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 multiple-input 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 results in extremely high-energy 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 ∼ 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 (e-mail: lh@ecs.soton.ac.uk).
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.

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-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 delay-domain 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.

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
Frobenius norm of matrix $A$. $\|a\|$ is the $\ell_2$ norm of vector $a$. \( [x] \) is the largest integer $n$ that $n \leq x$, \([x]\) is the smallest integer $n$ that $n \geq x$. $\Phi(g,m)$ is the $g$-th row and $m$-th column element of the matrix $\Phi$. $\ln(\cdot)$ and $\lg(\cdot)$ denotes the natural logarithm and logarithm to ten, respectively. $\text{vec}(A)$ denotes vectorizing a matrix as a vector. $\otimes$ denotes the Kronecker product. Finally, $\mathcal{N}(h; \mu, \nu)$ denotes the Gaussian distribution function of $h$ with mean $\mu$ and variance $\nu$. 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 $N$. $\text{Tr}(A)$ is the trace of a matrix. $\langle a \rangle$ is the empirical averaging of a vector $a \in \mathbb{R}^{N \times 1}$, i.e., $\langle a \rangle = \frac{1}{N} \sum a_i$.

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 [11], [12], [27]. The base station (BS) having $N_{BS}$ antennas and $N_{RF}^{BS}$ = $K$ RF chains serves $K$ user equipment (UE) having $N_{UE}$ antennas and $N_{RF}^{UE}$ = $1$ 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 $\mathcal{P} = \{1, 2, \ldots, P\}$, where we have $\Delta = [N/P]$. Furthermore, $\{k_{p}, \ell_{p} \mid k = 1, \ldots, K, g = 1, \ldots, G, p \in \mathcal{P}\}$ 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 $f_{BBk_p}$, followed by an RF precoder $f_{RFk_p} \in \mathbb{C}^{N_{UE} \times 1}$, the transmitted signal $x_{k_p} \in \mathbb{C}^{N_{UE} \times 1}$ can be written as

$$x_{k_p} = f_{BBk_p} f_{RFk_p} \in \mathbb{C}^{N_{UE} \times 1}$$

where $f_{k_p} = f_{RFk_p} f_{BBk_p} \in \mathbb{C}^{N_{UE} \times 1}$ 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 $L_{n_\alpha}$ 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]

$$H(f) = \sum_{i=1}^{L_{\text{path}}} \beta_i a_R(\theta_{Ri}) a_T^H(\theta_{Ti}) e^{-j2\pi f \tau_{i}},$$

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

$$a_R(\theta) = \begin{bmatrix} 1, e^{-j2\pi \theta}, \ldots, e^{-j2\pi (N_{BS}-1)\theta} \end{bmatrix}^T$$

$$a_T(\theta) = \begin{bmatrix} 1, e^{-j2\pi \theta}, \ldots, e^{-j2\pi (N_{UE}-1)\theta} \end{bmatrix}^T$$

denotes the receive and transmit steering vector with $\theta$ 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 $\phi_{Ti} \in (-\pi/2, \pi/2)$ by $\theta_{Ti} = \frac{1}{2} \sin(\phi_{Ti})$, 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]

$$H(f) = A_R W(f) A_T^H,$$

where

$$A_R = [\alpha_R(\theta_{R1}), \alpha_R(\theta_{R2}), \ldots, \alpha_R(\theta_{Rn_{BS}})] / \sqrt{N_{BS}},$$

$$A_T = [\alpha_T(\theta_{T1}), \alpha_T(\theta_{T2}), \ldots, \alpha_T(\theta_{Tn_{UE}})] / \sqrt{N_{UE}},$$

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

$$\beta_{R}^{nR} = \frac{1}{N_{BS}} [n_{R} - \bar{N}_{BS}], \quad \beta_{T}^{nT} = \frac{1}{N_{UE}} [n_{T} - \bar{N}_{UE}],$$

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

$$W_p = W(f_p) = \frac{1}{\sqrt{L}} \sum_{l=1}^{L} H_l e^{-j2\pi \theta (l-1)(\theta_p/\Delta)},$$

where $f_p = -B/2 \leq f_p \leq B/2$ is $p$-th subcarrier frequency, and $L = L + 1 + L$ with $L = [B_{\text{max}}]$, where $B$ and $B_{\text{max}}$ is the bandwidth and maximum path delay, respectively. The elements in $H_i$ is [28]–[30]

$$H_{n_{R} n_{T} l} = \sqrt{L} \sum_{i=1}^{L_{\text{path}}} \beta_i a_{n_{R}}(\theta_{n_{R} - \theta_{R}^{nR}}) a_{n_{T}}^H(\theta_{n_{T} - \theta_{T}^{nT}}) \times \text{sinc}(B_{\text{max}} (l - 1)),$$

where $f_N(\theta) = \frac{1}{\sqrt{N}} e^{-j2\pi \theta (N-1)} \sum_{n=0}^{N-1} e^{-j2\pi \theta (n/N)}$ and $\text{sinc}(x) = \frac{\sin(x)}{x}$.

