS} = 64$, $N_{UE} = 16$, P = 512, G = 500, SNR= 10 dB, K = 4 and B = 800 MHz.

V. CONCLUSIONS

By exploiting the 3-D clustered structure exhibited in the virtual AOA-AOD-delay domain with the aid of neighborhood parameter constraint, we proposed an algorithm termed as (V)AMP-NNSPL-DD for estimating broadband mmWave massive MIMO-OFDM channels. Our simulation results demonstrate that the proposed algorithm is capable of attaining a considerable performance gain, over the algorithms that only exploit the sparse clustered structure in the angular domain. The AMP-NNSPL-DD requires less storage space and has a lower complexity. The VAMP-NNSPL-DD provided an insight into interpreting the mismatch between the AMP-NNSPL-DD and its SE, while the SE of VAMP-NNSPL-DD characterizes the performance gains with shorter measuring time will be investigated in our future work.

Appendix

We fixed the parameters $\boldsymbol{\xi}$ to the values estimated at the previous iteration, i.e., $\boldsymbol{\xi} = \boldsymbol{\xi}^{t-1}$. We pass messages on the factor graph that is shown in Fig. 3 as follows. At the *t*th

iteration, by setting the message passed from the variable node $w_{\cdot p}$ to the factor node f_p as

$$u_{\boldsymbol{w}\cdot p}^{t} \to f_{p} \left(\boldsymbol{w} \cdot p \right) = \mathcal{N}_{\mathbb{C}} \left(\boldsymbol{w} \cdot p ; \boldsymbol{\omega}_{\cdot p}^{t}, \boldsymbol{\gamma}^{t} I_{M} \right),$$
(51)

then the belief of w_{p} at the factor node f_p is formulated as

$$\beta^{t}\left(\boldsymbol{w}_{\cdot p}\right) = u_{\boldsymbol{w}_{\cdot p} \to f_{p}}^{t}\left(\boldsymbol{w}_{\cdot p}\right) \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{y}_{p}; \boldsymbol{\Phi}_{p} \boldsymbol{w}_{\cdot p}, \boldsymbol{\sigma}^{t-1}\right)$$
$$\propto \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{w}_{\cdot p}; \hat{\boldsymbol{u}}_{\boldsymbol{w}_{\cdot p}}^{t}, \boldsymbol{v}_{\boldsymbol{w}_{\cdot p}}^{t} I_{P}\right),$$
(52)

with

$$\hat{\boldsymbol{u}}_{\boldsymbol{w},p}^{t} = \left(\frac{1}{\sigma^{t-1}}\boldsymbol{\Phi}_{p}^{\mathsf{H}}\boldsymbol{\Phi}_{p} + \frac{1}{\gamma^{t}}I_{P}\right)^{-1} \left(\frac{1}{\sigma^{t-1}}\boldsymbol{\Phi}_{p}^{\mathsf{H}}\boldsymbol{y}_{p} + \frac{\boldsymbol{\omega}_{p}^{t}}{\gamma^{t}}\right)$$

$$\triangleq g\left(\boldsymbol{\omega}_{p}^{t}, \boldsymbol{\gamma}^{t}, \boldsymbol{y}_{p}, \boldsymbol{\Phi}_{p}, \sigma^{t-1}\right), \qquad (53)$$

$$v_{\boldsymbol{w},p}^{t} = \frac{1}{M}\operatorname{Tr}\left[\left(\frac{1}{\sigma^{t-1}}\boldsymbol{\Phi}_{p}^{\mathsf{H}}\boldsymbol{\Phi}_{p} + \frac{1}{\gamma^{t}}I_{P}\right)^{-1}\right]$$

$$\triangleq \frac{1}{M}\operatorname{Tr}\left[\boldsymbol{\gamma}^{t}g^{\prime}\left(\boldsymbol{\omega}_{p}^{t}, \boldsymbol{\gamma}^{t}, \boldsymbol{y}_{p}, \boldsymbol{\Phi}_{p}, \sigma^{t-1}\right)\right], \qquad (54)$$

