Angular-Domain Selective Channel Tracking and Doppler Compensation for High-Mobility mmWave Massive MIMO

Guanying Liu, *Student Member, IEEE*, An Liu, *Senior Member, IEEE*, Rui Zhang, *Fellow, IEEE*, and Minjian Zhao, *Member, IEEE*

***Abstract*— In this paper, we consider a mmWave massive multiple-input multiple-output (MIMO) communication system with one static base station (BS) serving a fast-moving user, both equipped with a very large array. The transmitted signal arrives at the user through multiple paths, each with a different angle-of-arrival (AoA) and hence Doppler frequency offset (DFO), thus resulting in a fast time-varying multipath fading MIMO channel. In order to mitigate the Doppler-induced channel aging for reduced pilot overhead, we propose a new angular-domain selective channel tracking and Doppler compensation scheme at the user side. Specifically, we formulate the joint estimation of partial angular–domain channel and DFO parameters as a dynamic compressive sensing (CS) problem. Then we propose a Doppler-aware-dynamic variational Bayesian inference (DD-VBI) algorithm to solve this problem efficiently. Finally, we propose a practical DFO compensation scheme which selects the dominant paths of the fast time-varying channel for DFO compensation and thereby converts it into a slow time-varying effective channel. Compared with the existing methods, the proposed scheme can enjoy the huge array gain provided by the massive MIMO and also balance the tradeoff between the CSI signaling overhead and spatial multiplexing gain. Simulation results verify the advantages of the proposed scheme over various baseline schemes.**

***Index Terms*— Massive MIMO, channel tracking, Doppler compensation, high mobility.**

I. Introduction

WITH the development of the 5G wireless communication systems, high-mobility scenarios have gained increasingly more interest. Due to the high-speed relative motion between the transmitter and receiver, the transmitted signal, propagating through multiple different paths, arrives at the receiver with different Doppler frequency offsets (DFOs), thus resulting in a fast time-varying multipath fading channel. In this case, the link performance such as achievable data rate will be degraded significantly due to the Doppler-induced channel aging effect [1]. To overcome this challenge, several methods have been proposed in previous works, as elaborated below.

Direct Channel Estimation/Prediction: For line-of-sight (LOS) channels with a single LOS path, it is relatively easy to compensate the Doppler effect and resolve its induced channel (mainly phase) variation issue by estimating the DFO parameter of the LOS path. However, when there are multiple different paths due to rich scattering, it is challenging to compensate the Doppler effect because different paths with different DFOs are mixed together in the received signal, which causes both fast-varying signal phase and amplitude over time (or fading). Nevertheless, some works have proposed to directly estimate the fast–varying channels in the time/frequency domain [2], [3]. Existing channel estimators can be classified into two types. The first type approximates time-varying channels using a linearly time-varying (LTV) channel model [4], [5]. For example, a hybrid frequency and time-domain channel estimation algorithm is proposed in [4] based on the LTV model with two methods introduced to mitigate the Doppler effect. The second type of estimators adopts the basis expansion model (BEM) to convert the problem of estimating the channel impulse response to that of estimating the basis function weights [6]. For example, the channel estimation and Doppler mitigation are jointly considered by exploiting the correlations in time and frequency domains in [6], and the basis function coefficients are estimated via the linear minimum mean squared error approach. However, accurate knowledge on the maximum DFO is required to determine the minimum order of basis function and the computational burden is heavy for multi-antenna systems [7]. Moreover, the BEM inevitably introduces approximation error to channel estimation due to the assumed imperfect model.

Orthogonal Time Frequency Space (OTFS) Modulation: OTFS modulation [8] is an emerging technique which is able to handle the fast time-varying channels. This method modulates transmitted symbols in the delay-Doppler domain instead of time-frequency domain as in traditional modulation techniques such as OFDM. The idea is to transform the time-varying multipath channel into a time-invariant channel in

Manuscript received February 27, 2020; revised August 9, 2020; accepted December 6, 2020. Date of publication December 23, 2020; date of current version May 10, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1805005 and in part by the National Science Foundation of China under Grant 62071416. The associate editor coordinating the review of this article and approving it for publication was M. Ding. *(Corresponding authors: An Liu; Minjian Zhao.)*

Guanying Liu, An Liu, and Minjian Zhao are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: anliu@zju.edu.cn; mjzhao@zju.edu.cn).

Rui Zhang is with the Department of Electronics and Communication Engineering (ECE), National University of Singapore, Singapore 117583 (e-mail: elezhang@nus.edu.sg).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2020.3045272.

Digital Object Identifier 10.1109/TWC.2020.3045272

the delay-Doppler domain. Early works on OTFS modulation focused on the single-input single-output (SISO) systems [8]. Later, OTFS is extended to multiple-input multiple-output (MIMO) systems [9]. However, there is limited work on the channel estimation method for massive MIMO-OTFS systems since a large number of antennas may lead to large pilot overhead and computational complexity.

Angular-Domain DFO Estimation and Compensation: Since the different DFOs of multiple paths are resulted from their different angles of arrival (AoAs)/angles of departure (AoDs) to/from the receiver/transmitter, they can be separated in the angle domain. As such, angular-domain DFO estimation and compensation is another popular approach to address the Doppler-induced channel aging issue [1], [10]. For MIMO systems, prior work [11] first points out that channel time-variation can be slowed down through beamforming with a large number of transmit/receive antennas. Motivated by this, a small-scale uniform circular antenna-array (UCA) is adopted in [12], [13] to separate multiple DFOs via array beamforming. However, the DFO compensation methods in [12], [13] only apply to scenarios with very sparse channels due to the limited spatial resolution of small-scale MIMO. Recently, some works have exploited high-spatial resolution provided by massive MIMO to address the high-mobility induced challenges [14], [15]. For example, the authors of [7] propose to separate the DFOs in angular-domain by beamforming with a large-scale uniform linear antenna-array (ULA) at the mobile user side. After estimating and compensating the DFO in each angle, the resultant quasi-time-invariant channel can be estimated more efficiently. However, the signaling overhead for the adopted maximum likelihood (ML) based joint estimation of the massive MIMO channel matrix and DFO parameters is prohibitively high. Moreover, only a single data stream is transmitted from the BS to the mobile user via all possible channel directions, and thus it cannot enjoy both the spatial multiplexing gain as well as the huge array signal-to-noise ratio (SNR) gain provided by massive MIMO.

Codebook-based Beam Search and Beam Tracking: The above works have focused on channel estimation. It is also possible to directly search for the best beamforming vectors without explicit channel estimation. For example, in [16], the authors propose an exhaustive search (ES) scheme, which examines transmit/receive beam pairs in the codebook and determines the best pair that maximizes a given performance metric (e.g., beamforming gain). To reduce the training overhead of the ES scheme, the hierarchical search (HS) scheme proposed in [17] utilizes hierarchical codebooks and has a favorable performance at low SNRs. However, these codebook-based beam search methods suffer from the quantization error caused by the codebook and the channel aging effect. Besides, they cannot efficiently track the fast-varying channel directions in the high-mobility scenario. To overcome these drawbacks, beam tracking is proposed for high-mobility scenario. In [18], pilots are used to track the directions of dominant channel paths changes' under a hybrid mmWave beamforming architecture. The algorithm in [19] takes two measurements using beams perturbed in angle with respect to the current beamforming direction. However, these methods do not consider the Doppler effect.

In this paper, we consider a high-mobility mmWave massive MIMO system, where both the BS and users are equipped with massive antenna arrays to facilitate Doppler compensation, improve the spectrum and energy efficiency, as well as overcome the large path loss at high frequency band. As such, combining the mmWave and massive MIMO technologies has the potential to significantly improve the capacity and reliability of high-mobility communications. However, it is very challenging to design an efficient channel estimation scheme for this system. Specifically, since both the transmit and receiver ends have large number of antennas, the dimension of the channel matrix is huge and thus conventional channel estimation will lead to large signaling overhead. Although various pilot overhead reduction methods such as those based on compressive sensing (CS) have been proposed for the estimation of slow massive MIMO fading channels [20], [21], they do not consider the Doppler effect and thus cannot be applied to the high-mobility scenario. To overcome this challenge, we propose a novel angular-domain selective channel tracking and Doppler compensation scheme, which exploits the *dynamic sparsity* of the mmWave massive MIMO channel as well as precoded training in both the downlink and uplink to significantly reduce the signaling overhead. The main contributions of this paper are summarized as follows.

- **Angular-Domain Selective Channel Tracking:** We propose a selective channel tracking scheme to only estimate partial channel parameters at the user side that are sufficient for Doppler compensation to significantly reduce the pilot overhead. Moreover, we propose an efficient downlink training vector design at the BS side to strike a balance between *exploitation* of most promising channel directions for array (SNR) gain and *exploration* of unknown channel directions. Compared to the conventional training vector design, the proposed design can exploit the array gain to further enhance the channel tracking performance.
- **Angular-Domain Selective Doppler Compensation:** We propose an angular-domain selective DFO compensation scheme at the user side which selectively converts the dominant paths of the fast time-varying channel into a slow time-varying effective channel. Compared to the non-selective DFO compensation scheme in [22], the proposed scheme can enjoy the huge array (SNR) gain provided by the massive MIMO and also balance the tradeoff between the CSI signaling overhead reduction and spatial multiplexing gain maximization.
- **Channel Tracking Algorithm based on Dynamic VBI:** To further reduce the pilot overhead, the proposed partial channel tracking design is formulated as a dynamic CS problem with unknown Doppler parameters in the measurement matrix. Then, we adopt a three-layer hierarchical Markov model to capture the *dynamic sparsity* of the partial angular–domain channel. The existing methods, such as Variational Bayesian inference (VBI) [23] and Sparse Bayesian Learning (SBL) [24], cannot be directly applied to this three-layer prior. To address this challenge,

TABLE I
THE MAIN NOTATIONS USED IN THE PAPER

Notation	Meaning	Notation	Meaning
N_p	Number of downlink training vectors	$\theta_{T,t,q}\left(\theta_{R,t,q}\right)$	The AoD (AoA) of the q-th path
$\mathbf{v}_t$	Downlink training vector	η_t	Rotation angle of user's antenna array
$M\left(N\right)$	Number of BS (user) antennas	$\tilde{\theta}_{T,m}\left(\tilde{\theta}_{R,m}\right)$	m-th AoD grid (AoA grid)
L_t	Number of propagation paths	$\boldsymbol{\beta}_{T,t}\left(\boldsymbol{\beta}_{R,t}\right)$	The AoD(AoA) off-grid vector
$\alpha_{t,q}$	The path gain of the q-th path	N_b	Number of RF chains at the user
$f_{d,t}$	The maximum DFO	$\tilde{M}\left(\tilde{N}\right)$	Number of AoD (AoA) grid

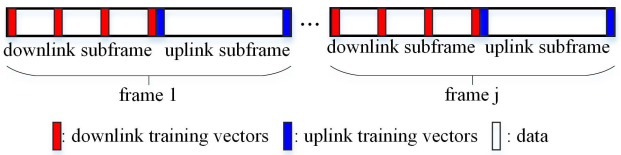


Fig. 1. Illustration of frame structure.

we propose a new Doppler-aware-dynamic Variational Bayesian inference (DD-VBI) algorithm, which combines the VBI and message-passing approaches to achieve superior channel tracking performance.

The rest of this paper is organized as follows. In Section II, we describe the system model and frame structure. In Section III, we give a brief introduction of the proposed angular-domain channel tracking and Doppler compensation scheme. In Sections IV and V, we present the three-layer hierarchical model for partial channel vector and the proposed DD-VBI algorithm, respectively. The simulation results and conclusions are given in Sections VI and VII, respectively.

Notations: For a set of scalars $\{x_1, \ldots, x_N\}$ and an index set $\mathcal{S} \subseteq \{1, \ldots, N\}$, we use $[x_n]_{n\in\mathcal{S}}$ to denote a $|\mathcal{S}| \times 1$ column vector consisting of the elements of $\{x_1, \ldots, x_N\}$ indexed by the set $\mathcal{S}$. For a set of matrices $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ with $\mathbf{X}_n \in \mathbb{C}^{N\times M}$, $[\mathbf{X}_n]_{n\in\mathcal{S}}$ denotes an $N \times M\,|\mathcal{S}|$ matrix consisting of the elements of $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ indexed by the set $\mathcal{S}$. When $M = 1$, $\mathbf{X}_n$ reduces to a vector and $[\mathbf{X}_n]_{n\in\mathcal{S}} \in \mathbb{C}^{N\times|\mathcal{S}|}$. We use $\mathbf{X}\left(:, j\right)$ to denote the j-th column of a matrix $\mathbf{X}$. The main notations used in this paper are summarized in Table I.

II. SYSTEM MODEL

A. System Architecture and Frame Structure

Consider a TDD mmWave massive MIMO system with one BS serving a fast-moving user[1]. The BS is equipped with M antennas. The user is equipped with N antennas. The time is divided into frames, with each frame containing a downlink subframe and an uplink subframe, as illustrated in Fig. 1. Each downlink/uplink subframe contains a large number of symbol durations. The length of the uplink/downlink subframe can be adjusted to accommodate non-symmetric uplink/downlink data traffic. Such a frame structure is consistent with that of practical TDD systems.

[1]For clarity, we focus on a single user system. However, the proposed selective channel tracking and Doppler compensation scheme can be readily extended to multi-user systems.

In the t-th downlink subframe, there are N_p uniformly distributed training vectors, which are set to be the same vector denoted as $\mathbf{v}_t$. For convenience, we use $\mathcal{N}_p$ to denote the symbol index set for the N_p training vectors. Note that inserting N_p identical training vectors uniformly in the downlink subframe facilitates the estimation of the (partial) channel parameters at the user side, as will be explained later. Based on the estimated parameters, the user applies a Doppler compensation matrix to mitigate the Doppler effect and essentially converts the fast time-varying channel into a slow time-varying effective channel. In the uplink, there are two sets of training vectors at the beginning and end of each uplink subframe, respectively. The uplink training vectors are used to estimate the slow time-varying effective channel after Doppler compensation. Specifically, the uplink transmission (e.g., beamforming and power allocation) in the t-th uplink subframe is optimized based on the slow time-varying effective channel estimated at the t-th uplink subframe. On the other hand, by making use of the channel reciprocity, the downlink transmission in the t-th downlink subframe is optimized based on the slow time-varying effective channel estimated at the end of the $(t-1)$-th uplink subframe. Since the effective channel after Doppler compensation changes slowly compared to the subframe duration, such a design can effectively overcome the channel aging issue caused by the Doppler effect.