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 $N_c(\beta_i; 0, 10 \text{power}/10)$, with the path-power being 0 dBm for the LOS component and $-5 \sim -20$ dBm for the NLOS component. The maximum path delay is about $B_{\text{max}} = 600$ ns [6], therefore, the path delays $\tau_i$ are uniformly selected from 0 $600$ ns. Other parameters are set as $f_0 = 28$ GHz, $B = 800$ MHz [6], $L = [B_{\text{max}}] = 480$, $N_{BS} = 64$, $N_{UE} = 16$, $[\phi_{R1}, \phi_{R2}, \phi_{R3}, \phi_{R4}] = [70, 20, -20, -70]/180\pi$ and $[\phi_{T1}, \phi_{T2}, \phi_{T3}, \phi_{T4}] = [60, 30, -30, -60]/180\pi$. With these parameters, the channel $H_i$ can be generated from (10), and the channel $W_p$ can be generated from (9). The normalized amplitude of $H_i$ 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 $[H_1 e_1, \cdots, H_{\text{BS}} e_1] \in \mathbb{C}^{64 \times 481}$ with $e_1 \in \mathbb{R}^{1 \times 61}$. Generally speaking, due to the effect of power leakage, angular spread and cluster duration, the channels $[H_1, \cdots, H_L]$ 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 $r_{sp} \in \mathbb{C}^{N_{\text{BS}}}^2$ at the $p$th subcarrier of the $g$th OFDM symbol from multiple users formulated as

$$r_{sp} = \sum_{k=1}^{K} A_R W_{kp} A_T^* x_{kp} + w_{sp}, \quad (11)$$

where $w_{sp} \sim \mathcal{N}_C\left(0, \sigma^2 \right)$ is the additive noise. The SNR is defined as $\text{SNR}=\frac{1}{\sigma^2}$ or under the condition that $\mathbb{E} \left[|s_{kp}|^2\right] = 1$, $\mathbb{E} \left[|W_{kp}|^2\right] = N_{\text{BS}}^2 \text{NUE}$ [27]. Then the received signal $r_{sp}$ is further processed by the combined matrix $Z_{sp} = Z_{RFg} Z_{BBSP} \in \mathbb{C}^{N_{\text{BS}} \times K^2}$ as follows

$$y_{sp} = Z_{sp}^T r_{sp} = Z_{sp}^T \sum_{k=1}^{K} A_R W_{kp} A_T^* x_{kp} + Z_{sp}^T w_{sp} \quad (12).$$

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

$$y_{sp} = Z_{sp}^T A_R \tilde{W}_p A_T^* \tilde{f}_{sp} + n_{sp}, \quad (13)$$

where $\tilde{W}_p = \left[ W_{1p}, W_{2p}, \cdots, W_{Kp} \right] \in \mathbb{C}^{N_{\text{BS}} \times K_{\text{NUE}}}$, $\tilde{A}_T = \text{diag} \left[ A_T, A_T, \cdots, A_T \right] \in \mathbb{C}^{K_{\text{NUE}} \times K_{\text{NUE}}}$, $\tilde{f}_{sp} = \left[ f_{1sp}^T, f_{2sp}^T, \cdots, f_{Ksp}^T \right] \in \mathbb{C}^{K_{\text{NUE}} \times 1}$ and $n_{sp} = Z_{sp}^T w_{sp} \in \mathbb{C}^{K \times 1}$. By vectorizing $\tilde{W}_p$, we have [1], [14]

$$y_{sp} = \text{vec} \left( Z_{sp}^T A_R \tilde{W}_p A_T^* \tilde{f}_{sp} \right) + n_{sp} = \left( A_T^* \tilde{f}_{sp} \right) \text{vec} \left( \tilde{W}_p \right) + n_{sp} \quad (14).$$

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

$$y_p = \Phi_p w_p + n_p, \quad (15)$$

where we have $y_p = \left[ y_{1p}^T, y_{2p}^T, \cdots, y_{Gp}^T \right] \in \mathbb{C}^{K_{G} \times 1}$, $\Phi_p = \left[ \Phi_{1p}^T, \Phi_{2p}^T, \cdots, \Phi_{Gp}^T \right] \in \mathbb{C}^{K_{G} \times K_{\text{NUE}}}$, and $n_p = \left[ n_{1p}^T, n_{2p}^T, \cdots, n_{Gp}^T \right] \in \mathbb{C}^{K_{G} \times 1}$.

Denote

$$H = \left[ \text{vec} \left( \tilde{H}_1 \right), \cdots, \text{vec} \left( \tilde{H}_L \right) \right] \in \mathbb{C}^{M \times L}, \quad (16)$$

$$W = [w_1, \cdots, w_p] \in \mathbb{C}^{M \times P}, \quad (17)$$

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

$$W = H \phi^T, \quad (18)$$

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

$$y_p = \text{vec} \left( \Phi_p H \phi^T e_p \right) + n_p = \left( \phi^T e_p \right) \text{vec}(H) + n_p$$

$$= \Psi_p h + n_p, \quad (19)$$

where $\Psi_p \in \mathbb{C}^{K_{G} \times K_{\text{NUE}}}$, $h \in \mathbb{C}^{K_{\text{NUE}}}$, and $e_p \in \mathbb{C}^{P \times 1}$. By stacking $P$ subchannel received signals $y_p$, we get the system model in the virtual AOA-AOD-delay domain as

$$y = \Psi h + n, \quad (20)$$

where $y = \left[ y_1^T, y_2^T, \cdots, y_p^T \right]^T \in \mathbb{C}^{K_{G} \times 1}$, $\Psi = \left[ \Psi_1^T, \Psi_2^T, \cdots, \Psi_p^T \right]^T \in \mathbb{C}^{K_{G} \times K_{\text{NUE}}}$, and $n = \left[ n_1^T, n_2^T, \cdots, n_p^T \right]^T \in \mathbb{C}^{K_{G} \times 1}$.