where $g\left(\omega_{\cdot p}^{t}, \gamma^{t}, \mathbf{y}_{p}, \mathbf{\Phi}_{p}, \sigma^{t-1}\right)$ denotes the LMMSE estimator, and $g'\left(\omega_{\cdot p}^{t}, \gamma^{t}, \mathbf{y}_{p}, \mathbf{\Phi}_{p}, \sigma^{t-1}\right)$ is the derivative of g with respect to $\omega_{\cdot p}^{t}$. This yields line 6 of Table III. According to the sum-product message passing rule of [36], the message passed from the factor node f_{p} to the variable node $w_{\cdot p}$ is calculated by $u_{f_{p}\to w_{\cdot p}}^{t}\left(w_{\cdot p}\right) = \beta^{t}\left(w_{\cdot p}\right)/u_{w_{\cdot p}\to f_{p}}^{t}\left(w_{\cdot p}\right) \propto \mathcal{N}_{\mathbb{C}}\left(w_{\cdot p}; \mathbf{r}_{\cdot p}^{t}, \zeta_{p}^{t}I_{M}\right)$, where the variance and mean are shown in line 7 of Table III. Again, according to the sum-product message passing rule, the message $u_{f_{p}\to w_{\cdot p}}^{t}\left(w_{\cdot p}\right)$ flows through the variable node $\mathbf{W} \in \mathbb{C}^{M \times P}$, and manifests itself as $u_{w_{m}\to\delta_{2}}^{t}\left(w_{m}\right) = \mathcal{N}_{\mathbb{C}}\left(w_{m}:\mathbf{r}_{m}^{t}, \zeta^{t}I_{P}\right)$ with $\zeta^{t} = \frac{1}{P}\sum_{p}\zeta_{p}^{t}$. Let us suppose that at the (t-1)th iteration, the message passed from the variable node $\mathbf{\check{h}}_{m}$ to the factor node δ_{2} is given by

$$u_{\check{\boldsymbol{h}}_{m}\to\delta_{2}}^{t-1}\left(\check{\boldsymbol{h}}_{m}\right)=\mathcal{N}_{\mathbb{C}}\left(\check{\boldsymbol{h}}_{m};\check{\boldsymbol{s}}_{m}^{t-1},\check{\boldsymbol{\kappa}}^{t-1}\boldsymbol{I}_{L}\right),\tag{55}$$

then the belief of $\check{\mathbf{h}}_m$ at the factor node δ_2 reads

$$\beta^{t} \left(\check{\boldsymbol{h}}_{m} \right) = \int u_{\check{\boldsymbol{h}}_{m} \to \delta_{2}}^{t-1} \left(\check{\boldsymbol{h}}_{m} \right) \delta \left(\boldsymbol{w}_{m} - \boldsymbol{\phi} \check{\boldsymbol{h}}_{m} \right) u_{\boldsymbol{w}_{m} \to \delta_{2}}^{t} \left(\boldsymbol{w}_{m} \right) d\boldsymbol{w}_{m}.$$

$$\propto \mathcal{N}_{\mathbb{C}} \left(\check{\boldsymbol{h}}_{m}; \hat{\boldsymbol{u}}_{\check{\boldsymbol{h}}_{m}}^{t}, v_{\check{\boldsymbol{h}}}^{t} I_{L} \right), \qquad (56)$$