B. Doppler and Multipath Channel Models

For the purpose of exposition, we focus on the case when both the BS and mobile user are equipped with a half-wavelength-space ULA, while the system is narrow-band (relative to the mmWave high carrier frequency) and thus the channel is flat-fading. To incorporate the DFO in practical mmWave channels, the downlink channel matrix for the i-th symbol in t-th frame is given by [21]

$$\boldsymbol{H}_{t,i} = \sum_{q=1}^{L_t} \alpha_{t,q} e^{j2\pi f_{d,t} i cos(\theta_{R,t,q}+\eta_t)} \boldsymbol{a}_R\left(\theta_{R,t,q}\right) \boldsymbol{a}_T^H\left(\theta_{T,t,q}\right), \tag{1}$$

where L_t is the total number of propagation paths, $\alpha_{t,q}$ is the complex path gain associated with the q-th path, $f_{d,t}$ is the normalized maximum DFO, $\theta_{T,t,q}$ and $\theta_{R,t,q}$ are the AoD and AoA of the q-th path, respectively, and η_t is rotation angle of the user's antenna array with respect to the moving direction. $\boldsymbol{a}_T\left(\theta_{T,t,q}\right)$ and $\boldsymbol{a}_R\left(\theta_{R,t,q}\right)$ are the array response vectors for the BS antenna array (associated with the q-th AoD $\theta_{T,t,q}$)

and user antenna array (associated with the q-th AoA $\theta_{R,t,q}$), respectively. For half-wavelength-space ULAs, $\boldsymbol{a}_T(\theta) = \frac{1}{\sqrt{M}}\left[1, e^{-j\pi sin(\theta)}, e^{-j2\pi sin(\theta)}, \ldots, e^{-j(M-1)\pi sin(\theta)}\right]^T$ and $\boldsymbol{a}_R(\theta) = \frac{1}{\sqrt{N}}[1, e^{-j\pi sin(\theta)}, e^{-j2\pi sin(\theta)}, \ldots, e^{-j(N-1)\pi sin(\theta)}]^T$.

Note that in (1), we have implicitly assumed that the channel parameters $L_t, \alpha_{t,q}, \theta_{R,t,q}, \theta_{T,t,q}, f_{d,t}, \eta_t$ are fixed within each frame but may change over different frames, which is usually true even for high-speed users [22]. However, the channel matrix $\boldsymbol{H}_{t,i}$ may change over symbols at a much faster timescale due to the fast changing phase terms of multiple paths $2\pi f_{d,t} icos\left(\theta_{R,t,q} + \eta_t\right)$, caused by their different DFOs.

C. Angular Domain Channel Representation

In this subsection, we describe the angular domain channel representation at t-th frame. For ease of notation, we drop the frame index t. In this case, the AoDs are denoted as $\{\theta_{T,1}, \ldots, \theta_{T,L}\}$ and the AoAs are denoted as $\{\theta_{R,1}, \ldots, \theta_{R,L}\}$. To obtain the angular domain channel representation, we introduce a uniform grid of $\tilde{M}$ AoDs and $\tilde{N}$ AoAs over $[-\pi/2, \pi/2]$:

$$\{\tilde{\theta}_{T,m} : sin\left(\tilde{\theta}_{T,m}\right) = \frac{2}{\tilde{M}}\left(m - \left\lfloor\frac{\tilde{M}-1}{2}\right\rfloor\right), \quad m = 0, \ldots, \tilde{M}-1\},$$

$$\{\tilde{\theta}_{R,n} : sin\left(\tilde{\theta}_{R,n}\right) = \frac{2}{\tilde{N}}\left(n - \left\lfloor\frac{\tilde{N}-1}{2}\right\rfloor\right), \quad n = 0, \ldots, \tilde{N}-1\}.$$

In practice, the true AoDs/AoAs usually do not lie exactly on the grid points. In this case, there will be mismatches between the true AoDs/AoAs and the nearest grid point. To overcome this issue, we introduce an off-grid basis for the angular domain channel representation, as in [24]. Specifically, let $\tilde{\theta}_{T,m_q}$ and $\tilde{\theta}_{R,n_q}$ denote the nearest grid point to $\theta_{T,q}$ and $\theta_{R,q}$, respectively. We introduce the AoD off-grid vector $\boldsymbol{\beta}_T = \left[\beta_{T,1}, \beta_{T,2}, \ldots, \beta_{T,\tilde{M}}\right]^T$ such that

$$\beta_{T,m} = \begin{cases} \theta_{T,q} - \tilde{\theta}_{T,m_q}, & m = m_q, \quad q = 1, 2, \ldots, L \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, let $\boldsymbol{\beta}_R = \left[\beta_{R,1}, \beta_{R,2}, \ldots, \beta_{R,\tilde{N}}\right]^T$ denote the AoA off-grid vector, such that

$$\beta_{R,n} = \begin{cases} \theta_{R,q} - \tilde{\theta}_{R,n_q}, & n = n_q, \quad q = 1, 2, \ldots, L \\ 0, & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{\varphi} = \{\boldsymbol{\beta}_R, f_d, \eta\}$ denote the set of channel parameters at the user side and $\boldsymbol{\varphi}^n = \{\beta_{R,n}, f_d, \eta\}$ denote the set of channel parameters associated with the n-th AoA grid at the user side. Define two matrices $\boldsymbol{A}_{R,i}(\boldsymbol{\varphi}) = [\tilde{\boldsymbol{a}}_{R,i}(\boldsymbol{\varphi}^1), \ldots, \tilde{\boldsymbol{a}}_{R,i}(\boldsymbol{\varphi}^{\tilde{N}})] \in \mathbb{C}^{N \times \tilde{N}}$ and $\boldsymbol{A}_T(\boldsymbol{\beta}_T) = [\boldsymbol{a}_T(\tilde{\theta}_{T,1} + \beta_{T,1}), \ldots, \boldsymbol{a}_T(\tilde{\theta}_{T,\tilde{M}} + \beta_{T,\tilde{M}})] \in \mathbb{C}^{M \times \tilde{M}}$ where $\tilde{\boldsymbol{a}}_{R,i}(\boldsymbol{\varphi}^n) = \boldsymbol{a}_R(\tilde{\theta}_{R,n} + \beta_{R,n}) \times e^{j2\pi f_d icos(\tilde{\theta}_{R,n} + \beta_{R,n} + \eta)}$. Furthermore, define $\tilde{\boldsymbol{X}}$ as the angular domain channel matrix with the (n, m)-th element given by

$$\tilde{x}_{n,m} = \begin{cases} \alpha_q, & (n, m) = (n_q, m_q), \quad q = 1, 2, \ldots, L \\ 0, & \text{otherwise.} \end{cases}$$

Given $\boldsymbol{\varphi}$ and $\boldsymbol{\beta}_T$, $\boldsymbol{H}_i$ can be expressed in a compact form as

$$\boldsymbol{H}_i\left(\boldsymbol{\varphi}, \boldsymbol{\beta}_T\right) = \boldsymbol{A}_{R,i}\left(\boldsymbol{\varphi}\right) \tilde{\boldsymbol{X}} \boldsymbol{A}_T^H\left(\boldsymbol{\beta}_T\right). \tag{2}$$

Note that we can also define the angular domain representation for general 2-dimensional antenna arrays. In this case, the array response vector $\boldsymbol{a}_T\left(\theta, \phi\right)$ (or $\boldsymbol{a}_R\left(\theta, \phi\right)$) can be expressed as a function of the azimuth angle θ and elevation angle ϕ. Please refer to [25] for the details.

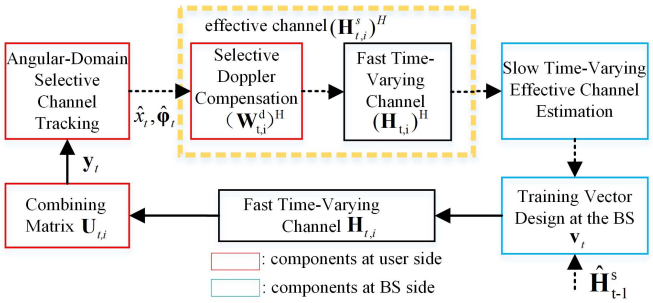


Fig. 2. A top-level diagram of the proposed scheme.

III. Angular-Domain Selective Channel Tracking and Doppler Compensation

In this section, we propose an efficient angular-domain selective channel tracking and Doppler compensation scheme at the user side. The proposed scheme can exploit both *dynamic sparsity* of mmWave massive MIMO channel and high-resolution of AoA at multi-antenna mobile users to accurately estimate the partial angular channel coefficients, AoAs, rotation angle and maximum DFO. Using these estimated parameters, a Doppler compensation matrix is applied at the user to convert the high-dimensional fast time-varying channel into a slow time-varying effective channel with lower dimension, based on which efficient downlink/uplink transmissions can be achieved. The proposed scheme includes four key components, namely, the Angular-Domain Selective Channel Tracking, Selective Doppler Compensation, Slow Time-Varying Effective Channel Estimation, and Downlink Training Vector Design. Fig. 2 illustrates a top-level diagram of the proposed scheme while the details of each component are elaborated as follows. The frame index t will be omitted in this section when there is no ambiguity.

A. Angular-Domain Selective Channel Tracking at the User

One important purpose of this component is to accurately estimate the parameters required by the Doppler compensation component. Observing from the channel model (1), the DFO only occurs at the user side and the AoA has a one-to-one relationship with the DFO parameters. Thanks to the high-spatial resolution provided by the large array at the user side, the user can distinguish DFOs associated with

different AoAs from multiple active paths. However, since both the BS and the user are equipped with a large array, the parameter space can be very large if we attempt to estimate the full angular–domain channel parameters (i.e., the full angular domain channel matrix $\tilde{\boldsymbol{X}}$, rotation angle η and maximum DFO f_d). In order to reduce signaling overhead and enhance the channel estimation performance, we propose to only estimate partial angular-domain channel parameters that are just sufficient to obtain channel coefficients, AoAs, rotation angle and maximum DFO for Doppler compensation.

Specifically, the product of channel $\boldsymbol{H}_i$ and downlink training vector $\mathbf{v}$ can be expressed as

$$\begin{aligned}\boldsymbol{H}_i\mathbf{v} &= \sum_{n=1}^{\tilde{N}}\sum_{m=1}^{\tilde{M}} \tilde{x}_{n,m}\tilde{\boldsymbol{a}}_{R,i}\left(\boldsymbol{\varphi}^n\right)\boldsymbol{a}_T^H\left(\tilde{\theta}_{T,m}+\beta_{T,m}\right)\mathbf{v} \\ &= \sum_{n=1}^{\tilde{N}} x_n\tilde{\boldsymbol{a}}_{R,i}\left(\boldsymbol{\varphi}^n\right) = \boldsymbol{A}_{R,i}\left(\boldsymbol{\varphi}\right)\boldsymbol{x}, \end{aligned} \tag{3}$$

where $\boldsymbol{x} = [x_1,\ldots,x_{\tilde{N}}]^T$ with $x_n = \sum_{m=1}^{\tilde{M}}\tilde{x}_{n,m}\boldsymbol{a}_T^H(\theta_{T,m}+\beta_{T,m})\mathbf{v}$ are called partial angular channel coefficients since the AoD parameters are integrated into $\boldsymbol{x}$ and $\boldsymbol{x}$ only contains partial information about the full angular channel $\tilde{\boldsymbol{X}}$. Specifically, a non-zero $|x_n|^2$ with value larger than the noise floor indicates that there is an active path to the n-th AoA direction at the user side. Therefore, we only need to estimate $\tilde{N}$ partial channel coefficients $\boldsymbol{x}$, the AoA off-grid vector $\boldsymbol{\beta}_R$, rotation angle η and maximum DFO f_d, which are much less than the original $\tilde{N}\tilde{M}$ full channel parameters. Note that if the N_p training vectors are different, there will be $\tilde{N}N_p$ partial angular channel coefficients, leading to a larger parameter space to be estimated. In addition, as detailed in Section III-D, the downlink training vector $\mathbf{v}$ is chosen such that the transmit power can focus on the directions associated with the active AoDs to exploit the massive array gain at the BS side, according to the available prior information on the AoDs. Using different trainning vectors $\mathbf{v}_i$'s cannot provide additional array gain, but will cause larger estimation error and/or higher signaling overhead due to the enlarged parameter space. Moreover, with uniformly distributed training vectors, the phase rotation due to the Doppler term $e^{j2\pi f_d i cos(\theta_{R,q}+\eta_t)}$ is larger compared to the case when the N_p training vectors are squeezed in the beginning of the downlink subframe, leading to a better estimation performance for the Doppler parameter f_d. Therefore, such a selective channel tracking design based on N_p uniformly distributed and identical training vectors can significantly reduce the number of downlink training vectors N_p required to achieve accurate estimation of partial channel coefficients, AoAs, rotation angle and maximum Doppler.

The aggregate received pilot signal (channel measurements) of all the N_p downlink pilot symbols (training vectors) in the t-th frame can be expressed in a compact form as

$$\boldsymbol{y} = \left[\boldsymbol{H}_i\mathbf{v}+\boldsymbol{n}_i\right]_{i\in\mathcal{N}_p}. \tag{4}$$

where $\boldsymbol{n}_i$ is the additive white Gaussian noise (AWGN) at the user with each element having zero mean and variance σ^2.

Based on the received pilot signal, the user obtains the estimated partial channel coefficients $\hat{\boldsymbol{x}}$, the AoA off-grid vector $\hat{\boldsymbol{\beta}}_R$, rotation angle $\hat{\eta}$ and maximum DFO $\hat{f}_d$ using a selective channel tracking algorithm. The detailed problem formulation and algorithm design for selective channel tracking scheme are postponed to section IV.