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 $\Phi_p \in \mathbb{C}^{K_{G} \times K_{\text{NUE}}}$, i.e., the scale of $O \left( K_{G}^2 P \right)$ [21]. However, directly applying the AMP-NNSPL to the model (20) would lead to high complexity of $O \left( K_{G}^2 P N_{\text{BS}}^2 \right)$, as the term $L$ is usually huge in broadband mmWave systems. Given the model of (15) and (18), we extend our low-complexity 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

$$p \left( H \mid y \right) = \frac{p \left( y \mid H \right) p \left( H \right)}{p \left( y \right)}, \quad (21)$$

where $p \left( y \right) = \int \int p \left( W \mid H \right) p \left( H \right) dW dH$. The numerator in (21) can be factored into

$$p \left( H \right) p \left( W \mid H \right) p \left( y \mid W \right) = \prod_{m=1}^{M} p \left( h_m \mid h_m \right) \times \prod_{p \in P} f_p \left( y_p \mid w_p \right). \quad (22)$$

where $h_m \in \mathbb{C}^{L \times 1}$ is the $m$th row of channels $H \in \mathbb{C}^{M \times L}$, $w_m \in \mathbb{C}^{P \times 1}$ and $w_p \in \mathbb{C}^{P \times 1}$ 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 $H \in \mathbb{C}^{M \times L}$.

$$p \left( H ; \xi \right) = \prod_{m=1}^{M} p \left( h_m \right)$$
message passing in the right part is efficiently implemented by the Gaussian message passing (GMP) algorithm proposed in [34], [35]. Passing messages from the left part to the right part is achieved by passing messages from the variable node $w_{mp}$ to the factor node $g_{mp}$, and vice versa. Again, we term the proposed message-passing scheme as AMP-NNPL-DD algorithm. Additionally, the SE is derived as part of our performance analysis.

1) The AMP-NNPL-DD Algorithm: The channels $\mathbf{H}$, $\mathbf{W}$ and the parameters $\xi$ can be iteratively estimated by message passing and by minimizing the Bethe free energy under a neighborhood constraint [21]. By fixing the parameters $\xi$ to the values estimated at the previous iteration, the terms $\zeta^t_{g_{mp}}$, $\bar{\zeta}^t_{g_{mp}}$ and $\omega^t_{mp}$ 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

$$
\xi^t_{mp} = \sum_{\omega_{mp}} \left( \phi_{pl} \right)^{-1} \bar{\zeta}^t_{w_{mp} \rightarrow h_{ml}},
$$

(28)

$$
\zeta^t_{g_{mp}} = \sum_{l=1}^{\bar{L}} v^t_{h_{ml} \rightarrow g_{ml}} \bar{v}^t_{m_{mp} \rightarrow h_{ml}},
$$

(29)

where $\bar{\zeta}^t_{g_{mp}}$ and $v^t_{h_{ml} \rightarrow g_{ml}}$ are the variance defined at the function node $g_{mp}$ and the variable node $h_{ml}$, respectively, and

$$
\hat{\nu}^t_{w_{mp} \rightarrow g_{mp}} = \hat{\nu}^t_{w_{mp} \rightarrow g_{mp}} - \sum_{l \neq 1}^{\bar{L}} \phi_{pl} \hat{h}_{ml}^{-1} - g_{mp},
$$

(30)

with $\hat{\nu}^t_{w_{mp} \rightarrow g_{mp}}$ being the mean of the message $u^t_{w_{mp} \rightarrow g_{mp}}$ passed from the variable node $w_{mp}$ to the factor node $g_{mp}$, and $\hat{\nu}^t_{h_{ml} \rightarrow g_{mp}}$ being the mean of the message passed from the variable node $h_{ml}$ to the factor node $g_{mp}$. $\xi^t_{mp}$ and $\zeta^t_{g_{mp}}$ 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 $u^t_{w_{mp} \rightarrow g_{mp}}$ is calculated by

$$
u^t_{w_{mp} \rightarrow g_{mp}} (w_{mp}) = \frac{\nu^t_{w_{mp} \rightarrow g_{mp}} (w_{mp})}{\mathbb{N} (w_{mp} ; \nu^t_{mp} ; \mu^t_{mp} ; \zeta^t_{g_{mp}})},
$$

(31)

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

$$
\hat{\nu}^t_{w_{mp} \rightarrow g_{mp}} = \omega^t_{mp}.
$$

(32)

Then the variance $\omega^t_{mp}$ and mean $\hat{\nu}^t_{g_{mp}}$ defined at the factor node $g_{mp}$, and the variance $\epsilon^t_m$ and mean $\mu^t_m$ 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

$$
p (h_{ml} | y, \xi^t) = \frac{1}{\int p (h_{ml}) \mathbb{N} (h_{ml} ; \mu^t_m ; \epsilon^t_m) dh_{ml} \times p (h_{ml}) \mathbb{N} (h_{ml} ; \mu^t_m ; \epsilon^t_m)}
$$

Figure 2: The factor graph representation of broadband mmWave massive MIMO-OFDM Systems.
= (1 - \pi^t_{m_l}) \delta(h_{m_l}) + \pi^t_{m_l} N_{C}(h_{m_l}; \tilde{h}^t_{m_l}, \nu^t_{m_l}), \tag{33}

where \pi^t_{m_l}, \tilde{h}^t_{m_l}, and \nu^t_{m_l} are shown in lines 13 and 14 of table I. The aposteriori mean and variance of \( h_{m_l} \) are calculated as

\[
g_{\text{mean}}(\mu^t_{m_l}, \epsilon^t_{m_l}; \xi^{-1}) = \int h_{m_l} p_{w}(h_{m_l} \mid y; \xi^{-1}) dh_{m_l}, \tag{34}
\]