where the variance and mean are shown in line 9 of Table III. The message passed from the factor node δ_2 to the variable node $\check{\mathbf{h}}_m$ is calculated as $u^t_{\delta_2 \to \check{\mathbf{h}}_m} (\check{\mathbf{h}}_m) = \beta^t (\check{\mathbf{h}}_m) / u^{t-1}_{\check{\mathbf{h}}_m \to \delta_2} (\check{\mathbf{h}}_m) \propto \mathcal{N}_{\mathbb{C}} (\check{\mathbf{h}}_m; s^t_m, \kappa^t I_L)$. In a sequel, the message $u^t_{\delta_2 \to \check{\mathbf{h}}_m} (\check{\mathbf{h}}_m)$ flows leftward, and manifests itself as $u^t_{\delta_1 \to \mathbf{h}_m} (\mathbf{h}_m) = \mathcal{N}_{\mathbb{C}} (\mathbf{h}_m; s^t_m, \kappa^t I_L)$, where the variance and mean are shown in line 10 of Table III. Similar to (33)-(35), the *aposteriori* distribution of \mathbf{h}_m reads

$$p\left(\boldsymbol{h}_{m}|\boldsymbol{y};\boldsymbol{\xi}^{t-1}\right) = \frac{1}{\int p\left(\boldsymbol{h}_{m}\right)u_{\delta_{1}\to\boldsymbol{h}_{m}}^{t}\left(\boldsymbol{h}_{m}\right)d\boldsymbol{h}_{m}} \times p\left(\boldsymbol{h}_{m}\right)u_{\delta_{1}\to\boldsymbol{h}_{m}}^{t}\left(\boldsymbol{h}_{m}\right), \quad (57)$$

and the *aposteriori* variance $v_{\mathbf{h}}^{t}$ and mean $\hat{u}_{\mathbf{h}_{m}}^{t}$ of \mathbf{h}_{m} are shown in line 12 of Table III, where the function $g_{\text{mean}}\left(\mathbf{s}_{m}^{t}, \kappa^{t}; \boldsymbol{\xi}^{t-1}\right)$ and $g_{\text{var}}\left(\mathbf{s}_{m}^{t}, \kappa^{t}; \boldsymbol{\xi}^{t-1}\right)$ are defined in (34) and (35), respectively, which are the element-wise functions of vector \mathbf{s}_{m}^{t} . Next, the message passed from the variable node \mathbf{h}_{m} to the factor node δ_{1} is calculated by $u_{\mathbf{h}_{m}\to\delta_{1}}^{t}\left(\mathbf{h}_{m}\right) =$ $\mathcal{N}_{\mathbb{C}}\left(\mathbf{h}_{m}; \hat{u}_{\mathbf{h}_{m}}^{t}, v_{\mathbf{h}}^{t}\right)/u_{\delta_{1}\to\mathbf{h}_{m}}^{t}\left(\mathbf{h}_{m}\right) \propto \mathcal{N}_{\mathbb{C}}\left(\mathbf{h}_{m}; \mathbf{s}_{m}^{t}, \mathbf{k}^{t}I_{L}\right)$. Then the message $u_{\mathbf{h}_{m}\to\delta_{1}}^{t}$ passes rightward, and manifests itself as $u_{\mathbf{h}_{m}\to\delta_{2}}^{t}\left(\mathbf{h}_{m}\right) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{h}_{m}; \mathbf{s}_{m}^{t}, \mathbf{k}^{t}I_{L}\right)$, where the variance and mean are shown in line 13 of Table III, which updates the message defined in (55). Next, the belief of \mathbf{w}_{m} . at the factor node δ_{2} is calculated as

$$\beta^{t} (\boldsymbol{w}_{m}.) = \int u_{\boldsymbol{\check{h}}_{m}\to\delta_{2}}^{t} \left(\boldsymbol{\check{h}}_{m}\right) \delta\left(\boldsymbol{w}_{m}.-\boldsymbol{\phi}\boldsymbol{\check{h}}_{m}\right) u_{\boldsymbol{w}_{m}.\to\delta_{2}}^{t} \left(\boldsymbol{w}_{m}.\right) d\boldsymbol{\check{h}}_{m},$$
(58)

and then the message passed from the factor node δ_2 to the variable node w_m . reads $u_{\delta_2 \to w_m}^t (w_m) = \beta^t (w_m) / u_{w_m \to \delta_2}^t (w_m) = \mathcal{N}_{\mathbb{C}} (w_m :; \omega_m^{t+1}, \gamma^{t+1} I_P)$, where the variance and mean are shown in line 15 of Table III, which updates the message initialized in (51).