B. Angular-Domain Selective Doppler Compensation at the User

This component is used to convert the high-dimensional fast time-varying channel into a slow time-varying effective channel with lower dimension, using the estimated partial channel parameters $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{\beta}}_R$, $\hat{\eta}$ and $\hat{f}_d$ obtained from the Angular-Domain Selective Channel Tracking component. We first select a set of N_d most significant AoA directions (beams) with the largest energy, where the energy of the n-th AoA direction $\tilde{\theta}_{R,n}+\hat{\beta}_{R,n}$ is defined as $|\hat{x}_n|^2$. The parameter $N_d \le N$ is used to control the tradeoff between the spatial multiplexing gain and the effective CSI signaling overhead (i.e., the CSI signaling overhead required to obtain effective channel $\boldsymbol{H}_i^s$ in (5)). Let $\mathcal{N}_d \subseteq \{1,\ldots,N\}$ denote the index set of the selected N_d most significant AoA directions. Since the n-th selected AoA direction, where $n \in \mathcal{N}_d$, is affected by a single DFO term $e^{j2\pi f_d i cos\left(\tilde{\theta}_{R,n}+\beta_{R,n}+\eta\right)}$, the per-AoA DFO compensation for this direction can be achieved by left-multiplying the channel matrix $\boldsymbol{H}_i$ with the DFO compensation vector $\tilde{\boldsymbol{a}}_{R,i}^H\left(\hat{\boldsymbol{\varphi}}^n\right) = \boldsymbol{a}_R^H\left(\tilde{\theta}_{R,n}+\hat{\beta}_{R,n}\right)\times e^{-j2\pi\hat{f}_d i cos\left(\tilde{\theta}_{R,n}+\hat{\beta}_{R,n}+\eta\right)}$. To perform the DFO compensation for all the N_d significant AoA directions, a DFO compensation matrix $\mathbf{W}_i^d = \left[\tilde{\boldsymbol{a}}_{R,i}\left(\hat{\boldsymbol{\varphi}}^n\right)\right]_{n\in\mathcal{N}_d} \in \mathbb{C}^{N\times N_d}$, which also serves as beamforming matrix, is applied at the user side. In this way, we can convert the fast time-varying channel $\boldsymbol{H}_i$ into a slow time-varying effective channel $\boldsymbol{H}_i^s \in \mathbb{C}^{N_d\times M}$ as

$$\boldsymbol{H}_i^s = \left(\mathbf{W}_i^d\right)^H\boldsymbol{H}_i. \tag{5}$$

In the following, we explain why the Doppler effect can be alleviated by applying the above DFO compensation matrix to obtain an effective channel $\boldsymbol{H}_i^s$. For half-wavelength-space ULA, if there is no estimation error for the partial channel parameters, we have

$$\begin{aligned}\boldsymbol{H}_i^s = \sum_{m=1}^{\tilde{M}}\left[\tilde{x}_{n,m}+\sum_{\tilde{n}=1,\tilde{n}\neq n}^{\tilde{N}}\tilde{x}_{\tilde{n},m}\tilde{\boldsymbol{a}}_{R,i}^H\left(\boldsymbol{\varphi}^n\right)\tilde{\boldsymbol{a}}_{R,i}\left(\boldsymbol{\varphi}^{\tilde{n}}\right)\right]_{n\in\mathcal{N}_d} \\ \times\boldsymbol{a}_T^H(\tilde{\theta}_{T,m}+\beta_{T,m}) = \boldsymbol{H}^s+\triangle\boldsymbol{H}_i\end{aligned}$$

where $\boldsymbol{H}^s = \sum_{m=1}^{\tilde{M}}\left[\tilde{x}_{n,m}\right]_{n\in\mathcal{N}_d}\boldsymbol{a}_T^H\left(\theta_{T,m}+\beta_{T,m}\right)$ is the slow time-varying component whose Doppler effect is completely eliminated (i.e., $\boldsymbol{H}^s$ remain constant within a frame),

$$\begin{aligned}\triangle\boldsymbol{H}_i = \sum_{m=1}^{\tilde{M}}\left[\sum_{\tilde{n}=1,\tilde{n}\neq n}^{\tilde{N}}\tilde{x}_{\tilde{n},m}\tilde{\boldsymbol{a}}_{R,i}^H\left(\boldsymbol{\varphi}^n\right)\tilde{\boldsymbol{a}}_{R,i}\left(\boldsymbol{\varphi}^{\tilde{n}}\right)\right]_{n\in\mathcal{N}_d} \\ \times\boldsymbol{a}_T^H\left(\theta_{T,m}+\beta_{T,m}\right)\end{aligned}$$

is the fast time-varying component due to the energy leakage effect (i.e., $\triangle\boldsymbol{H}_i$ changes over different symbol durations).

It follows from a direct calculation that the energy (the second order moment) of each element of $\triangle \boldsymbol{H}_i$ is $\mathcal{O}\left(\frac{L}{N^2}\right)$. Therefore, for large N, $\triangle \boldsymbol{H}_i$ is negligible.

From the above discussions, we can see that the proposed angular-domain selective Doppler compensation exploits the asymptotical features of massive MIMO. Specifically, the high-spatial resolution of the massive antenna array provides the opportunity to perform the per-AoA DFO compensation in angle domain. In addition, the energy of the fast time-varying component $\triangle \boldsymbol{H}_i$ is negligible due to the fact that when the number of user antennas N is large, the array response vectors $\boldsymbol{a}_R\left(\tilde{\theta}_{R,n}+\beta_{R,n}\right)$ associated with different AoAs are nearly orthogonal to each other.

C. Slow Time-Varying Effective Channel Estimation at the BS

Under typical cases when the number of RF chains at the BS is one half or a quarter of the number of BS antennas, the user can simply transmit $2N_d$ to $4N_d$ orthogonal pilots in the uplink training stage for the BS to estimate $\boldsymbol{H}_i^s$ using the conventional Least Square based channel estimation method. Based on $\hat{\boldsymbol{H}}_i^s$, the BS can optimize the precoder for both uplink and downlink transmissions. The optimization of MIMO precoder is a standard problem and there are many existing solutions with different performance and complexity tradeoff. Then, the optimized uplink precoder is fed back to the user for uplink transmission. Note that the coherence time of $\boldsymbol{H}_i^s$ is much larger than the symbol durations after Doppler compensation [22], and therefore each frame, which should be chosen to be smaller than the coherence time of $\boldsymbol{H}_i^s$, can contain a large number of symbols (much larger than N_d since $N_d = O(1) \ll N$). Therefore, the uplink trainning overhead and the feedback overhead for the optimized uplink precoder is acceptable for practice.

D. Training Vector Design at the BS

The downlink training vector $\mathbf{v}_t$ is designed according to the slow time-varying effective channel $\hat{\boldsymbol{H}}_{t-1}^s$ estimated at the end of the $(t-1)$-th uplink subframe. The basic idea for training vector design is to strike a balance between *exploitation* of known channel directions (i.e., transmitting training signal over the most promising channel directions with large effective channel energy to achieve beamforming gain) and *exploration* of unknown channel directions (i.e., transmitting training signal over other channel directions to detect unknown channel directions). Since the effective channel $\boldsymbol{H}_i^s$ changes slowly, the effective channel $\hat{\boldsymbol{H}}_{t-1}^s$ estimated at the end of the $(t-1)$-th uplink subframe is expected to provide valuable information for the most promising channel directions. On the other hand, the information about the most promising channel directions extracted from the previously estimated slow time-varying effective channel $\hat{\boldsymbol{H}}_{t-1}^s$ may not be perfect due to the estimation error and CSI delay. In addition, some new direction may arise in the next frame. Therefore, other channel directions should also be incorporated into the training vector to facilitate the detection of unknown channel directions.

Specifically, we first project the estimated effective channel $\hat{\boldsymbol{H}}_{t-1}^s$ onto an orthogonal basis $\mathbf{B}^s = [\mathbf{b}_1^s, \ldots, \mathbf{b}_M^s] \in \mathbb{C}^{M\times M}$

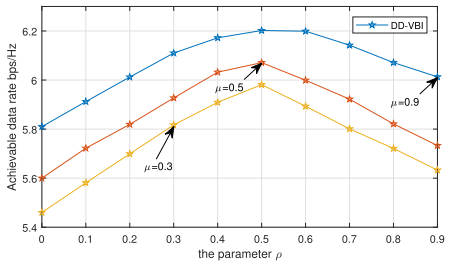


Fig. 3. Achievable data rate versus the parameter ρ with different values of the parameter μ.

to obtain the effective channel energy on each basis vector (quantized channel direction) as $\lambda_m^s = \left\|\hat{\boldsymbol{H}}_{t-1}^s \mathbf{b}_m^s\right\|^2, \forall m$. The basis matrix is chosen such that the projection vector $\boldsymbol{\lambda}^s = [\lambda_1^s, \ldots, \lambda_M^s]^T$ is as sparse as possible. For half-wavelength ULAs, we can simply choose the basis matrix $\mathbf{B}^s$ as an $M \times M$ DFT matrix. Then we find the index set $\mathcal{M}$ of the most promising channel directions as

$$\mathcal{M}^* = \text{argmin}_{\mathcal{M}} |\mathcal{M}|, \quad \text{s.t.} \sum_{m\in\mathcal{M}} \lambda_m^s / \sum_{m=1}^{M} \lambda_m^s \geq \mu, \quad (6)$$

where μ is a threshold which is chosen to be closed to 1. In other words, the most promising channel directions contain μ fraction of the total effective channel energy. Let $N_s = |\mathcal{M}^*|$. Finally, the training vector is given by

$$\mathbf{v}_t = \frac{\sqrt{\rho}}{\sqrt{N_s}} \sum_{m\in\mathcal{M}^*} e^{j\theta_m^s} \mathbf{b}_m^s + \frac{\sqrt{1-\rho}}{\sqrt{M-N_s}} \sum_{m\in\{1,\ldots,M\}\setminus\mathcal{M}^*} e^{j\theta_m^s} \mathbf{b}_m^s. \quad (7)$$

where the first term in (7) exploits the information about the most promising N_s channel directions extracted from the previously estimated slow time-varying effective channel $\hat{\boldsymbol{H}}_{t-1}^s$, ρ is a system parameter which determines the proportion of transmit power used to exploit the most promising channel directions, the second term is used to detect the other $M-N_s$ unknown channel directions, θ_m^s is randomly generated from $[0, 2\pi]$. Note that $\mathbf{v}_t$ is a linear combination of all possible directions $\mathbf{b}_m^s$'s that span the whole space and there is always some energy transmitted on any path. The phase term θ_m^s does not affect the distribution of the transmit energy on different directions $\mathbf{b}_m^s$'s. Consequently, the choice of the θ_m^s has little impact on the performance of proposed selective channel tracking algorithm.

In Fig. 3, we illustrate how the achievable data rate is affected by changing the parameters ρ and μ to achieve different tradeoffs between the exploration and exploitation. It can be seen that setting $\mu = 0.9$ and $\rho = 0.5$ can strike a good balance between *exploitation* and *exploration*.

IV. Problem Formulation for Angular-Domain Selective Channel Tracking

A. Three-Layer Hierarchical Markov Model for Partial Angular Channel Vector

In the massive MIMO regime, due to the limited scattering, only a few dimensions in the angular domain are occupied,

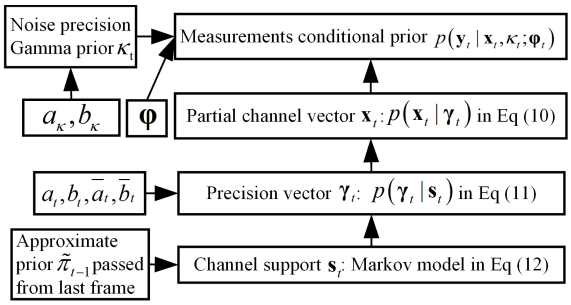


Fig. 4. Three-layer hierarchical Markov model for partial angular channel coefficients.

which, in return, brings a sparse representation for the massive MIMO channels [26], [27]. In addition, the support of the partial angular channel coefficients $\boldsymbol{x}_t$ also has temporal correlation [28]. Such a *dynamic sparsity* of $\boldsymbol{x}_t$ (i.e., sparsity with support having temporal correlation) is captured using a three-layer hierarchical Markov model, as illustrated in Fig. 4. The first layer of random variable is the channel support vector $\boldsymbol{s}_t \in \{0,1\}^{\tilde{N}}$, whose n-th element, denoted by $s_{t,n}$, indicates whether there is an active path in the n-th AoA direction in the t-th frame ($s_{t,n} = 1$) or not ($s_{t,n} = 0$). The second layer of random variable is the precision vector $\boldsymbol{\gamma}_t = \left[\gamma_{t,1}, \cdots, \gamma_{t,\tilde{N}}\right]^T$, where $\gamma_{t,n}$ represents the precision (inverse of the variance) of $x_{t,n}$. The third layer of random variables are partial angular channel coefficients $\boldsymbol{x}_t$. For convenience, we denote a time series of vectors $\{\boldsymbol{x}_\tau\}_{\tau=1}^t$ as $\boldsymbol{x}_{1:t}$ (same for $\boldsymbol{\gamma}_{1:t}$, $\boldsymbol{s}_{1:t}$, $f_{d,1:t}$, $\eta_{1:t}$, $\boldsymbol{\beta}_{R,1:t}$, $\boldsymbol{\varphi}_{1:t}$). Then the three-layer hierarchical Markov prior distribution (joint distribution of $\boldsymbol{x}_{1:t}$, $\boldsymbol{\gamma}_{1:t}$ and $\boldsymbol{s}_{1:t}$) is given by

$$p\left(\boldsymbol{x}_{1:t}, \boldsymbol{\gamma}_{1:t}, \boldsymbol{s}_{1:t}\right) = \prod_{\tau=1}^{t} p\left(\boldsymbol{s}_\tau | \boldsymbol{s}_{\tau-1}\right) p\left(\boldsymbol{\gamma}_\tau | \boldsymbol{s}_\tau\right) p\left(\boldsymbol{x}_\tau | \boldsymbol{\gamma}_\tau\right), \tag{8}$$

where $p\left(\boldsymbol{s}_1 | \boldsymbol{s}_0\right) \triangleq p\left(\boldsymbol{s}_1\right)$, the conditional probability $p\left(\boldsymbol{x}_\tau | \boldsymbol{\gamma}_\tau\right)$ has a product form $p\left(\boldsymbol{x}_\tau | \boldsymbol{\gamma}_\tau\right) = \prod_{n=1}^{\tilde{N}} p\left(x_{\tau,n} | \gamma_{\tau,n}\right)$ and each is modeled as a Gaussian prior distribution.

$$p\left(x_{\tau,n} | \gamma_{\tau,n}\right) = CN\left(x_{\tau,n}; 0, \gamma_{\tau,n}^{-1}\right), \tag{9}$$

The conditional prior of precision vector $\boldsymbol{\gamma}_\tau$ is given by

$$p\left(\boldsymbol{\gamma}_\tau | \boldsymbol{s}_\tau\right) = \prod_{n=1}^{\tilde{N}} \Gamma\left(\gamma_{\tau,n}; a_\tau, b_\tau\right)^{s_{\tau,n}} \Gamma\left(\gamma_{\tau,n}; \overline{a}_\tau, \overline{b}_\tau\right)^{1-s_{\tau,n}}, \tag{10}$$

$\Gamma\left(\gamma; a_\gamma, b_\gamma\right)$ is a Gamma hyperprior. a_τ, b_τ are the shape and rate parameters of the channel precision $\gamma_{\tau,n}$ conditioned on $s_{\tau,n} = 1$ and they should be chosen such that $\frac{a_\tau}{b_\tau} = E\left[\gamma_{\tau,n}\right] = \Theta\left(1\right)$, since the variance $\gamma_{\tau,n}^{-1}$ of $x_{\tau,n}$ is $\Theta(1)$ when it is active ($s_{\tau,n} = 1$) and the corresponding entries $x_{\tau,n}$ will have large probability to be deviated from zero. $\overline{a}_\tau, \overline{b}_\tau$ are the shape and rate parameters, conditioned on the opposite event. In this case, the shape and rate parameters $\overline{a}_\tau, \overline{b}_\tau$ of the precision $\gamma_{\tau,n}$ should be chosen such that $\frac{\overline{a}_\tau}{\overline{b}_\tau} = E\left[\gamma_{\tau,n}\right] \gg 1$, since the variance $\gamma_{\tau,n}^{-1}$ of $x_{\tau,n}$ is close to zero when it is inactive.