\[
g_{\text{var}}(\mu^t_{m_l}, \epsilon^t_{m_l}; \xi^{-1}) = \int |h_{m_l}|^2 p_{w}(h_{m_l} \mid y; \xi^{-1}) dh_{m_l}
- |g_{\text{mean}}(\mu^t_{m_l}, \epsilon^t_{m_l}; \xi^{-1})|^2, \tag{35}
\]

and they are shown in line 15 of table I. According to the sum-product message passing rule, the aposteriori distributions of \( \nu^t_{w_{m_p}} \) are obtained by

\[
p(\nu^t_{w_{m_p}} \mid y_p) = \frac{u^t_{\nu^t_{w_{m_p}} \rightarrow y_p}(w_{m_p}) u^t_{\nu^t_{w_{m_p}} \rightarrow \nu^t_{w_{m_p}}}(w_{m_p})}{N_{C}(w_{m_p}; \bar{w}^t_{\nu^t_{w_{m_p}} \rightarrow y_p}, \nu^t_{\nu^t_{w_{m_p}} \rightarrow \nu^t_{w_{m_p}}}) N_{C}(w_{m_p}; \bar{w}^t_{\nu^t_{w_{m_p}} \rightarrow y_p}, \nu^t_{\nu^t_{w_{m_p}} \rightarrow \nu^t_{w_{m_p}}})}
\approx N_{C}(w_{m_p}; \bar{w}^t_{\nu^t_{w_{m_p}}}, \nu^t_{\nu^t_{w_{m_p}}}), \tag{36}\]

where \( \nu^t_{\nu^t_{w_{m_p}} \rightarrow \nu^t_{w_{m_p}}}, \nu^t_{\nu^t_{w_{m_p}} \rightarrow \nu^t_{w_{m_p}}}, \) and \( \nu^t_{\nu^t_{w_{m_p}} \rightarrow \nu^t_{w_{m_p}}} \) are calculated by GMP as shown in line 18 of table I, \( \nu^t_{w_{m_p}} \) and \( \nu^t_{w_{m_p}} \) 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 \( m = (k-1)N_{BS}N_{UE} + (i-1)N_{BS} + j \), where \( m = 1, \ldots, M, k = 1, \ldots, K, i = 1, \ldots, N_{BS} \) and \( j = 1, \ldots, N_{UE} \). Then we denote \( h_{ijl} \equiv h_{m_l} \), 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 \( h_{ijl} \) tend to be either simultaneously small value or large value, where \( N_{ijl} \) denotes the set of neighbor indices of element \( h_{ijl} \), and is defined as follows\(^1\)

\[
N_{ijl} = [(i-1, j, l), (i+1, j, l), (i, j - 1, l), (i, j + 1, l), (i, j, l - 1), (i, j, l + 1)]. \tag{37}
\]

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

\[
\xi^t = \arg \min_{\xi} Q(\xi), \tag{38}
\]

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

\[
\frac{\sum_{p \in \mathcal{P}} \sum_{k = 1}^{\frac{K}{G}} p(\epsilon_{p} \Phi_{p} w_{p} | y_{p}; \xi^{-1}) \ln p(y_{p} | w_{p}; \xi)}{\int_{H} p(H | y; \xi^{-1}) \ln p(H; \xi) + \text{Const}}. \tag{40}\]

The proposed algorithm is summarized in table I. The initialization of the parameters \( \xi \) is consistent with \([21]\) and are shown in line 2 of table I. Note that \( \xi_{\nu^t_{w_{m_p}} \rightarrow y_p}, \nu^t_{\nu^t_{w_{m_p}} \rightarrow \nu^t_{w_{m_p}}}, \) and \( \nu^t_{\nu^t_{w_{m_p}} \rightarrow \nu^t_{w_{m_p}}} \) 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 \( O(K^2GPN_{BS}N_{UE} + KN_{BS}N_{UE} \log_2 P) \), 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 \( O(K^2GPN_{BS}N_{UE}) \).

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 \( \nu_{w_{m_p}} \) and \( H \) are defined as

\[
\nu_{w} = \frac{1}{MP} \sum_{m=1}^{M} \sum_{p \in \mathcal{P}} (\nu^t_{w_{m_p}} - \nu_{w_{m_p}})^2, \nu_{h} = \frac{1}{ML} \sum_{m=1}^{M} \sum_{l=1}^{L} (h_{m_l} - \nu_{h_{m_l}})^2, \tag{41}\]

\[
\nu^t_{w_{m_p}} = \frac{1}{MP} \sum_{m=1}^{M} \sum_{p \in \mathcal{P}} (\nu^t_{w_{m_p}} - \nu^t_{w_{m_p}})^2, \nu^t_{h_{m_l}} = \frac{1}{ML} \sum_{m=1}^{M} \sum_{l=1}^{L} (h_{m_l} - \nu^t_{h_{m_l}})^2, \tag{42}\]

respectively. It is shown in \([31]\) that in the large \( M \) limit and when the elements of the measuring matrix \( \Phi_{p} \) are drawn from \( N_{C}(x; 0, 1/M) \), \( \theta_{m_p} \) and \( \omega_{m_p} \) can be expressed as

\[
\theta^t = \sigma^{-1} + \nu^{-1}, \omega_{m_p} = \nu_{w_{m_p}} + \sqrt{\frac{\sigma_0 + \nu^{-1}_{w_{m_p}}}{KG/M}}, \tag{43}\]

\[
\nu^t_{w_{m_p}} = \nu^t_{h_{m_l}} = \frac{v_{0}}{L} \left( \theta^t + \nu_{h} \right), \tag{44}\]

\(^1\)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 rectangles.
From (30), (32), (43) and (18), we have

\[
\eta_m = \frac{1}{p_m} \sum \phi_p h_{mp} - \frac{\phi_p h_{mp}}{\|\Phi_p\|_F},
\]