Given the *aposteriori* distribution of $p(h_m|y; \xi^{t-1})$ and $p(\Phi_p w_{\cdot p}|y_p; \xi^{t-1})$, parameters ξ are updated by minimizing the Bethe free energy under a neighborhood constraint, which is similar to (38). From (52), the *aposteriori* distributions of $\Phi_p w_{\cdot p}$ are obtained as follows

$$p\left(\boldsymbol{\Phi}_{p}\boldsymbol{w}_{\cdot p}|\boldsymbol{y}_{p};\boldsymbol{\xi}^{t-1}\right)$$
$$= \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{\Phi}_{p}\boldsymbol{w}_{\cdot p};\boldsymbol{\Phi}_{p}\boldsymbol{\hat{u}}_{\boldsymbol{w}_{\cdot p}}^{t}, v_{\boldsymbol{w}_{\cdot p}}^{t}\frac{1}{KG}\mathsf{Tr}\left(\boldsymbol{\Phi}_{p}\boldsymbol{\Phi}_{p}^{\mathsf{H}}\right)I_{KG}\right).$$
(59)

Hence, the variance of noise is updated in line 18 of Table III, where $|\mathbf{y}_p - \mathbf{\Phi}_p \hat{\mathbf{u}}_{\mathbf{w},p}^t|^2$ refers to the element-wise square of a vector. The sparsity ratio and the *apriori* variance of the channel coefficients are updated similar to line 23 of Table I.

REFERENCES

- A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, 2014.
- [2] O. El Ayach, R. W. Heath, S. Rajagopal, and Z. Pi, "Multimode precoding in millimeter wave MIMO transmitters with multiple antenna sub-arrays," in *IEEE GLOBECOM Commun. Conf.*, 2013, pp. 3476– 3480.
- [3] C. Kim, T. Kim, and J. Y. Seol, "Multi-beam transmission diversity with hybrid beamforming for MIMO-OFDM systems," in *IEEE GLOBECOM Workshops*, 2013, pp. 61–65.
- [4] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heath, "MIMO precoding and combining solutions for millimeter wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, 2014.
- [5] H. Xu, V. Kukshya, and T. S. Rappaport, "Spatial and temporal characteristics of 60-GHz indoor channels," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 3, pp. 620–630, 2002.
- [6] T. S. Rappaport, S. Sun, R. Mayzus, and H. Zhao, "Millimeter wave mobile communications for 5G cellular: It will work!" *Access IEEE*, vol. 1, no. 1, pp. 335–349, 2013.
- [7] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, June 2014.