Remark 1: In practice, the exact distribution of each element of the sparse channel $\boldsymbol{x}_\tau$ is usually unknown. In this case, it is reasonable to choose a prior distribution such that the devised algorithm can exploit sparsity with low complexity and achieve robust performance to different channel distributions. The hierarchical Gaussian prior with Gamma distributed variance considered in (9) and (10) has been widely used to exploit sparsity for sparse signals with unknown sparse distributions [23], [29]. By controlling the parameters in the Gamma distribution of $\gamma_{\tau,n}$, one can easily promote sparsity based on the knowledge of channel support $\boldsymbol{s}_\tau$. Moreover, since the Gamma distribution for $\gamma_{\tau,n}$ is the conjugate probability distribution of the Gaussian distribution for $x_{\tau,n}$, the above hierarchical prior for $x_{\tau,n}$ and $\gamma_{\tau,n}$ facilitates low-complexity VBI algorithm design with closed-form update equations [23]. Finally, the VBI-type algorithms derived from such a hierarchical prior are well known to be insensitive to the true distribution of the sparse signals [23], [29].

Due to the slowly changing propagation environment, the channel supports often change slowly over time, which implies that $s_{\tau,n}$ depends on $s_{\tau-1,n}$, e.g., if $s_{\tau-1,n} = 1$, then there is a higher probability that $s_{\tau,n}$ is also 1 [28]. Such temporal correlation of support vectors can be naturally modeled as a temporal Markov model in (11) with an initial prior distribution $p\left(\boldsymbol{s}_1\right)$ and a transition probability:

$$p\left(\boldsymbol{s}_\tau | \boldsymbol{s}_{\tau-1}\right) = \prod_{n=1}^{\tilde{N}} p\left(s_{\tau,n} | s_{\tau-1,n}\right), \tag{11}$$

where transition probability is given by $p(s_{\tau,n} = 1 | s_{\tau-1,n} = 0) = \rho_{0,1}$, and $p(s_{\tau,n} = 0 | s_{\tau-1,n} = 1) = \rho_{1,0}$. The Markov parameters $\{\rho_{1,0}, \rho_{0,1}\}$ characterize the degree of temporal correlation of the channel supports. Specifically, smaller $\rho_{1,0}$ or $\rho_{0,1}$ leads to highly correlated supports across time, which means the propagation environment between the user and BS is changing slowly. Larger $\rho_{1,0}$ or $\rho_{0,1}$ can allow support to change substantially across time, which means the propagation environment is changing significantly. Moreover, the statistic parameters $\{\rho_{1,0}, \rho_{0,1}\}$ could be automatically learned based on the EM framework during the recovery process [30], as detailed in Appendix A. The initial distribution $p\left(s_{1,n}\right), \forall n$ is set to be the steady state distribution of the Markov chain in (11), i.e., $\lambda \triangleq p\left(s_{1,n}\right) = \frac{\rho_{0,1}}{\rho_{0,1}+\rho_{1,0}}$, where λ reflects the sparsity of $\boldsymbol{H}$. This ensures that all elements of $s_{\tau,n}$ have the same distribution $p\left(s_{\tau,n}\right) = \lambda^{s_{\tau,n}}(1-\lambda)^{1-s_{\tau,n}}$.

In practice, the noise precision $\kappa_\tau = \sigma_\tau^{-2}$ is unknown and we model it as a Gamma hyperpiror $p\left(\kappa_\tau\right) = \Gamma\left(\kappa_\tau; a_{\kappa,\tau}, b_{\kappa,\tau}\right)$, where we set $a_{\kappa,\tau}, b_{\kappa,\tau} \to 0$ so as to obtain a broad hyperprior.

In summary, the above three-layer hierarchical Markov channel model is a probabilistic model with sufficient degrees of freedom in the model parameters (such as $\rho_{0,1}$, $\rho_{1,0}$), which can be adjusted to cover various channel statistics that may occur in practice. Moreover, the algorithm does not require knowledge of these parameters, which can be automatically learned based on the EM method. As a result, the three-layer hierarchical Markov model also provides more flexibility to model various realistic channels in practice. Fig. 5 shows

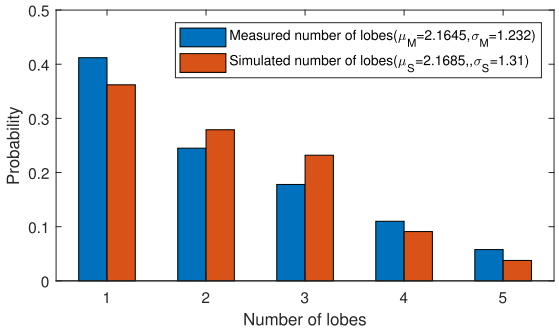


Fig. 5. Comparison of the measured channel property extracted from the 28-GHz mm-SSCM [31] and the simulated channel property extracted from the three-layer hierarchical Markov model. We set $M = N = 128$.

a typical empirical histogram plot of the number of AoA spatial lobes (SLs) extracted from the 28-GHz millimeter-wave statistical spatial channel model (mm-SSCM) proposed in [31], next to the simulated histograms extracted from the three-layer hierarchical Markov model. It can be seen that the three-layer hierarchical Markov channel prior yields good agreement with the practical mmWave channels.

B. Selective Channel Tracking Problem Formulation

Using the partial channel representation (3), the received signal at t-th frame $\boldsymbol{y}_t \in \mathbb{C}^{NN_p}$ can be rewritten as a CS model with an unknown $\boldsymbol{\varphi}_t$ in the measurement matrix as

$$\boldsymbol{y}_t = \boldsymbol{F}_t \boldsymbol{x}_t + \boldsymbol{n}_t, \tag{12}$$

where $\boldsymbol{F}_t = \left[\boldsymbol{F}_{t,i}^T\right]_{i\in\mathcal{N}_p}^T \in C^{NN_p\times\tilde{N}}$, $\boldsymbol{F}_{t,i} = \boldsymbol{A}_{R,i}(\boldsymbol{\varphi}_t)$, $\boldsymbol{n}_t = [\boldsymbol{n}_{t,i}]_{i\in\mathcal{N}_p}$.

In each frame t, the user needs to estimate the partial channel parameters $\boldsymbol{x}_t$ and $\boldsymbol{\varphi}_t$, given the observations up to t frame $\boldsymbol{y}_{1:t}$ in model (12) and $\hat{\boldsymbol{\varphi}}_{1:t-1} = \left\{\hat{\boldsymbol{\beta}}_{R,1:t-1}, \hat{f}_{d,1:t-1}, \hat{\eta}_{1:t-1}\right\}$ up to $(t-1)$-th frame. In particular, for given $\boldsymbol{\varphi}_t$, we are interested in computing minimum mean-squared error (MMSE) estimates of $x_{t,n}$, $\hat{x}_{t,n} = \mathrm{E}\left[x_{t,n}|\boldsymbol{y}_{1:t}; \hat{\boldsymbol{\varphi}}_{1:t-1}, \boldsymbol{\varphi}_t\right]$, where the expectation is over the marginal posterior:

$$p(x_{t,n}|\boldsymbol{y}_{1:t}; \hat{\boldsymbol{\varphi}}_{1:t-1}, \boldsymbol{\varphi}_t) \propto \int_{-x_{t,n}} p(\boldsymbol{y}_{1:t}, \boldsymbol{v}_t; \hat{\boldsymbol{\varphi}}_{1:t-1}, \boldsymbol{\varphi}_t), \tag{13}$$

where $\boldsymbol{v}_t = \{\boldsymbol{x}_t, \boldsymbol{s}_t, \boldsymbol{\gamma}_t, \kappa_t\}$, $-x_{t,n}$ denotes the vector collections integration over $\boldsymbol{v}_t$ except for the element $x_{t,n}$ and $\propto$ denotes equality after scaling.

On the other hand, the optimal $\boldsymbol{\varphi}_t$ at the t-th frame is obtained by ML as follows [24]:

$$\begin{aligned}\hat{\boldsymbol{\varphi}} &= \arg\max_{\boldsymbol{\varphi}_t} \ln p\left(\boldsymbol{y}_{1:t}; \hat{\boldsymbol{\varphi}}_{1:t-1}, \boldsymbol{\varphi}_t\right) \\ &= \arg\max_{\boldsymbol{\varphi}_t} \ln \int_{\boldsymbol{v}_t} p\left(\boldsymbol{y}_{1:t}, \boldsymbol{v}_t; \hat{\boldsymbol{\varphi}}_{1:t-1}, \boldsymbol{\varphi}_t\right) d\boldsymbol{v}_t. \end{aligned} \tag{14}$$

Once we obtain the ML estimate of $\hat{\boldsymbol{\varphi}}_t$, and the associated conditional marginal posterior $p\left(x_{t,n}|\boldsymbol{y}_{1:t}; \hat{\boldsymbol{\varphi}}_{1:t-1}, \boldsymbol{\varphi}_t\right)$, we can obtain the MMSE estimates of $\{x_{t,n}\}$.

One challenge in computing the MMSE estimate is the calculation of the exact posterior in (13) whose factor graph has loops. In the next section, we propose a DD-VBI algorithm to approximately calculate the marginal posteriors $p\left(x_{t,n}|\boldsymbol{y}_{1:t}; \hat{\boldsymbol{\varphi}}_{1:t-1}, \boldsymbol{\varphi}_t\right)$ by combining the message passing and VBI approaches, and use the in-exact majorization-minimization (MM) method (which is a generalization of the EM method) [24] to find an approximate solution for (14). The DD-VBI algorithm is shown in the simulations to achieve a good performance.

V. Doppler-Aware-Dynamic-VBI Algorithm

In (13) and (14), we need to estimate $\boldsymbol{\varphi}_t$ and $\boldsymbol{x}_t$ based on all the observations $\boldsymbol{y}_{1:t}$. One possible solution is to store all the available observations $\boldsymbol{y}_{1:t}$ and perform a joint estimation of $\boldsymbol{\varphi}_t$ and $\boldsymbol{x}_t$ at each frame. However, the memory cost and computational complexity of such a brute-force solution would become unacceptable for large t. To address this challenge, the proposed algorithm decomposes the joint distribution associated with the estimation problem (13) and (14) into two components, such that one component only involves the variables associated with the t-th frame, and the other component, which is the message passed from the previous frame, summarizes the available information provided by the observations $\boldsymbol{y}_{1:t-1}$, as detailed in Section V-E. Then, for given message from the previous frame and the observation $\boldsymbol{y}_t$, we propose the DD-VBI algorithm to calculate the MMSE estimates of $\{x_{t,n}\}$, as detailed in Section V-B.

A. Decomposition and Approximation of Joint Probability Distribution

This section is to decompose and approximate the joint probability distribution in (14) such that the joint probability distribution at the t-th frame only involves the probability density function (PDF) of the current hidden variables $\boldsymbol{v}_t$, $\boldsymbol{y}_t$, and the messages $\hat{p}\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1}, \hat{\boldsymbol{\varphi}}_{1:t-1}\right)$ passed from the previous frame, based on which a more efficient algorithm can be designed.

The joint probability distribution in (14) and (13) can be written as

$$\begin{aligned} &p\left(\boldsymbol{y}_{1:t}, \boldsymbol{v}_t; \hat{\boldsymbol{\varphi}}_{1:t-1}, \boldsymbol{\varphi}_t\right) \\ &\quad\propto \sum_{\boldsymbol{s}_{t-1}} p\left(\boldsymbol{s}_{t-1}|\boldsymbol{y}_{1:t-1}; \hat{\boldsymbol{\varphi}}_{1:t-1}\right) p\left(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}\right) \\ &p\left(\boldsymbol{y}_t|\boldsymbol{x}_t, \kappa_t; \boldsymbol{\varphi}_t\right) p\left(\boldsymbol{x}_t|\boldsymbol{\gamma}_t\right) p\left(\boldsymbol{\gamma}_t|\boldsymbol{s}_t\right) p\left(\kappa_t\right) \\ &\quad\approx \sum_{\boldsymbol{s}_{t-1}} q\left(\boldsymbol{s}_{t-1}|\boldsymbol{y}_{1:t-1}; \hat{\boldsymbol{\varphi}}_{1:t-1}\right) p\left(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}\right) \\ &p\left(\boldsymbol{y}_t|\boldsymbol{x}_t, \kappa_t; \boldsymbol{\varphi}_t\right) p\left(\boldsymbol{x}_t|\boldsymbol{\gamma}_t\right) p\left(\boldsymbol{\gamma}_t|\boldsymbol{s}_t\right) p\left(\kappa_t\right) \\ &\quad= \hat{p}\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1}; \hat{\boldsymbol{\varphi}}_{1:t-1}\right) p\left(\boldsymbol{y}_t|\boldsymbol{x}_t, \kappa_t; \boldsymbol{\varphi}_t\right) \\ &p\left(\boldsymbol{x}_t|\boldsymbol{\gamma}_t\right) p\left(\boldsymbol{\gamma}_t|\boldsymbol{s}_t\right) p\left(\kappa_t\right), \end{aligned}$$

where $\hat{p}\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1}; \hat{\boldsymbol{\varphi}}_{1:t-1}\right) = \sum_{\boldsymbol{s}_{t-1}} q\left(\boldsymbol{s}_{t-1}|\boldsymbol{y}_{1:t-1}; \hat{\boldsymbol{\varphi}}_{1:t-1}\right) p\left(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}\right)$ can be calculated based on the messages passed from the previous frame, $q\left(\boldsymbol{s}_{t-1}|\boldsymbol{y}_{1:t-1}; \hat{\boldsymbol{\varphi}}_{1:t-1}\right)$ is a tractable approximation for the posterior $p\left(\boldsymbol{s}_{t-1}|\boldsymbol{y}_{1:t-1}; \hat{\boldsymbol{\varphi}}_{1:t-1}\right)$ and $p\left(\boldsymbol{y}_t|\boldsymbol{x}_t, \kappa_t; \boldsymbol{\varphi}_t\right) = \mathcal{CN}\left(\boldsymbol{y}_t; \boldsymbol{F}_t\boldsymbol{x}_t, \kappa_t^{-1}\boldsymbol{I}\right)$. We will elaborate how to calculate $q\left(\boldsymbol{s}_{t-1}|\boldsymbol{y}_{1:t-1}; \hat{\boldsymbol{\varphi}}_{1:t-1}\right)$ and

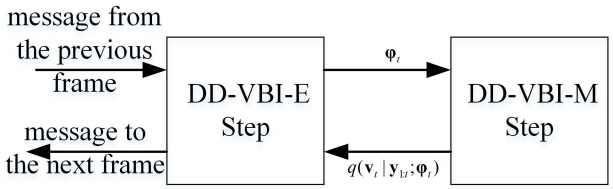


Fig. 6. Interaction between the two modules of the DD-VBI algorithm within a frame.