(46), we have

\[
e_m = \frac{1}{v} \sum \phi_p h_{mp} + \sum \phi_p \left( \frac{1}{v_{mp}} \sum \phi_p \right)
\]

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

\[
e_m = \frac{1}{v} \sum \phi_p h_{mp} + \sum \phi_p \left( \frac{1}{v_{mp}} \sum \phi_p \right)
\]

where step \((\approx)\) is by the assumption that \(\phi_p h_{mp}\) are drawn from \(N_C(\phi_p h_{mp}; 0, 1/L)\), then in the large \(L\) limit, the third term of (45) admits a Gaussian random with zero mean and variance \(e_{h}^{-1}\) according to the central limit theorem, and the variable \(r\) in (46) admits the distribution \(N_C(r; 0, 1)\). From (28) and
Table II: The State Evolution of AMP-NNSPL-DD algorithm.

1: Initialization: $\forall m,l: \rho_{ml}^0 = 0.5, \sigma_{ml}^0 = 1, \eta_{ml}^0 = 1$.
2: for $t = 1, \cdots T$
3:    // AMP: passing message $w_{mp} \rightarrow f_{kp} \rightarrow w_{mp}$
4:    $\theta_t = \frac{\sigma_{t-1}^{p} + \epsilon_t^{c - 1}}{KG/M}$,
5:    // GMP: passing message $w_{mp} \rightarrow g_{mp} \rightarrow h_{ml}$
6:    $\nu_t = L (\theta_t + \nu_t^{-1})$,
7:  $\forall m,l: e_{ml}^t = \sum_{\rho \in \rho} (\phi_{ml})^{-1} \left( \frac{\sigma_{m}^{p} + \epsilon_t^{c - 1} \nu_t^{-1}}{KG/M} + \frac{\epsilon_t^{c - 1} \nu_t^{-1}}{KG/M} \right)$
8:    $\forall l: \epsilon_t^l = \frac{\nu_t}{\rho_t}$.
9: // The aposteriori: passing message $h_{ml} \rightarrow p (h_{ml}) \rightarrow h_{ml}$
10: $\forall l, m, \hat{h}_{ml}^t = \ln \frac{e_{ml}^t}{\nu_t \epsilon_{ml}^t} - \frac{\nu_t \epsilon_{ml}^t}{\nu_t}$
11: $\hat{h}_{ml}^t = \sigma_{ml}^t \epsilon_{ml}^t$
12: $\forall m, l, v_{hml}^t = \pi_{ml}^t \left( (\hat{h}_{ml}^t)^2 (1 - \pi_{ml}^t) + \hat{v}_{ml}^t \right)$
13: $v_{hml}^t = \int D h_{ml} \int D z \int D r \hat{h}_{ml}^t$
14:  // GMP: passing message $h_{ml} \rightarrow g_{mp} \rightarrow w_{mp}$
15:  $\forall m, p: \hat{v}_{mp}^t = v_{hml}^t w_{mp} + L \left( \frac{\sigma_{mp}^p + \epsilon_t^{c - 1} \nu_t^{-1}}{KG/M} + \frac{\epsilon_t^{c - 1} \nu_t^{-1}}{KG/M} \right)$
16:  $\forall m, p: v_{mp}^t = \frac{1}{\nu_t \epsilon_{mp}^t}$
17:  $v_{wp}^t = \int D w_{mp} \int D z \int D r v_{hml}^t$.
18: //Simultaneously update $\rho_{ml}$ and $\eta_{ml}$ as
19:  Denote: $m \equiv (k-1)N_{BS}N_{UE} + (l-1)N_{BS} + j$.
20: $\forall k, l, j: \pi_{kij}^t = \pi_{ml}^t$
21: $\forall k, l, j: \pi_{kij}^t = \frac{1}{N_{BS}N_{UE}} \sum_{q, r} \pi_{kqrs}^t \pi_{lqr}s$
22: $\forall k, l, j: \hat{v}_{kij}^t = \int D z \int D r \pi_{kij}^t (z, r)$,
23: $\forall m, l: \rho_{ml}^t = \rho_{ml}^t$
24: Update $\sigma$ as
25: $\sigma = \frac{\sigma_0 + \epsilon_t^{c - 1}}{(1 + \nu_t \epsilon_{ml}^t)} + \frac{\sigma_0 + \epsilon_t^{c - 1} \nu_t^{-1}}{\sigma_0 + \epsilon_t^{c - 1} \nu_t^{-1} + \nu_t}$
26: end
27: Output: $\epsilon_{h_m}^t$.

$Dh_{ml}$ are implemented by Monte Carlo method with $w_{mp}$ and $h_{ml}$ generated from (18) and (10), respectively. The parameters $\rho_{ij}(z, r)$ and $\eta_{ml}(z, r)$ are calculated by the AMP-NNSPL-DD algorithm as shown in line 22 of Table II. Then $\rho_{ij}$ and $\eta_{ml}$ 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 $\delta_{1} (-) = \delta_{2} (-) = \delta (-)$ is the Dirac delta function and the node with “+” represents $w_{p} = \left[ w_{1p}, w_{2p}, \cdots, w_{MP} \right]^{T}$ or $w_{m} = \left[ w_{m1}, w_{m2}, \cdots, w_{mP} \right]^{T}$. The channel $W \in \mathbb{C}^{M \times P}$ in the virtual AO-AD-frequency domain is represented by a matrix; the factor of $\prod_{p \in P} f_{p}(x_{p} |w_{p})$ represents $P$ independent measurements along the column of the matrix $W \in \mathbb{C}^{M \times P}$, and the factor of $\prod_{m=1}^{M} p(w_{m} | h_{m})$ describes the $M$ relationships among the channels in the delay-domain and frequency-domain 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 $N_{\mathbb{C}}(w_{p}; \omega_{l}^{p}, \gamma_{l}^{p} I_{MP})$ (input information), and measurements of $y_{p} = \Phi_{w}w_{p} + n_{p}$ with $n_{p} \sim N_{\mathbb{C}}(n_{p}; 0, \sigma_{l_{KG}}^{-1} I_{KG})$,
Table III: The VAMP-NNSPL-DD algorithm.