- [8] K. Ramachandran, N. Prasad, K. Hosoya, K. Maruhashi, and S. Rangarajan, "Adaptive beamforming for 60 GHz radios: challenges and preliminary solutions," in ACM International Workshop on Mmwave Communications: From Circuits To Networks, 2010, pp. 33–38.
- [9] B. Li, Z. Zhou, W. Zou, X. Sun, and G. Du, "On the efficient beamforming training for 60GHz wireless personal area networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 504–515, 2013.
- [10] D. E. Berraki, S. M. D. Armour, and A. R. Nix, "Application of compressive sensing in sparse spatial channel recovery for beamforming in mmwave outdoor systems," in *Wireless Commun. and NETWORKING Conf.*, 2014, pp. 887–892.
- [11] A. Alkhateeby, G. Leusz, and R. W. Heath, "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?" in 2015 IEEE ICASSP, April 2015, pp. 2909–2913.
- [12] R. W. Heath, N. González-Prelcic, S. Rangan, and W. Roh, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, 2016.
- [13] L. Dai, X. Gao, S. Han, I. Chih-Lin, and X. Wang, "Beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array," in 2016 IEEE/CIC International Conference on Communications in China (ICCC), July 2016, pp. 1–6.
- [14] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeterwave massive MIMO with hybrid precoding over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1259–1262, 2016.
- [15] J. Mo, P. Schniter, N. G. Prelcic, and R. W. Heath, "Channel estimation in millimeter wave MIMO systems with one-bit quantization," in 2014 48th Asilomar Conference on Signals, Systems and Computers, Nov 2014, pp. 957–961.
- [16] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2008.
- [17] J. Yang, C. Wen, S. Jin, and F. Gao, "Beamspace channel estimation in mmwave systems via cosparse image reconstruction technique," arXiv:1707.09097, Jul 2017.
- [18] Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli, "Estimation of sparse MIMO channels with common support," *IEEE Trans. Commun.*, vol. 60, no. 12, pp. 3705–3716, December 2012.
- [19] M. K. Samimi and T. S. Rappaport, "Ultra-wideband statistical channel model for non line of sight millimeter-wave urban channels," in 2014 IEEE Global Communications Conference, Dec 2014, pp. 3483–3489.
- [20] J. Mo, P. Schniter, and R. W. Heath, "Channel estimation in broadband millimeter wave MIMO systems with few-bit ADCs," arXiv:1610.02735, Apr. 2017.
- [21] X. Lin, S. Wu, L. Kuang, Z. Ni, X. Meng, and C. Jiang, "Estimation of sparse massive MIMO-OFDM channels with approximately common support," *IEEE Communications Letters*, vol. 21, no. 5, pp. 1179–1182, May 2017.
- [22] S. Rangan, P. Schniter, and A. Fletcher, "Vector approximate message passing," arXiv:1610.03082, Oct 2016.
- [23] S. S. Chen and M. A. Saunders, "Atomic decomposition by basis pursuit," in *Siam J. Sci. Comput*, 2001, pp. 33–61.
- [24] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.
- [25] Z. Gao, L. Dai, Z. Wang, and S. Chen, "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.
- [26] J. Vila and P. Schniter, "Expectation-maximization Bernoulli-Gaussian approximate message passing," in 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Nov 2011, pp. 799–803.
- [27] 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, March 2014.
- [28] Y. Zhou, M. Herdin, A. M. Sayeed, and E. Bonek, "Experimental study of MIMO channel statistics and capacity via virtual channel representation," Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep, 2007.
- [29] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1058–1076, June 2010.
- [30] W. U. Bajwa, "New information processing theory and methods for exploiting sparsity in wireless systems," Ph.D. dissertation, Univ. Wisconsin-Madison, Madison, 2009.

- [31] F. Krzakala, M. Mézard, Y. S. F. Sausset, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," *Journal of Statistical Mechanics Theory and Experiment*, vol. 8, no. 4, p. P08009, 2012.
- [32] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: II. analysis and validation," in 2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo), Jan 2010, pp. 1–5.
- [33] C. K. Wen, S. Jin, K. K. Wong, J. C. Chen, and P. Ting, "Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1356–1368, March 2015.
- [34] S. Wu, L. Kuang, Z. Ni, D. Huang, Q. Guo, and J. Lu, "Messagepassing receiver for joint channel estimation and decoding in 3D massive MIMO-OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8122–8138, Dec 2016.
- [35] S. Wu, L. Kuang, Z. Ni, J. Lu, D. D. Huang, and Q. Guo, "Expectation propagation approach to joint channel estimation and decoding for OFDM systems," in 2014 IEEE ICASSP, May 2014, pp. 1941–1945.
- [36] H. Wymeersch, Iterative receiver design. Cambridge University Press, 2007.
- [37] X. Meng, S. Wu, L. Kuang, and J. Lu, "Concise derivation of complex bayesian approximate message passing via expectation propagation," arXiv:1509.08658V2, Jan. 2016.
- [38] J. Ma, X. Yuan, and L. Ping, "Turbo compressed sensing with partial DFT sensing matrix," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 158–161, Feb 2015.