$\hat{p}\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1};\hat{\boldsymbol{\varphi}}_{1:t-1}\right)$ later in Sections V-B and V-E. When $t=1$, $\hat{p}\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1};\hat{\boldsymbol{\varphi}}_{1:t-1}\right)$ is reduced to $\hat{p}\left(\boldsymbol{s}_t|\boldsymbol{y}_1;\hat{\boldsymbol{\varphi}}_1\right)=p\left(\boldsymbol{s}_1\right)$.

For simplicity, we define

$$\begin{aligned}&\hat{p}\left(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\hat{\boldsymbol{\varphi}}_{1:t-1},\boldsymbol{\varphi}_t\right)=\hat{p}\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1};\hat{\boldsymbol{\varphi}}_{1:t-1}\right)\\&p\left(\boldsymbol{y}_t|\boldsymbol{x}_t,\kappa_t;\boldsymbol{\varphi}_t\right)p\left(\boldsymbol{x}_t|\boldsymbol{\gamma}_t\right)p\left(\gamma_t|\boldsymbol{s}_t\right)p\left(\kappa_t\right).\end{aligned}\tag{15}$$

In the rest of this section, we will omit $\hat{\boldsymbol{\varphi}}_{1:t-1}$ in the PDFs when there is no ambiguity.

B. Doppler-Aware-Dynamic-VBI Algorithm for Frame t

The basic idea of the DD-VBI algorithm is to simultaneously approximate the marginal posterior $\{p\left(x_{t,n}|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)\}$ and maximize the log-likelihood $\ln p\left(\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ with respect to $\boldsymbol{\varphi}_t$, based on the noisy measurements $\boldsymbol{y}_t$ of the t-th frame and the messages $\hat{p}\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1}\right)$ passed from the previous frame. In summary, at the t-th frame, the DD-VBI algorithm performs iterations between the following two major steps until convergence, as shown in Fig. 6.

- **DD-VBI-E Step:** Given $\boldsymbol{\varphi}_t$ at the t-th frame and messages $\hat{p}\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1}\right)$ from the $(t-1)$-th frame, E Step calculates the approximate posterior $p\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)\approx q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ by combining the message passing and VBI approaches, as elaborated in Section V-D.
- **DD-VBI-M Step:** Given $q(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t)\approx p(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t)$,E Step constructs a surrogate function (lower bound) for the objective function $\ln p\left(\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$, then maximize the surrogate function with respect to $\boldsymbol{\varphi}_t$, as elaborated in Section V-C.

After convergence, the messages $\hat{p}\left(\boldsymbol{s}_{t+1}|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ are calculated based on $q\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ and passed to the next frame. In the following, we first elaborate the M step, which is a variation of the in-exact MM method in [24]. After that, we will elaborate how to approximately calculate the posterior $p\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)\approx q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ in the E step, which is required to construct the surrogate function in the M step.

C. DD-VBI-M Step

It is difficult to directly maximize the log-likelihood function $\ln p\left(\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$, because there is no closed-form expression due to the multi-dimensional integration over $\boldsymbol{v}_t$ as in (14). To make the problem tractable, in the DD-VBI-M Step, we adopt an in-exact MM method in [24], [32], which maximizes a surrogate function of $\ln p\left(\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ with respect to $\boldsymbol{\varphi}_t$ to find an approximate solution of (14). Specifically, let $u\left(\boldsymbol{\varphi}_t;\dot{\boldsymbol{\varphi}}_t\right)$ be the surrogate function constructed at some fixed point $\dot{\boldsymbol{\varphi}}_t$, which satisfies the following properties:

$$\begin{aligned}u\left(\boldsymbol{\varphi}_t;\dot{\boldsymbol{\varphi}}_t\right)&\le\ln p\left(\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right),\\u\left(\dot{\boldsymbol{\varphi}}_t;\dot{\boldsymbol{\varphi}}_t\right)&=\ln p\left(\boldsymbol{y}_{1:t};\dot{\boldsymbol{\varphi}}_t\right),\\\frac{\partial u\left(\boldsymbol{\varphi}_t;\dot{\boldsymbol{\varphi}}_t\right)}{\partial\boldsymbol{\varphi}_t}|_{\boldsymbol{\varphi}_t=\dot{\boldsymbol{\varphi}}_t}&=\frac{\partial\ln p\left(\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)}{\partial\boldsymbol{\varphi}_t}|_{\boldsymbol{\varphi}_t=\dot{\boldsymbol{\varphi}}_t}.\end{aligned}\tag{16}$$

Inspired by the EM algorithm [32], we use the following surrogate function:

$$u\left(\boldsymbol{\varphi}_t;\dot{\boldsymbol{\varphi}}_t\right)=\int q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)\ln\frac{p\left(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\boldsymbol{\varphi}_t\right)}{q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)}d\boldsymbol{v}_t,\tag{17}$$

where $q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ is a tractable approximation of $p\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$, which is calculated in the E-step. In Section V-A, we have approximated the joint probability distribution in (17) using $\hat{p}\left(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\boldsymbol{\varphi}_t\right)$ in (15). Therefore, the surrogate function can be approximated as

$$\hat{u}\left(\boldsymbol{\varphi}_t;\dot{\boldsymbol{\varphi}}_t\right)=\int q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\dot{\boldsymbol{\varphi}}_t\right)\ln\frac{\hat{p}\left(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\boldsymbol{\varphi}_t\right)}{q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\dot{\boldsymbol{\varphi}}_t\right)}d\boldsymbol{v}_t,\tag{18}$$

When there is no approximation error for the associated PDFs, i.e., $q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)=p\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ and $\hat{p}\left(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\boldsymbol{\varphi}_t\right)=p\left(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\boldsymbol{\varphi}_t\right)$, it can be verified $\hat{u}\left(\boldsymbol{\varphi}_t;\dot{\boldsymbol{\varphi}}_t\right)$ satisfies the properties in (16).

In the M step of the j-th iteration, we update $\boldsymbol{\varphi}_t$ as

$$\boldsymbol{\varphi}_t^{j+1}=\arg\max_{\boldsymbol{\varphi}_t}\hat{u}\left(\boldsymbol{\varphi}_t;\boldsymbol{\varphi}_t^j\right),\tag{19}$$

where $(\cdot)^j$ stands for the j-th iteration. In our problem, $\hat{u}\left(\boldsymbol{\varphi}_t;\dot{\boldsymbol{\varphi}}_t\right)$ is a non-convex function and it is difficult to find its optimal solution. We use a gradient update as in [24], i.e.

$$\boldsymbol{\varphi}_t^{j+1}=\boldsymbol{\varphi}_t^j+\boldsymbol{\tau}^j\frac{\partial\hat{u}\left(\boldsymbol{\varphi}_t;\boldsymbol{\varphi}_t^j\right)}{\partial\boldsymbol{\varphi}_t},\tag{20}$$

where $\boldsymbol{\tau}^j$ is the step sizes determined by the Armijo rule [24]. The initial value of $\boldsymbol{\varphi}_t$ is set based on the available prior knowledge. In particular, the initial value of f_d is calculated according to the phase rotation of the received signals between the two adjacent pilot symbols. Considering the true AoAs are around their nearest grid points, the initial value of $\boldsymbol{\beta}_{R,t}$ is set as zero vector. When $t=1$, the rotation angle η_t is set to be zero; otherwise, it is set to be the estimated rotation angle in the $(t-1)$-th frame.

The approximate posterior $q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ has a factorized form as

$$\begin{aligned}&q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)\\&\quad=q\left(\boldsymbol{x}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)q\left(\boldsymbol{\gamma}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)q\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right).\end{aligned}\tag{21}$$

Therefore, after the convergence of the DD-VBI algorithm, we not only obtain an approximate stationary solution $\hat{\boldsymbol{\varphi}}_t$ of (14), but also the associated marginal conditional posterior $q\left(\boldsymbol{x}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)\approx p\left(\boldsymbol{x}_t|\boldsymbol{y}_{1:t},\boldsymbol{\varphi}_t\right)$.

D. DD-VBI-E Step

DD-VBI-E Step performs the sparse VBI to approximate the conditional marginal posteriors $p(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t)$ based on the following joint prior distribution:

$$\hat{p}\left(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\boldsymbol{\varphi}_t\right)=p\left(\boldsymbol{x}_t|\boldsymbol{\gamma}_t\right)p\left(\boldsymbol{\gamma}_t|\boldsymbol{s}_t\right)p\left(\kappa_t\right)\hat{p}\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1}\right)p\left(\boldsymbol{y}_t|\boldsymbol{x}_t,\kappa_t;\boldsymbol{\varphi}_t\right) \quad (22)$$

The corresponding approximate posterior distributions $q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)\approx p\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ will be given by (25)-(31).

1) Outline of Sparse VBI Within a Frame: For convenience, we use $\boldsymbol{v}_{t,n}$ to denote an individual variable in $\boldsymbol{v}_t$. Let $\mathcal{H}=\{n|\forall\boldsymbol{v}_{t,n}\in\boldsymbol{v}_t\}$. Moreover, we use $q\left(\boldsymbol{v}_t\right)$ as a simplified notation for $q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ when there is no ambiguity. The approximate conditional marginal posterior $q\left(\boldsymbol{v}_t\right)$ could be calculated by minimizing the Kullback-Leibler divergence (KLD) between $p\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ and $q\left(\boldsymbol{v}_t\right)$ subject to a factorized form constraint on $q\left(\boldsymbol{v}_t\right)$ as

$$\begin{aligned}&\mathscr{A}_{\mathrm{VBI}}:\ q^*\left(\boldsymbol{v}_t\right)\\&=\arg\min_{q(\boldsymbol{v}_t)}\int q\left(\boldsymbol{v}_t\right)\ln\frac{q\left(\boldsymbol{v}_t\right)}{p\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)}d\boldsymbol{v}_t\\&\quad\text{s.t. }q\left(\boldsymbol{v}_t\right)=\prod_{n\in\mathcal{H}}q\left(\boldsymbol{v}_{t,n}\right),\quad\int q\left(\boldsymbol{v}_{t,n}\right)d\boldsymbol{v}_{t,n}=1,\quad\forall n\in\mathcal{H}.\end{aligned}$$

Problem $\mathscr{A}_{\mathrm{VBI}}$ is non-convex, we aim at finding a stationary solution (denoted by $q^*\left(\boldsymbol{v}_t\right)$) of $\mathscr{A}_{\mathrm{VBI}}$, as defined below.

Definition 1 (Stationary Solution): $q^*(\boldsymbol{v}_t)=\prod_{n\in\mathcal{H}}q^*\left(\boldsymbol{v}_{t,n}\right)$ is called a stationary solution of Problem $\mathscr{A}_{\mathrm{VBI}}$ if it satisfies all the constraints in $\mathscr{A}_{\mathrm{VBI}}$ and $\forall n\in\mathcal{H}$,

$$\begin{aligned}&q^*\left(\boldsymbol{v}_{t,n}\right)\\&=\arg\min_{q(\boldsymbol{v}_{t,n})}\int\prod_{l\neq n}q^*(\boldsymbol{v}_{t,l})q(\boldsymbol{v}_{t,n})\ln\frac{\prod_{l\neq n}q^*(\boldsymbol{v}_{t,l})q(\boldsymbol{v}_{t,n})}{p(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t)}d\boldsymbol{v}_t.\end{aligned}$$

By finding a stationary solution $q^*\left(\boldsymbol{v}_t\right)$ of $\mathscr{A}_{\mathrm{VBI}}$, we could obtain the approximate posterior $q^*\left(\boldsymbol{v}_{t,n}\right)\approx p\left(\boldsymbol{v}_{t,n}|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right),\forall n\in\mathcal{H}$.

A stationary solution of $\mathscr{A}_{\mathrm{VBI}}$ can be obtained via alternately optimizing each individual density $q\left(\boldsymbol{v}_{t,n}\right),n\in\mathcal{H}$. For given $q\left(\boldsymbol{v}_{t,l}\right),\forall l\neq n$, the optimal $q\left(\boldsymbol{v}_{t,n}\right)$ that minimizes the KLD in $\mathscr{A}_{\mathrm{VBI}}$ is given by

$$q\left(\boldsymbol{v}_{t,n}\right)\propto\exp\left(\left\langle\ln p(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\boldsymbol{\varphi}_t)\right\rangle_{\prod_{l\neq n}q(\boldsymbol{v}_{t,l})}\right), \quad (23)$$

where $\left\langle f\left(x\right)\right\rangle_{q(x)}=\int f\left(x\right)q(x)dx$. However, $p(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\boldsymbol{\varphi}_t)$ is intractable. Since $p\left(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\boldsymbol{\varphi}_t\right)\approx\hat{p}\left(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\boldsymbol{\varphi}_t\right)$ in (15), (23) can be approximated as

$$q\left(\boldsymbol{v}_{t,n}\right)\propto\exp\left(\left\langle\ln\hat{p}(\boldsymbol{y}_{1:t},\boldsymbol{v}_t;\boldsymbol{\varphi}_t)\right\rangle_{\prod_{l\neq n}q(\boldsymbol{v}_{t,l})}\right). \quad (24)$$

Based on (24), the update equations of all variables are given in the subsequent subsections. The detailed derivation can be found in Appendix B. Note that the operator $\left\langle\cdot\right\rangle_{\boldsymbol{v}_{t,l}}$ is equivalent to $\left\langle\cdot\right\rangle_{q(\boldsymbol{v}_{t,l})}$ and expectation $\left\langle f\left(\boldsymbol{v}_{t,l}\right)\right\rangle_{q(\boldsymbol{v}_{t,l})}$ w.r.t. its own posterior is simplified as $\left\langle f\left(\boldsymbol{v}_{t,l}\right)\right\rangle$.