1: Input: \( y_p, \Phi_p \).

2: Initialization: \( \forall m, l: \rho_{ml}^0 = 0.5, \sigma_0^0 = \frac{\sum_{p \in \mathcal{P}} \| y_p \|^2}{\| K \|}, \eta_{ml}^0 = \frac{1}{\rho_{ml}^0} \sum_{p \in \mathcal{P}} \frac{(\| y_p \|^2 - KG \sigma_0^0)}{\| \Phi_p \|^2} \).

3: \( \forall m: \omega_{ml} = 0, \eta_{ml} = 0 \).

4: for \( t = \frac{1}{T} \ldots T \), \( T \) is the number of iterations:

5: \( / \text{LMMSE: transform message} \ w_p \rightarrow f_p \rightarrow w_p \): 

6: \( \forall p: v_{w, p}^t = \frac{1}{\mathcal{M}} \text{Tr} \left( y_p' \omega_{ml}^t y_p', y_p, \Phi_p, \sigma_{l+1}^{-1} \right) \).

7: \( \forall p: \zeta_{ml}^t = \left( \frac{1}{\sqrt{\rho_{ml}^t}} - 1 \right)^{-1} \).

8: \( / \text{LMMSE: transform message} \ w_m \rightarrow \delta_m \rightarrow \hat{h}_m \rightarrow h_m \): 

9: \( v_{h, m}^t = \frac{1}{\mathcal{M}} \text{Tr} \left( \kappa_1^t \kappa_2^t \kappa_3^t \kappa_4^t \right) \).

10: \( \forall m: \hat{u}_{h, m}^t = \frac{\hat{u}_{h, m}^t}{\sqrt{\rho_{ml}^t} - s_m^t / \kappa_1^t} \).

11: \( \text{The aposteriori: transform message} \ h_m \rightarrow p (h_m) \rightarrow h_m \rightarrow \delta_m \): 

12: \( v_{h, m}^t > g_{\text{mean}} \left( s_m^t, \kappa_2^t \right) \).

13: \( \kappa_3^t = \left( \frac{1}{\sqrt{\rho_{ml}^t}} - 1 \right)^{-1} \).

14: \( / \text{transform message} \ h_m \rightarrow w_m \): 

15: \( \gamma_{l+1}^t = \kappa_3^t \).

16: Updating \( \rho_{ml}^t \) and \( \eta_{ml}^t \) similar to lines 21-24 of Table I, respectively.

17: /Updating \( \sigma \) as

18: \( \sigma_{l+1}^t = \frac{1}{\mathcal{P}} \sum_{p} \left( \left( y_p - \Phi_p \hat{u}_{w, p}^t \right)^2 \right) + \frac{1}{K_G} \text{Tr} \left( \Phi_p \Phi_p^t \right) \).

19: end

20: Output: \( \forall m: \hat{u}_{h, m}^t \).

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

The refined information \( \mathcal{N}_C \left( w_p, \hat{u}_{w, p}, v_{w, p}, l_m \right) \) (output information) is obtained for \( w_p \) by LMMSE estimation. Next, the extrinsic information \(^2\) of \( w_p \) is given by \( \mathcal{N}_C \left( w_p, r_{p}^t, \zeta_{p}^t l_m \right) \) (line 7), which is fed into the next module as the apriori information. The first module (lines 6-7) corresponds to passing 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-AOD-delay domain to the virtual AOA-AOD-frequency domain under the constraint \( \delta (w_m - \Phi \hat{h}_m) \), which updates the apriori information in the first module (lines 6-7). Lines 16-18 update the parameter \( \xi \) by minimizing the Bethe free energy under neighborhood constraints, which is similar to (38). Line 9 implies that passing the message \( \mathcal{N}_C \left( w_m: r_{m}^t, \frac{1}{\mathcal{P}} \sum_{p} \zeta_{p}^t l_p \right) \) of \( w_m \) to \( \hat{h}_m \) corresponds to applying the LMMSE estimation on the following model

\[
\gamma_{l+1}^t = \frac{1}{\mathcal{P}} \sum_{p} \left( \left( y_p - \Phi_p \hat{u}_{w, p}^t \right)^2 \right) + \frac{1}{K_G} \text{Tr} \left( \Phi_p \Phi_p^t \right)
\]