Linling Kuang (S'01-M'06) received the B.S. and M.S. degrees from the National University of Defense Technology, Changsha, China, in 1995 and 1998, respectively, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2004. Since 2007, she has been with the Tsinghua Space Center, Tsinghua University. Her research interests include wireless broadband communications, signal processing, and satellite communication. She is a member of the IEEE Communications Society.



Jian Yan received the B.S., M.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 1998, 2000, and 2010, respectively. From 2001 to 2011, he was with the Department of Electronic Engineering, Tsinghua University. Since August 2011, he has been working at the Tsinghua Space Center, Tsinghua University, as an associate researcher. His research interests lie in the area of satellite communications.



Xincong Lin received the B.S. degree in electronic and information engineering from Beijing Jiaotong University, Beijing, China, in 2015. He is currently working towards Ph.D. degree in the Department of Aerospace Engineering at Tsinghua University, Beijing, China. His research interests include sparse signal processing, iterative detection, massive MIMO and millimeter-wave communications.



Sheng Wu (S'13-M'14) received the B.E. and M.E. degrees from Beijing University of Post and Telecommunications, Beijing, China, in 2004 and 2007, respectively, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2014. Currently, he is a postdoctoral researcher in the Tsinghua Space Center at Tsinghua University, Beijing, China. His research interests are mainly in iterative detection and decoding, channel estimation, massive MIMO, and satellite communications.

Chunxiao Jiang (S'09-M'13-SM'15) received the B.S. in information engineering from Beihang University in Jun. 2008 and the Ph.D. in electronic engineering from Tsinghua University in Jan. 2013, both with the highest honors. From Feb. 2013 - Jun. 2016, Dr. Jiang was a Postdoc in the Department of Electronic Engineering Tsinghua University, during which he visited University of Maryland College Park and University of Southampton. He is currently an assistant professor in Tsinghua Space Center, Tsinghua University. He is a recipient of the IEEE

Globecom Best Paper Award in 2013, the IEEE GlobalSIP Best Student Paper Award in 2015, the IEEE IWCMC Best Paper Award in 2017, and the IEEE Communications Society Young Author Best Paper Award in 2017. Since 2015, Dr. Jiang became a IEEE Senior Member.



Lajos Hanzo (http://www-mobile.ecs.soton.ac.uk) FREng, FIEEE, FIET, Fellow of EURASIP, DSc received his degree in electronics in 1976 and his doctorate in 1983. In 2009 he was awarded an honorary doctorate by the Technical University of Budapest and in 2015 by the University of Edinburgh. In 2016 he was admitted to the Hungarian Academy of Science. During his 40-year career in telecommunications he has held various research and academic posts in Hungary, Germany and the UK. Since 1986 he has been with the School of

Electronics and Computer Science, University of Southampton, UK, where he holds the chair in telecommunications. He has successfully supervised 111 PhD students, co-authored 18 John Wiley/IEEE Press books on mobile radio communications totalling in excess of 10 000 pages, published 1700+ research contributions at IEEE Xplore, acted both as TPC and General Chair of IEEE conferences, presented keynote lectures and has been awarded a number of distinctions. Currently he is directing a 60-strong academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC) UK, the European Research Council's Advanced Fellow Grant and the Royal Society's Wolfson Research Merit Award. He is an enthusiastic supporter of industrial and academic liaison and he offers a range of industrial courses. He is also a Governor of the IEEE ComSoc and VTS. During 2008 - 2012 he was the Editor-in-Chief of the IEEE Press and a Chaired Professor also at Tsinghua University, Beijing. For further information on research in progress and associated publications please refer to http://www-mobile.ecs.soton.ac.uk Lajos has 34 000+ citations.