2) Initialization of Sparse VBI: In order to trigger the alternating optimization (AO) algorithm, we use the following initializations for the distribution functions $q\left(\boldsymbol{x}_t\right)$, $q\left(\boldsymbol{\gamma}_t\right)$ in the first iteration of t-th frame and $q\left(\boldsymbol{s}_1\right)$ in the first iteration of first frame. In the rest iterations, we initialize $q\left(\boldsymbol{x}_t\right)$, $q\left(\boldsymbol{s}_t\right)$, $q\left(\boldsymbol{\gamma}_t\right)$ to the (approximate) posterior calculated in the previous iteration.

Initialize $q\left(\boldsymbol{s}_1\right)=\hat{p}\left(\boldsymbol{s}_1|\boldsymbol{y}_1;\boldsymbol{\varphi}_t\right)=\prod_{n=1}^{\tilde{N}}q\left(s_{1,n}\right)$ with $q\left(s_{1,n}\right)=\left(\tilde{\pi}_{1,n}\right)^{s_{1,n}}\left(1-\tilde{\pi}_{1,n}\right)^{1-s_{1,n}}$. For given $\hat{p}\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)=\prod_{n=1}^{\tilde{N}}\left(\tilde{\pi}_{t,n}\right)^{s_{t,n}}\left(1-\tilde{\pi}_{t,n}\right)^{1-s_{t,n}}$, initialize a gamma distribution for $\boldsymbol{\gamma}_t$: $q\left(\boldsymbol{\gamma}_t\right)=\prod_{n=1}^{\tilde{N}}\Gamma\left(\gamma_{t,n};\tilde{a}_{\gamma,t,n},\tilde{b}_{\gamma,t,n}\right)$, where $\tilde{a}_{\gamma,t,n}=\tilde{\pi}_{t,n}a_t+\left(1-\tilde{\pi}_{t,n}\right)\overline{a}_t$, $\tilde{b}_{\gamma,t,n}=\tilde{\pi}_{t,n}b_t+\left(1-\tilde{\pi}_{t,n}\right)\overline{b}_t$. Initialize a Gaussian distribution for $\boldsymbol{x}_t$: $q\left(\boldsymbol{x}_t\right)=\mathcal{CN}\left(\boldsymbol{x}_t;\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t\right)$, where $\boldsymbol{\Sigma}_t=\left(\mathrm{diag}\left(\left\langle\boldsymbol{\gamma}_t\right\rangle\right)+\left(\boldsymbol{F}_t\right)^H\boldsymbol{F}_t\right)^{-1}$, $\boldsymbol{\mu}_t=\boldsymbol{\Sigma}_t\left(\boldsymbol{F}_t\right)^H\boldsymbol{y}_t$.

3) Update for $q\left(\kappa_t\right)$: From (24), $q\left(\kappa_t\right)$ can be derived as

$$q\left(\kappa_t\right)=\Gamma(\kappa_t;\tilde{a}_{\kappa,t},\tilde{b}_{\kappa,t}). \quad (25)$$

where $\tilde{a}_{\kappa,t}=a_\kappa+NN_p$, $\tilde{b}_{\kappa,t}=b_\kappa+\left\langle\left\|\boldsymbol{y}_t-\boldsymbol{F}_t\boldsymbol{x}_t\right\|^2\right\rangle_{\boldsymbol{x}_t}=b_\kappa+\left\|\boldsymbol{y}_t-\boldsymbol{F}_t\boldsymbol{\mu}_t\right\|^2+\mathrm{tr}\left(\boldsymbol{F}_t\boldsymbol{\Sigma}_t\left(\boldsymbol{F}_t\right)^H\right)$.

4) Update for $q(\boldsymbol{x}_t)$: $q\left(\boldsymbol{x}_t\right)$ can be derived as

$$q\left(\boldsymbol{x}_t\right)=\mathcal{CN}\left(\boldsymbol{x}_t;\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t\right). \quad (26)$$

$\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ can be calculated through

$$\begin{aligned}\boldsymbol{\Sigma}_t&=\left(\mathrm{diag}\left(\left\langle\boldsymbol{\gamma}_t\right\rangle\right)+\left\langle\kappa_t\right\rangle\left(\boldsymbol{F}_t\right)^H\boldsymbol{F}_t\right)^{-1} \quad (27)\\&=\mathbf{R}-\upsilon\mathbf{R}\left(\boldsymbol{F}_t\right)^H\left(\mathbf{I}+\upsilon\boldsymbol{F}_t\mathbf{R}\left(\boldsymbol{F}_t\right)^H\right)^{-1}\boldsymbol{F}_t\mathbf{R},\\\boldsymbol{\mu}_t&=\left\langle\kappa_t\right\rangle\boldsymbol{\Sigma}_t\left(\boldsymbol{F}_t\right)^H\boldsymbol{y}_t. \quad (28)\end{aligned}$$

where $\left\langle\boldsymbol{\gamma}_t\right\rangle=\frac{\tilde{a}_{\gamma,t,n}}{\tilde{b}_{\gamma,t,n}}$, $\left\langle\kappa_t\right\rangle=\frac{\tilde{a}_{\kappa,t}}{\tilde{b}_{\kappa,t}}$, $\mathbf{R}=\mathrm{diag}\left(\left[\frac{\tilde{b}_{\gamma,t,1}}{\tilde{a}_{\gamma,t,1}},\cdots,\frac{\tilde{b}_{\gamma,t,\tilde{N}}}{\tilde{a}_{\gamma,t,\tilde{N}}}\right]\right)$, $\upsilon=\left\langle\kappa_t\right\rangle=\frac{\tilde{a}_{\kappa,t}}{\tilde{b}_{\kappa,t}}$.

5) Update for $q\left(\boldsymbol{\gamma}_t\right)$: $q\left(\boldsymbol{\gamma}_t\right)$ can be derived as

$$q\left(\boldsymbol{\gamma}_t\right)=\prod_{n=1}^{\tilde{N}}\Gamma\left(\gamma_{t,n};\tilde{a}_{\gamma,t,n},\tilde{b}_{\gamma,t,n}\right), \quad (29)$$

where $\tilde{a}_{\gamma,t,n},\tilde{b}_{\gamma,t,n}$ are given by:

$$\begin{aligned}\tilde{a}_{\gamma,t,n}&=\left\langle s_{t,n}\right\rangle a_t+\left\langle 1-s_{t,n}\right\rangle\overline{a}_t+1,\\\tilde{b}_{\gamma,t,n}&=\left\langle s_{t,n}\right\rangle b_t+\left\langle 1-s_{t,n}\right\rangle\overline{b}_t+\left\langle\left|x_{t,n}\right|^2\right\rangle. \quad (30)\end{aligned}$$

where $\left\langle s_{t,n}\right\rangle=\tilde{\pi}_{t,n},\left\langle 1-s_{t,n}\right\rangle=1-\tilde{\pi}_{t,q}$, $\left\langle\left|x_{t,n}\right|^2\right\rangle=\left|\mu_{t,n}\right|^2+\Sigma_{t,n}$, $\mu_{t,n}$ is the n-th element of $\boldsymbol{\mu}_t$, $\Sigma_{t,n}$ is the n-th diagonal element of $\boldsymbol{\Sigma}_t$.

6) Update for $q\left(\boldsymbol{s}_t\right)$: $q\left(\boldsymbol{s}_t\right)$ can be derived as

$$q\left(\boldsymbol{s}_t\right)=\prod_{n=1}^{\tilde{N}}\left(\pi_{t,n}\right)^{s_{t,n}}\left(1-\pi_{t,n}\right)^{1-s_{t,n}}, \quad (31)$$

where $\pi_{t,n}=\frac{1}{C}\frac{\tilde{\pi}_{t,n}b_t^{a_t}}{\Gamma(a_t)}e^{(a_t-1)\left\langle\ln\gamma_{t,n}\right\rangle-b_t\left\langle\gamma_{t,n}\right\rangle}$, $\left\langle\ln\gamma_{t,n}\right\rangle=\psi\left(\tilde{a}_{\gamma,t,n}\right)$, $C=\frac{\tilde{\pi}_{t,n}b_t^{a_t}}{\Gamma(a_t)}e^{(a_t-1)\left\langle\ln\gamma_{t,n}\right\rangle-b_t\left\langle\gamma_{t,n}\right\rangle}+\frac{\left(1-\tilde{\pi}_{t,n}\right)\overline{b}_t^{\overline{a}_t}}{\Gamma(\overline{a}_t)}e^{(\overline{a}_t-1)\left\langle\ln\gamma_{t,n}\right\rangle-\overline{b}_t\left\langle\gamma_{t,n}\right\rangle}$ is the normalization constant, and $\psi\left(x\right)=\frac{d}{dx}\ln\left(\Gamma\left(x\right)\right)$ is the digamma function, defined as the logarithmic derivative of the gamma function.

TABLE II
COMPLEXITY ORDERS/SIGNALING OVERHEADS FOR DIFFERENT ALGORITHMS

Algorithms	Complexity order	Typical complexity order	Signaling overhead	Typical signaling overhead
Proposed	$\mathcal{O}\left(\left(N_b N_p\right)^2 \tilde{N}\right)$	$\mathcal{O}\left(L_D^2 N\right)$	$\mathcal{O}\left(\frac{L_D \ln(N/L_D)}{N_b}\right)+\frac{MN_d}{M_b}$	$\mathcal{O}\left(\frac{L_D \ln(N/L_D)}{N_b}\right)+\frac{MN_d}{M_b}$
ML-partial	$\mathcal{O}\left(\tilde{N}^3\right)$	$\mathcal{O}\left(N^3\right)$	N^2/N_b	N^2/N_b
HS	$\mathcal{O}\left(SL_D^4K_H^2\right)$	$\mathcal{O}\left(\log_{K_H}\left(N/L_D\right)L_D^4K_H^2\right)$	$K_H^2\left(L_D\right)^3S$	$K_H^2\left(L_D\right)^3\log_{K_H}\left(N/L_D\right)$
BT	$\mathcal{O}\left(L_DMN\right)$	$\mathcal{O}\left(L_DMN\right)$	$4L_D+\mathcal{O}\left(L_D+2\right)$	$6L_D+4$

E. Messages $\hat{p}\left(\boldsymbol{s}_{t+1}|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ Passed to the Next Frame

After the convergence of the DD-VBI iterations in frame t, let $\hat{f}_{d,t}$ and $q\left(\boldsymbol{s}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)=\prod_{n=1}^{\tilde{N}}\left(\pi_{t,n}\right)^{s_{t.n}}\left(1-\pi_{t,n}\right)^{1-s_{t,n}}$ denote the converged Doppler parameter and the associated approximate posterior for $\boldsymbol{s}_t$, respectively. Then, the messages $\hat{p}\left(\boldsymbol{s}_{t+1}|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ passed to the next frame are calculated as

$$\begin{aligned}\hat{p}\left(\boldsymbol{s}_{t+1}|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)&=\prod_{n=1}^{\tilde{N}}\sum_{\mathbf{s}_{t,n}}q\left(s_{t,n}|\boldsymbol{y}_{1:t},\boldsymbol{\varphi}_t\right)p\left(s_{t+1,n}|s_{t,n}\right)\\&=\prod_{n=1}^{\tilde{N}}\left(\tilde{\pi}_{t+1,n}\right)^{s_{t+1,n}}\left(1-\tilde{\pi}_{t+1,n}\right)^{1-s_{t+1,n}}.\end{aligned}\tag{32}$$

where $\tilde{\pi}_{t+1,n}=\left(1-\pi_{t,n}\right)\rho_{0,1}+\pi_{t,n}\left(1-\rho_{1,0}\right)$.

Finally, messages $\hat{p}\left(\boldsymbol{s}_{t+1}|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$, as the prior to channel support, are passed to the next frame. The overall DD-VBI algorithm is summarized in Algorithm 1. '

Algorithm 1 Doppler-Aware-Dynamic-VBI Algorithm

1: **for** $t=1,2,\ldots$ **do**
2: Initialize the distribution according to Section V-D.2.
3: **while** not converge **do**
4: **while** not converge **do**
5: **%DD-VBI-E Step:**
6: Update $q\left(\kappa_t\right),q\left(\boldsymbol{x}_t\right),q\left(\boldsymbol{\gamma}_t\right),q\left(\boldsymbol{s}_t\right)$ using (25), (26), (29) and (31).
7: **end while**
8: **%DD-VBI-M Step:**
9: Construct the surrogate function $\hat{u}$ in (18) using the output of DD-VBI-E Step $q\left(\boldsymbol{v}_t|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$.
10: Update $\boldsymbol{\varphi}_t$ using (19).
11: **end while**
12: Calculate $\hat{p}\left(\boldsymbol{s}_{t+1}|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ using (32). Pass messages $\hat{p}\left(\boldsymbol{s}_{t+1}|\boldsymbol{y}_{1:t};\boldsymbol{\varphi}_t\right)$ to frame $t+1$.
13: Estimate $x_{t,n}$ using (28).
14: **end for**

Remark 2: In practical mmWave massive MIMO systems, the number of RF chains can be less than the number of antennas at both the BS and user sides to reduce the hardware cost and power consumption. The proposed scheme can be easily extended to the case with limited RF chains. For example, suppose there are only $M_b<M$ RF chains at the BS and $N_b<N$ RF chains at the user. In this case, when the BS transmits the training vector $\mathbf{v}=\mathbf{F}\mathbf{g}\in\mathbb{C}^M$ for downlink channel tracking in the i-th symbol duration, the user employs $\mathbf{U}_i=\mathbf{W}_i\mathbf{G}_i\in\mathbb{C}^{N\times N_b}$ as a combining matrix to combine the received signal into N_b baseband channel measurements, where $\mathbf{F}\in\mathbb{C}^{M\times M_b}$ and $\mathbf{g}\in\mathbb{C}^{M_b}$ are the RF training matrix and baseband training vector at the BS, respectively, and $\mathbf{W}_i\in\mathbb{C}^{N\times N_b}$ and $\mathbf{G}_i\in\mathbb{C}^{N_b\times N_b}$ are the RF and baseband combining matrix at the mobile user in the i-th symbol duration. We can still write the received pilot signal as a CS model as in (12), but with $\boldsymbol{F}_t=\left[\left(\boldsymbol{U}_{t,i}^H\boldsymbol{A}_{R,i}(\boldsymbol{\varphi}_t)\right)\right]_{i\in\mathcal{N}_p}$, $\boldsymbol{n}_t=\left[\boldsymbol{U}_{t,i}^H\boldsymbol{n}_{t,i}\right]_{i\in\mathcal{N}_p}$.