where \( n_m \sim \mathcal{N}_C \left( n_m; 0, \frac{1}{\mathcal{P}} \sum_{p} \zeta_{p}^t l_p \right) \). For \( m = 1, \ldots, M \), (49) is a multiple-measurement-vector (MMV) problem [25], which represents simultaneously recovering multiple vectors \( \hat{h}_m \) from multiple measurements \( r_{m}^t \) with the aid of a common measurement matrix \( \Phi \). 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 \( \Phi_p = \mathcal{U}_p \text{diag} \left( s_p \right) \mathcal{V}_p^H \) and \( \phi = \mathcal{U}_\phi \text{diag} \left( s_\phi \right) \mathcal{V}_\phi^H \). For more details, we refer the motivated reader to [22]. The SVD is pre-computed off-line and corresponding results are saved. The
complexity of the proposed algorithm is on the order of \(O(K^2 R_p^2 P_{NB\text{S},N_{UE}} + K N_{BS} N_{UE} R_p L_0})\), where \(R_p\) and \(R_\Phi\) are the rank of \(\Phi_p\) and \(\Phi\), 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 \(\Phi_p\) is large and right-rotationally 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 \(\zeta_p^t\), which is shown in line 5 of Table IV. In line 5 of Table IV, \(\zeta_p^t\) is calculated by substituting the SVD of \(\Phi_p = U_p \text{diag} \left( s_p \right) V_p^H\) 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 \(\mathcal{N}(z_1; 0, I_M)\). Similarly, line 9 of Table IV is obtained from lines 9-10 of Table III, where \(z_2\) obeys the distribution of \(\mathcal{N}(z_2; 0, I_M)\). Next, other quantities are calculated by the VAMP-NNSPL-DD and are shown in Table IV. The operations \(Dz_1\) and \(Dz_2\) in lines 6 and 12 refer to \(\int \mathcal{N}(z_1; 0, I_M)dz_1\) and \(\int \mathcal{N}(z_2; 0, I_M)dz_2\), respectively. Furthermore, \(Dh = \frac{p(H)dh}{\left\|H\right\|_2^2}\) in line 13 is implemented by the classic Monte Carlo method with \(H\) generated from (10). The quantities of \(\nu_w^t, \hat{u}_w^t, \bar{s}_m, \) and \(\omega_m^{t+1}\) (right part of lines 12 and 15, respectively) are used for updating the noise variance \(\sigma^t\) (line 18). But if noise variance \(\sigma^t\) is known, these quantities could be removed. In such a case, the SE is not dependent on \(y_p, \Phi_p\) and \(\Phi\), but dependent on \(s_p\) and \(s_\Phi\) (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 algorithm [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

\[
\text{NMSE}[\text{dB}] = 10 \log \left( \mathbb{E} \left( \frac{\|\text{vec} (\hat{H} - H)\|_2^2}{\|\text{vec} (H)\|_2^2} \right) \right),
\]

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

Another 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.
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 ~ 10 physical paths, where the complex path gains are drawn from $N_c \left( \beta; 0, 10^{\text{Power}/10} \right)$, 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 $N_{BS} = 64$ and $N_{UE} = 16$, respectively. The elements of the precoders $(f_{RFk}, f_{BBmnt})$ and combiners $(W_{RF}, W_{BB})$ are of the form of $e^{j\phi}$, where $\phi$ 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-AOD-frequency 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 $r^t_m = A_m \hat{h}_m + n_m$, 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 gains [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

Figure 4: NMSE performance versus measuring time duration $G$ for 4 clusters with angular spread 15 degrees and cluster duration 9 ns. Simulation setting: $N_{BS} = 64$, $N_{UE} = 16$, $P = 512$, SNR= 10 dB, $K = 4$ and $B = 800$ MHz.
that both of them characterize the performance bound when elements of $\Psi$ obey i.i.d. Gaussian distribution. However, the MMV problem results in elements in the equivalent measurement matrix $\Psi$ 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 angular-delay 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 support 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 $\alpha \in (0, 1)$, and assume that the support only includes the location of those elements whose amplitudes are higher than the threshold $\alpha$, but not the other elements whose amplitudes are lower than the threshold $\alpha$. 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-DD 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 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.

---

6. The location of non-zero values in a sparse vector.
7. Without loss of generality, supposing that the largest amplitude of the channel elements is normalized to 1.
Figure 6: Simulation setting: domain. The AMP-NNSPL-DD requires less storage space that only exploit the sparse clustered structure in the angular attaining a considerable performance gain, over the algorithms =

We pass messages on the $N$, $\text{OMP}$, DSAMP $\text{BP}$, AMP−NNSPL−FD, AMP−EM−BG, DSAMP, SCAMPI, AMP−NNSPL−FD, SE (21), SE, VAMP−NNSPL−DD where $\delta$ is formulated as $N_{\text{BB}}$. $N_{\text{UE}} = 16$, $P = 512$, $G = 500$, 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 bound more accurately. How to attain potential performance gains with shorter measuring time will be investigated in our future work.

Appendix

We fixed the parameters $\xi$ to the values estimated at the previous iteration, i.e., $\xi = \xi^{t-1}$. 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 $w_p$ to the factor node $f_p$ as

$$u_{w_p \rightarrow f_p}^t(w_p) = N_{\mathbb{C}}(w_p; \omega_{w_p}, \gamma I_M),$$

then the belief of $w_p$ at the factor node $f_p$ is formulated as

$$\beta_f^t(w_p) = u_{w_p \rightarrow f_p}^t(w_p) N_{\mathbb{C}}(y_p; \Phi_p w_p, \sigma^t I),$$

$$\propto N_{\mathbb{C}}(w_p; \tilde{u}_{w_p}^t, v_{w_p}^t I_p),$$

with

$$\tilde{u}_{w_p}^t = \frac{1}{\sigma^t I_p} \int \exp \left( \frac{1}{\sigma^t} \Phi_p^* w_p, \gamma I_p \right) + \omega_{w_p} \right)$$

$$\beta_f^t(w_p) = \frac{1}{M} \text{Tr} \left( \left( \frac{1}{\sigma^t} \Phi_p^* \Phi_p + \gamma I_p \right)^{-1} \right)$$