F. Complexity and Signaling Overhead Comparison

The computational complexity of the proposed algorithm is dominated by the update of $q\left(\boldsymbol{x}_t\right)$. Assuming the arithmetic with individual elements has complexity $\mathcal{O}\left(1\right)$, the computational complexity of matrix inversion in (27) is $\mathcal{O}\left(N_b^3N_p^3\right)$ and the total number of multiplications to update $q(\boldsymbol{x}_t)$ is $3\tilde{N}+2N_bN_p\tilde{N}^2+2\left(N_bN_p\right)^2\tilde{N}+\left(N_bN_p\right)^2$. As such, the complexity order of the proposed method is $\mathcal{O}\left(\left(N_bN_p\right)^2\tilde{N}\right)$, considering that $\tilde{N}$ is usually larger than N_bN_p.

The signaling overhead is defined as the number of the downlink training vectors required to achieve the stable estimation of the partial channel parameters plus the uplink training vectors required to achieve the stable estimation of the effective uplink channel, where stable estimation means that the estimation error tends to zero as the noise power tends to zero [33]. Suppose the mmWave channel has L_D dominant paths, the number of required measurements N_bN_p to accurately recover the downlink partial channel parameters is proportional to $\mathcal{O}\left(L_D\ln\left(N/L_D\right)\right)$. MN_d/M_b orthogonal pilots are used to estimate the slow time-varying effective channel by LS channel estimation scheme in the uplink training stage.In Table II, we compare the complexity and signaling overhead of the proposed algorithm with the HS algorithm in [17] and the following baseline algorithms:

- **Baseline 1** (ML) [7]: This is the ML based joint DFO and channel estimation algorithm proposed in [7]. We also consider a modified version of ML algorithm in which only partial channel coefficients are estimated as in this paper. The modified ML will be called ML-partial. Since ML-partial has both lower signaling overhead and higher performance, we will focus on the comparison with ML-partial.
- **Baseline 2** (BT) [19]: This is an improved version of the beam tracking (BT) algorithm proposed in [19]. Since Doppler estimation is not considered in the original BT,

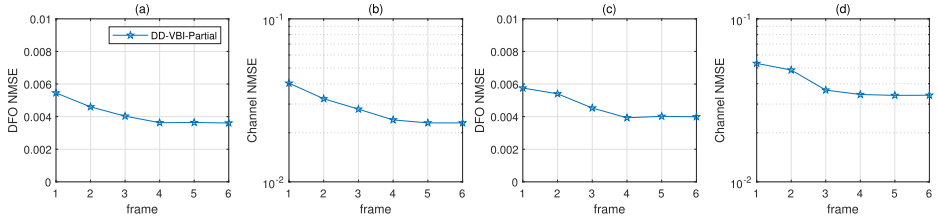


Fig. 7. The DFO NMSE and channel NMSE for different frame. SNR $= 0$ dB; (a) and (b): Full RF chains with $N = N_b = 128$; (c) and (d): Limited RF chains with $N = 128$, $N_b = 32$.

we assume perfect knowledge of Doppler parameter is available for BT. Moreover, the initial beam in the first frame is obtained by assuming perfect knowledge about channel directions.

In Table II, S is the total level of hierarchical codebook in HS [17], and K_H denotes the number of beamforming vectors of each codebook level in HS. Note that $N_b N_p$ is chosen to be $\mathcal{O}(L_D)$ and we have $\tilde{N} = \mathcal{O}(N)$. For a resolution $\mathcal{O}(2\pi/N)$ of the HS algorithm, we usually have $\mathcal{S} = \mathcal{O}\left(\log_{K_H}(N/L_D)\right)$ [17]. In this typical case, the complexity of the proposed and the baseline algorithms are shown in the third column of Table II. The typical complexity order of the proposed algorithm is less than the ML baseline and similar to the HS and BT baselines in Table II. From the the simulations (see Fig. 11), we find that the proposed algorithm consumes less computational time to achieve the same performance compared to all baselines. On the other hand, the proposed algorithm has the lowest signaling overhead, as shown in Table II.

VI. Simulation Results

In this section, we compare the performance of the proposed algorithm with the baseline algorithms described in Section V-F. The channel parameters are based on the mm-SSCM as specified in [31], which was developed according to the 28- and 73-GHz ultrawideband propas_{t+1,n}|s_{t,n}gation measurements in New York City. The main parameters (i.e., number of AoD/AoA spatial lobes, the central AoD/AoA angle of each AoD/AoA spatial lobe and the angular spread) of the mm-SSCM are generated according to the statistics in Table II of [31]. The signal bandwidth is 50MHz and the frame duration is set as $T_b = 0.2$ms. The carrier frequency is 28GHz. The BS employs a ULA of $M = 128$ antennas and the inter antenna spacing is $\lambda/2$, and the user also employs a ULA. The number of the uniform grid $\tilde{N}$ is set $\tilde{N} = N$. The user velocity is assumed to be 380km/h, which translates to $f_{d,t} \approx 10$KHz. Besides the baselines considered in V-F, we added the following two baselines for performance comparison.

- **Baseline 3** (DD-VBI-Random): This is similar to the proposed algorithm, except that the training vector is generated randomly.
- **Baseline 4** (DD-VBI-Full): This is similar to the Baseline 3, except that the full channel matrix instead of the partial channel is estimated.

Note that we do not compare the performance with the HS algorithm because it cannot work well in the high-mobility scenario. In the simulations, we will consider both cases when the user is equipped with a full set of RF chains and limited RF chains. The DFO estimation normalized MSE (NMSE), the channel estimation NMSE and the achievable data rate are adopted as the performance metrics. The DFO NMSE is defined as $\frac{\left\|\hat{f}_{d,t}-f_{d,t}\right\|^2}{\|f_{d,t}\|^2}$ and the channel NMSE is defined as $\frac{\|\hat{\boldsymbol{x}}_t-\boldsymbol{x}_t\|^2}{\|\boldsymbol{x}_t\|^2}$. The achievable rate is defined as $R_u = \frac{1}{T_u}\sum_{i=1}^{T_u}\log_2\left|\boldsymbol{I}+\frac{\overline{P}_u(\boldsymbol{H}_i^s)^H\boldsymbol{H}_i^s}{N_d\sigma_u^2}\right|$,where T_u is the number of data symbols in the uplink subframe, $\overline{P}_u$ is chosen to satisfy the transmit power constraint $\overline{P}_u Tr\left(\mathbf{W}_i^d\left(\mathbf{W}_i^d\right)^H\right) = P_u$ with P_u denoting the transmit power in the uplink, and σ_u^2 is the per-antenna noise variance in the uplink. For convenience, the SNR P_u/σ_u^2 in the uplink is set to be the same as that in the downlink. Clearly, for different algorithms, the estimated channel parameters are different, and thus both $\boldsymbol{H}_i^s$ and the achievable rate in the uplink are different. When the estimated channel parameters are more accurate, the achievable rate is also larger. Note that the achievable rate in the downlink is similar due to the uplink-downlink duality and therefore is not shown for conciseness. The number of selected beams N_d is set to be the same for all algorithms for fair comparison. In particular, we set $N_d = 6$ unless specified otherwise, which covers most of the channel energy. In the first frame, the proposed algorithm still works except that there is no prior information about the channel support passed from the previous frame. In this case, we may slightly send more pilots to ensure a good channel estimation quality of the first frame. In particular, 8 pilots are send for the proposed scheme in the first frame.

A. Channel Tracking Performance

In Fig. 7, we show the channel and DFO NMSE over time. In the first frame, there is no prior information passed from the previous frame and the channel estimation performance is poor. As the number of frame increases, the channel estimation performance becomes better because we can exploit the time correlation of the channel supports (as captured by the message passed from the previous frame). Finally, the channel estimation performance will saturate after the channel tracking process enters into the steady state.

B. DFO NMSE Performance

The DFO NMSE performance of different algorithms versus the number of user antennas and SNR is shown in Fig. 8.

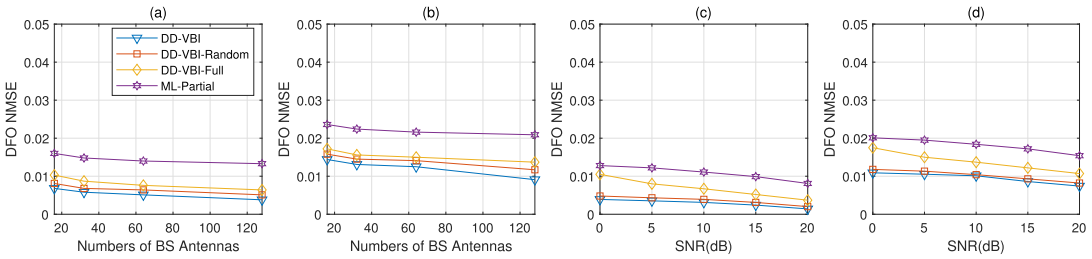


Fig. 8. DFO NMSE performance versus the number of user antennas and SNR. (a) and (b): SNR $= 0$ dB; (a): Full RF chains with $N_b = N$; (b): Limited RF chains with $N_b = N/4$. (c) and (d): $N = 128$; (c): Full RF chains with $N_b = 128$; (d): Limited RF chains with $N_b = 32$.

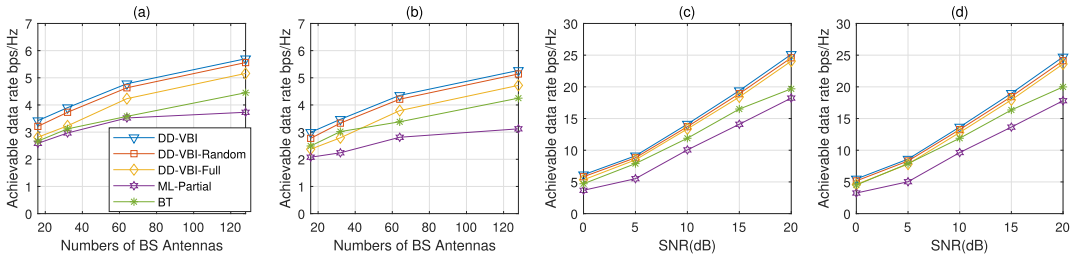


Fig. 9. Achievable data rate versus the number of user antennas and SNR. (a): SNR $= 0$ dB, full RF chains with $N_b = N$. (b): SNR= 0 dB, limited RF chains with $N_b = N/4$. (c): Full RF chains with $N = N_b = 128$. (d): Limited RF chains with $N = 128$, $N_b = 32$.

Note that the number of pilots consumed by different algorithms is different (as labeled in the simulation figures), and the proposed algorithm consumes much less pilot resource than all baselines. It can be seen that the proposed algorithm achieves large performance gain over all the baseline algorithms, under both full and limited RF chains. By using the proposed training vector design, the DD-VBI algorithm could further improve the NMSE performance compared to the case with a random training vector. This demonstrates that the proposed algorithm can effectively estimate the maximum DFO by selective channel tracking and efficient training vector design. Fig. 8 demonstrates that the proposed algorithm is a powerful method for accurate DFO estimation, even when both BS and user are equipped with massive MIMO, and the number of RF chains at the user side is limited.

C. Achievable Data Rate Performance

The achievable data rates of different algorithms versus the SNR and number of user antennas are shown in Fig. 9. Note that the BT baseline can only work with a full set of RF chains. It can be observed that the performance of all algorithms increases with SNR and number of user antennas. The proposed DD-VBI algorithm can achieve large performance gain over various baselines under both full and limited RF chains. Moreover, by using the proposed training vector design, the proposed DD-VBI algorithm could further improve the achievable data rate. Note that since the performance loss caused by the channel estimation error in the Baseline 3 (DD-VBI-Random) is much smaller than that in the Baseline 4 (DD-VBI-Full), a higher achievable rate can be achieved by Baseline 3. The performance of the beam tracking baseline is worse than the proposed algorithm. This is because after each beam initialization, it only searches for the channel directions around the previously estimated directions and cannot track the newly emerging paths in a real-time way. The results in Fig. 9 verify that the proposed selective channel tracking and Doppler compensation scheme can also enhance the achievable data rate with low pilot overhead, under different SNRs and numbers of user antennas.

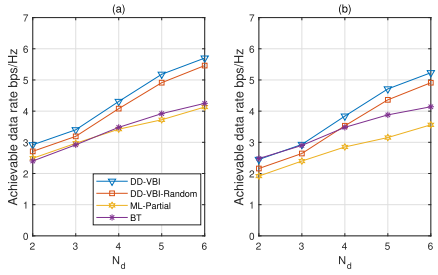


Fig. 10. Achievable data rate versus the number of selected beams N_d. (a): SNR $= 0$ dB; full RF chains with $N = N_b = 128$. (b): SNR $= 0$ dB; limited RF chains with $N = 128$, $N_b = 32$.

D. Impact of the Number of Selected Beams N_d

The achievable data rates of different algorithms versus the number of selected beams N_d are shown in Fig. 10. It can be observed that the performance of all algorithms increases with N_d. Again, the proposed DD-VBI algorithm can achieve large performance gain over various baselines under both the full and limited RF chains.

E. Complexity Versus Realized Gain

In practice, we can control the tradeoff between the complexity and realized gain of the proposed algorithm by adjusting the number of iterations. In Fig. 11, we plot the achievable rate versus the CPU time. It can be seen that the proposed algorithm can achieve a better performance than the baseline

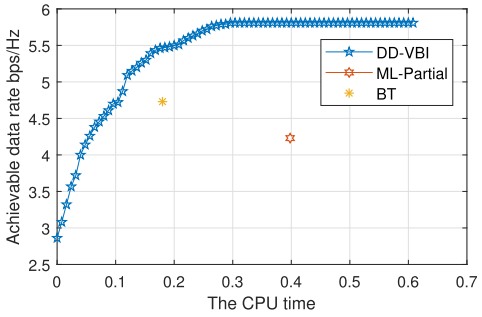


Fig. 11. The tradeoff between the complexity and realized gain of the proposed algorithm. SNR $= 0$ dB, full RF chains with $N_b = N = 128$.

algorithms for the same CPU time. Moreover, the performance gain increases with the CPU time. Therefore, the proposed algorithm provides a better and more flexible tradeoff between the performance and computational power.