$$\tilde{v}_{w_p}^t = \frac{1}{M} \text{Tr} \left( \left( \frac{1}{\sigma^t} \Phi_p^* \Phi_p + \gamma I_p \right)^{-1} \right)$$

where $g \left( \omega_{w_p}^t, \gamma, y_p, \Phi_p, \sigma^t \right)$ denotes the LMMSE estimator, and $g' \left( \omega_{w_p}^t, \gamma, y_p, \Phi_p, \sigma^t \right)$ is the derivative of $g$ with respect to $\omega_{w_p}^t$. 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 $w_p$ is calculated by $u_{w_p \rightarrow f_p}^t(w_p) = \beta_f^t(w_p) u_{w_p \rightarrow f_p}^t(w_p) \propto N_{\mathbb{C}}(w_p; r_{w_p}^t, \xi^t I_M)$, where the variance and mean are shown in line 7 of Table III. According to the sum-product message passing rule, the message $u_{h_m \rightarrow w_p}^t(w_p)$ flows through the variable node $W \in \mathbb{C}^{M \times P}$, and manifests itself as $u_{w_p \rightarrow h_m}^t(w_p) = N_{\mathbb{C}}(w_p; r_{w_p}^t, \xi^t I_M)$ with $\xi^t = \frac{1}{P} \sum \xi_p$. Let us suppose that at the $(t-1)$th iteration, the message passed from the variable node $h_m$ to the factor node $\delta_2$ is given by

$$u_{h_m \rightarrow \delta_2}^{t-1}(h_m) = N_{\mathbb{C}}(h_m; \hat{s}_{m}, \hat{s}_{m}^{-1}, \hat{k}_{m}^{-1} I_L),$$

then the belief of $h_m$ at the factor node $\delta_2$ reads

$$\beta_\delta^t(h_m) = u_{h_m \rightarrow \delta_2}^{t-1}(h_m) \hat{\delta}_2 \left( w_{m} \right) u_{w_m \rightarrow \delta_2}^{t-1} (w_{m}) dw_{m},$$

where the variance and mean are shown in line 9 of Table III. The message passed from the factor node $\delta_2$ to the variable node $h_m$ is calculated as $u_{\delta_2 \rightarrow h_m}^{t}(h_m) = \beta_\delta^t(h_m) / u_{h_m \rightarrow \delta_2}^{t-1}(h_m) \propto N_{\mathbb{C}}(h_m; \hat{s}_{m}, \hat{k}_{m}^{-1} I_L)$. In a sequel, the message $u_{\delta_2 \rightarrow h_m}^{t}(h_m)$ flows leftward, and manifests itself as $u_{\delta_1 \rightarrow h_m}^{t}(h_m) = N_{\mathbb{C}}(h_m; s_{m}^t, \hat{k}_{m} I_L)$, where the variance and mean are shown in line 10 of Table III. Similar to (33)-(35), the aposteriori distribution of $h_m$ reads

$$p \left( h_m \mid y, \xi^{t-1} \right) = \frac{1}{\int p \left( h_m \right) u_{\delta_1 \rightarrow h_m}^{t}(h_m) dh_m} \times p \left( h_m \right) u_{\delta_1 \rightarrow h_m}^{t}(h_m),$$
and the *aposteriori* variance $v_p^2$ and mean $\hat{u}_p^m$ of $h_m$ are shown in line 12 of Table III, where the function $g_{\text{mean}}(s_m^t, k^t; \xi^{-1})$ and $g_{\text{var}}(s_m^t, k^t; \xi^{-1})$ are defined in (34) and (35), respectively, which are the element-wise functions of vector $s_m^t$. Next, the message passed from the variable node $h_m$ to the factor node $\delta_1$ is calculated by $u_{h_m}^t = N_C(h_m; \hat{u}_{h_m}^t, v_{h_m}^t) / u_{\delta_1\rightarrow h_m}^t(h_m) \propto N_C(h_m; \hat{s}_m^t, k^t \cdot I_L)$. Then the message $u_{h_m}^t$ passes rightward, and manifests itself as $u_{\hat{h}_m; \delta_2}^t(h_m) = N_C(h_m; \hat{s}_m^t, k^t \cdot I_L)$, where the variance and mean are shown in line 13 of Table III, which updates the message defined in (55). Next, the belief of $w_m$, at the factor node $\delta_2$ is calculated as

$$\beta'(w_m) = \int u_{h_m; \delta_2}^t(h_m) \delta(w_m - \phi(h_m)) u_{w_m; \delta_1}^t(w_m) d\hat{h}_m,$$

and then the message passed from the factor node $\delta_2$ to the variable node $w_m$ reads $u_{\delta_2\rightarrow w_m}^t(w_m) = \beta'(w_m) / u_{w_m; \delta_2}^t(w_m) = N_C(w_m; \omega^t + 1, \eta^t + 1, \gamma^t + 1, \Gamma^t)$, 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 $p(h_m | y; \xi^{-1})$ and $p(\Phi_p w_p | y_p; \xi^{-1})$, parameters $\xi$ are updated by minimizing the Bethe free energy under a neighborhood constraint, which is similar to (38). From (52), the *aposteriori* distributions of $\Phi_p w_p$ are obtained as follows

$$p(\Phi_p w_p | y_p; \xi^{-1}) = N_C(\Phi_p w_p; \hat{\Phi}_p \hat{u}_{w_p}^t, v_{w_p}^t, \frac{1}{K_G} \text{Tr}(\Phi_p \Phi_p^H) I_{K_G}).$$

Hence, the variance of noise is updated in line 18 of Table III, where $[y_p - \Phi_p \hat{u}_{w_p}^t]^2$ 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


**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 at 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 at 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.

**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 communications. 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.

**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.