VII. Conclusion

In this paper, we propose an angular-domain selective channel tracking and Doppler compensation scheme for high-mobility mmWave massive MIMO systems. Firstly, we propose a selective channel tracking scheme and the associated Doppler-aware-dynamic-VBI algorithm to accurately estimate the DFO and partial angular–domain channel parameters with reduced pilot overhead. Then, we propose an angular-domain selective DFO compensation scheme to convert the dominant paths of the fast time-varying channel into a slow time-varying effective channel, based on which efficient uplink and downlink transmissions can be achieved. Simulations verify that the proposed scheme not only can mitigate the Doppler and channel aging effect with much less pilots than existing schemes, but also can achieve a desired tradeoff between the CSI signaling overhead and spatial multiplexing/array gains.

Appendix

A. EM Estimation of the Parameters $\rho_{1,0}$ and $\rho_{0,1}$

We apply the EM method in [32] to obtain an estimate of the parameter $\rho_{1,0}$ by performing the following iterations until convergence:

$$\rho_{1,0}^{i+1} = \arg\max_{\rho_{1,0}} E_{\hat{p}^i(\boldsymbol{s}_{1:t})}\left\{\ln p(\boldsymbol{s}_{1:t}, \boldsymbol{y}_{1:t}; \rho_{1,0}) | \rho_{1,0}^i\right\}, \quad (33)$$

where $\rho_{1,0}^i$ stands for the i-th iteration, $E_{\hat{p}^i(\boldsymbol{s}_{1:t})}\{\cdot\}$ denotes expectation over the posterior distribution $\hat{p}^i(\boldsymbol{s}_{1:t}) = p(\boldsymbol{s}_{1:t}|\boldsymbol{y}_{1:t}, \rho_{1,0}^i)$ conditioned on $\boldsymbol{y}_{1:t}$ and $\rho_{1,0}^i$. Solving the problem (33), we can get a closed-form expression for updating the $\rho_{1,0}$ as

$$\rho_{1,0}^{i+1} = \frac{\sum_{\tilde{t}=2}^{t}\sum_{n=1}^{\tilde{N}} E_{\hat{p}^i(\boldsymbol{s}_{1:t})}\left[s_{\tilde{t},n}\right] - E_{\hat{p}^i(\boldsymbol{s}_{1:t})}\left[s_{\tilde{t}-1,n}s_{\tilde{t},n}\right]}{\sum_{\tilde{t}=2}^{t}\sum_{n=1}^{\tilde{N}} E_{\hat{p}^i(\boldsymbol{s}_{1:t})}\left[s_{\tilde{t}-1,n}\right]} \quad (34)$$

Since the channel support vector $\boldsymbol{s}_{1:t}$ forms a Markov chain, the posterior distribution $\hat{p}^i(\boldsymbol{s}_{1:t})$ can be approximately calculated using the sum-product message passing over the factor graph of the associated Markov chain with the priors of $s_{\tilde{t},n}$'s given by the output of the DD-VBI algorithm at each frame $\tilde{t}$ [30]. After the iterations in (34) convergence, we can obtain the EM estimation of the parameter $\rho_{1,0}$ in the t-th frame. The EM estimation of the parameter $\rho_{0,1}$ can be obtained in a similar way.

B. Derivation of (25)-(31)

From (24), $q(\kappa_t)$ in (25) can be obtained a

$$\begin{aligned}\ln q(\kappa_t) &\propto \langle \ln p(\boldsymbol{y}_t|\boldsymbol{x}_t, \kappa_t; \boldsymbol{\varphi}_t)\rangle_{\boldsymbol{x}_t} + \ln p(\kappa_t)\\ &\propto -\kappa_t \left\langle \|\boldsymbol{y}_t - \boldsymbol{F}_t\boldsymbol{x}_t\|^2 \right\rangle_{\boldsymbol{x}_t} + NN_p \ln \kappa_{t,i}\\ &\quad + (a_\kappa - 1)\ln\kappa_t - b_\kappa\kappa_t \propto (\tilde{a}_\kappa - 1)\ln\kappa - \tilde{b}_\kappa\kappa.\end{aligned}$$

$q(\boldsymbol{x}_t)$ in (26) can be obtained as

$$\begin{aligned}\ln q(\boldsymbol{x}_t) &\propto \langle \ln p(\boldsymbol{y}_t|\boldsymbol{x}_t, \kappa_t; \boldsymbol{\varphi}_t)\rangle_{\kappa_t} + \langle \ln p(\boldsymbol{x}_t|\boldsymbol{\gamma}_t)\rangle_{\boldsymbol{\gamma}_t}\\ &\propto -\langle\kappa_t\rangle \|\boldsymbol{y}_t - \boldsymbol{F}_t\boldsymbol{x}_t\|^2 - \boldsymbol{x}_t^H \mathrm{diag}(\langle\boldsymbol{\gamma}_t\rangle)\boldsymbol{x}_t\\ &\propto -(\boldsymbol{x}_t - \boldsymbol{\mu}_t)^H(\boldsymbol{\Sigma}_t)^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_t).\end{aligned}$$

$q(\boldsymbol{\gamma}_t)$ in (29) can be obtained as

$$\begin{aligned}&\ln q(\boldsymbol{\gamma}_t)\\ &\quad\propto \langle \ln p(\boldsymbol{x}_t|\boldsymbol{\gamma}_t)\rangle_{\boldsymbol{x}_t} + \langle \ln p(\boldsymbol{\gamma}_t|\boldsymbol{s}_t)\rangle_{\boldsymbol{s}_t}\\ &\quad\propto \sum_n (\tilde{a}_{\gamma,t,n} - 1)\ln\boldsymbol{\gamma}_{t,n} - \tilde{b}_{\gamma,t,n}\boldsymbol{\gamma}_{t,n}.\end{aligned}$$

$q(\boldsymbol{s}_t)$ in (31) can be obtained as

$$\begin{aligned}&\ln q(\boldsymbol{s}_t)\\ &\propto \langle \ln p(\boldsymbol{\gamma}_t|\boldsymbol{s}_t)\rangle_{\boldsymbol{\gamma}_t} + \ln\hat{p}(\boldsymbol{s}_t)\\ &\propto \sum_n s_t\left(\ln\frac{b_t^{a_t}}{\Gamma(a_t)} + (a_t - 1)\langle\ln\boldsymbol{\gamma}_{t,n}\rangle - b_t\langle\boldsymbol{\gamma}_{t,n}\rangle + \ln\tilde{\pi}_{t,n}\right)\\ &\quad + (1 - s_t)\left(\ln\frac{\bar{b}_t^{\bar{a}_t}}{\Gamma(\bar{a}_t)} + (\bar{a}_t - 1)\times\langle\ln\boldsymbol{\gamma}_{t,n}\rangle - \bar{b}_t\langle\boldsymbol{\gamma}_{t,n}\rangle\right.\\ &\quad \left. + \ln(1 - \tilde{\pi}_{t,n})\right) \propto \ln\prod_{n=1}^{\tilde{N}}(\pi_{t,n})^{s_{t.n}}(1 - \pi_{t,n})^{1 - s_{t,n}}.\end{aligned}$$

References

[1] M. Souden, S. Affes, J. Benesty, and R. Bahroun, "Robust Doppler spread estimation in the presence of a residual carrier frequency offset," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 4148–4153, Oct. 2009.

[2] C. R. Berger, Z. Wang, J. Huang, and S. Zhou, "Application of compressive sensing to sparse channel estimation," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 164–174, Nov. 2010.

[3] X. Wang, G. Wang, R. Fan, and B. Ai, "Channel estimation with expectation maximization and historical information based basis expansion model for wireless communication systems on high speed railways," *IEEE Access*, vol. 6, pp. 72–80, 2018.

[4] Y. Mostofi and D. C. Cox, "ICI mitigation for pilot-aided OFDM mobile systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 765–774, Mar. 2005.

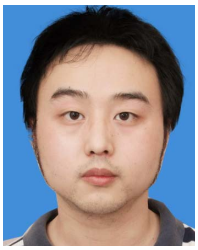
[5] T.-L. Liu, W.-H. Chung, S.-Y. Yuan, and S.-Y. Kuo, "On the banded approximation of the channel matrix for mobile OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 64, no. 8, pp. 3526–3535, Aug. 2015.

[6] T. Y. Al-Naffouri, K. M. Zahidul Islam, N. Al-Dhahir, and S. Lu, "A model reduction approach for OFDM channel estimation under high mobility conditions," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2181–2193, Apr. 2010.

[7] Y. Ge, W. Zhang, F. Gao, and H. Minn, "Angle-domain approach for parameter estimation in high-mobility OFDM with fully/partly calibrated massive ULA," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 591–607, Jan. 2019.

[8] A. Monk, R. Hadani, M. Tsatsanis, and S. Rakib, "OTFS–orthogonal time frequency space," 2016, *arXiv:1608.02993*. [Online]. Available: http://arxiv.org/abs/1608.02993
[9] M. Kollengode Ramachandran and A. Chockalingam, "MIMO-OTFS in high-Doppler fading channels: Signal detection and channel estimation," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 206–212.
[10] F. Bellili and S. Affes, "A low-cost and robust maximum likelihood Doppler spread estimator," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2013, pp. 4325–4330.
[11] D. Chizhik, "Slowing the time-fluctuating MIMO channel," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1554–1565, Sep. 2004.
[12] Y. Zhang, Q. Yin, P. Mu, and L. Bai, "Multiple Doppler shifts compensation and ICI elimination by beamforming in high-mobility OFDM systems," in *Proc. Int. ICST Conf. Commun. Netw. China (CHINACOM)*, 2011, pp. 170–175.
[13] W. Guo, P. Mu, Q. Yin, and H.-M. Wang, "Multiple Doppler frequency offsets compensation technique for high-mobility OFDM uplink," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Aug. 2013, pp. 1–5.
[14] X. Chen, J. Lu, T. Li, P. Fan, and K. Ben Letaief, "Directivity-beamwidth tradeoff of massive MIMO uplink beamforming for high speed train communication," *IEEE Access*, vol. 5, pp. 5936–5946, 2017.
[15] W. Guo, W. Zhang, P. Mu, F. Gao, and B. Yao, "Angle-domain Doppler pre-compensation for high-mobility OFDM uplink with massive ULA," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.
[16] J. Wang *et al.*, "Beam codebook based beamforming protocol for multi-gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, Oct. 2009.
[17] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[18] J. Palacios, D. De Donno, and J. Widmer, "Tracking mm-Wave channel dynamics: Fast beam training strategies under mobility," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2017, pp. 1–9.
[19] D. Zhu, J. Choi, Q. Cheng, W. Xiao, and R. W. Heath, Jr., "High-resolution angle tracking for mobile wideband millimeter-wave systems with antenna array calibration," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7173–7189, Nov. 2018.
[20] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1259–1262, Jun. 2016.
[21] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proc. IEEE*, vol. 98, no. 6, pp. 1058–1076, Jun. 2010.
[22] W. Guo, W. Zhang, P. Mu, F. Gao, and H. Lin, "High-mobility wideband massive MIMO communications: Doppler compensation, analysis and scaling laws," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3177–3191, Jun. 2019.
[23] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131–146, Jan. 2008.
[24] J. Dai, A. Liu, and V. K. N. Lau, "FDD massive MIMO channel estimation with arbitrary 2D-array geometry," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2584–2599, May 2018.
[25] C. B. Dietrich, Jr., "Adaptive arrays and diversity antenna configurations for handheld wireless communication terminals," Ph.D. dissertation, Virginia Polytech. Inst. State Univ., Blacksburg, VA, USA, 2000.
[26] J.-C. Shen, J. Zhang, E. Alsusa, and K. B. Letaief, "Compressed CSI acquisition in FDD massive MIMO: How much training is needed?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4145–4156, Jun. 2016.
[27] Z. Chen and C. Yang, "Pilot decontamination in wideband massive MIMO systems by exploiting channel sparsity," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5087–5100, Jul. 2016.
[28] X. Zhu, L. Dai, G. Gui, W. Dai, Z. Wang, and F. Adachi, "Structured matching pursuit for reconstruction of dynamic sparse channels," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–5.
[29] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, Jun. 2008.
[30] J. Ziniel and P. Schniter, "Dynamic compressive sensing of time-varying signals via approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5270–5284, Jul. 2013.
[31] M. K. Samimi and T. S. Rappaport, "3-D millimeter-wave statistical channel model for 5G wireless system design," *IEEE Trans. Microw. Theory Tech.*, vol. 64, no. 7, pp. 1–19, Jul. 2016.
[32] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, Jan. 1977.
[33] G. Wunder, S. Stefanatos, A. Flinth, I. Roth, and G. Caire, "Low-overhead hierarchically-sparse channel estimation for multiuser wideband massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2186–2199, Apr. 2019.

Guanying Liu (Student Member, IEEE) received the B.Eng. degree in information science and technology from Southwest Jiaotong University in 2017. She is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University. Her research interests include signal processing and compressive sensing.



An Liu (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Peking University, China, in 2004 and 2011, respectively. From 2008 to 2010, he was a Visiting Scholar with the Department of Electrical, Computer & Energy Engineering (ECEE), University of Colorado Boulder. He was a Post-Doctoral Research Fellow from 2011 to 2013, a Visiting Assistant Professor in 2014, and a Research Assistant Professor from 2015 to 2017, with the Department of Electronics and Communication Engineering (ECE), HKUST. He is currently a Distinguished Research Fellow with the College of Information Science and Electronic Engineering, Zhejiang University. His research interests include wireless communications, stochastic optimization, and compressive sensing.

Rui Zhang (Fellow, IEEE) received the B.Eng. (Hons.) and M.Eng. degrees from the National University of Singapore, Singapore, and the Ph.D. degree from Stanford University, Stanford, CA, USA, all in electrical engineering. From 2007 to 2010, he worked with the Institute for Infocomm Research, A*STAR, Singapore. Since 2010, he has been working with the National University of Singapore, where he is currently a Professor with the Department of Electrical and Computer Engineering. He has published over 200 journal articles and over 180 conference papers. He has also been listed as a Highly Cited Researcher by Thomson Reuters/Clarivate Analytics since 2015. His current research interests include UAV/satellite communications, wireless power transfer, reconfigurable MIMO, and optimization methods.

Minjian Zhao (Member, IEEE) received the M.Sc. and Ph.D. degrees in communication and information systems from Zhejiang University, Hangzhou, China, in 2000 and 2003, respectively. He was a Visiting Scholar with the University of York, U.K., in 2010. He is currently a Professor and the Deputy Director of the College of Information Science and Electronic Engineering, Zhejiang University. His research interests include modulation theory, channel estimation and equalization, MIMO, signal processing for wireless communications, anti-jamming technology for wireless transmission and networking, and communication SOC chip design.