

Channel Estimation for Massive MIMO: An Information Geometry Approach

Jiyuan Yang, *Student Member, IEEE*, An-An Lu, *Member, IEEE*, Yan Chen, *Student Member, IEEE*, Xiqi Gao, *Fellow, IEEE*, Xiang-Gen Xia, *Fellow, IEEE*, and Dirk Slock, *Fellow, IEEE*

Abstract—In this paper, we investigate the channel estimation for massive multi-input multi-output orthogonal frequency division multiplexing (MIMO-OFDM) systems. Using the sampled steering vectors in the space and frequency domain, we first establish a space-frequency (SF) beam based statistical channel model. The accuracy of the channel model can be guaranteed with sufficient sampling steering vectors. With the channel model, the channel estimation is formulated as obtaining the *a posteriori* information of the beam domain channel. We solve this problem by calculating an approximation of the *a posteriori* distribution's marginals within the information geometry framework. Specifically, by viewing the set of Gaussian distributions and the set of the marginals as a manifold and its *e*-flat submanifold, we turn the calculation of the marginals into an iterative projection process between submanifolds with different constraints. We derive the information geometry approach (IGA) for channel estimation by calculating the solutions of projections. We prove that the mean of the approximate marginals at the equilibrium of IGA is equal to that of the *a posteriori* distribution. Simulations demonstrate that the proposed IGA can accurately estimate the beam domain channel within limited iterations.

Index Terms—Massive MIMO, beam based channel model, channel estimation, information geometry.

I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) [1]–[3] is known as one of the key techniques of the fifth generation (5G) cellular systems. In a massive MIMO system, the base station (BS) equipped with a large number of antennas can serve tens of users on the same time and frequency resource simultaneously, which provides tremendous capacity gains potentially and increases the energy efficiency significantly. Among all the types of the antenna array, the uniform planar array (UPA) is a good choice for practical applications since it has a compact size and three-dimensional (3-D) coverage ability. Orthogonal frequency division multiplexing (OFDM) [4] is a multicarrier modulation technique, which can lessen

the severe effects of frequency selective fading for wideband wireless communications. Massive MIMO-OFDM plays an essential role in 5G systems and receives increasing attention for the future sixth generation (6G) systems.

In massive MIMO-OFDM systems, channel estimation plays a vital role since the system performance is highly dependent on the quality of the estimated channel. In realistic systems, pilot-aided channel estimation, i.e., the transmitter periodically sends the pilot signals, and the receiver obtains channel state information (CSI) based on the received pilot signals, is the common channel estimation approach [5]. Given the received pilot signals, channel estimation is to obtain the *a posteriori* information of the channel parameters. When the prior distribution of channel parameters is Gaussian, the *a posteriori* distribution of them is also Gaussian, of which the *a posteriori* information is given by the mean and covariance matrix. Nevertheless, calculating the *a posteriori* mean and covariance is challenging due to the large dimension of the channel in the massive MIMO-OFDM systems. The calculation of the conventional estimators, such as MMSE estimator, is not affordable since a large dimension matrix inverse is usually required.

Numerous approaches have been investigated to obtain the *a posteriori* information with relatively low complexity. Among them, methods of statistical inference, such as belief propagation (BP) [6], expectation propagation (EP) [7], generalized approximate message passing (GAMP) [8] and their variants [9]–[11], have been widely researched. These methods aim to calculate the marginals (or the approximations of them) of the *a posteriori* distributions. Then, the *a posteriori* mean and variance can be obtained. [12] proposes an estimation algorithm for massive MIMO systems based on the expectation-maximization Gaussian-mixture approximate message passing [11]. Turbo orthogonal approximate message passing (Turbo-OAMP) algorithm is proposed and applied to downlink channel estimation in massive MIMO systems in [13]. [14] proposes an algorithm for downlink channel estimation of massive MIMO systems based on EP [7].

The space defined by the parameters of the *a posteriori* PDF can be regarded as a differentiable manifold with a Riemannian structure. Hence, the definitions and tools of differential geometry can be well applied. This is exactly one of the subjects of information geometry [15]–[17]. Thus, it is appropriate to apply information geometry into the channel estimation. The main idea of information geometry is to investigate the intrinsic geometrical structures of the specific sets of PDFs by regarding the parametric space of them as

This work was supported by the National Key R&D Program of China under Grant 2018YFB1801103, the Jiangsu Province Basic Research Project under Grant BK2019200, and the Huawei Cooperation Project. Part of the material in this paper was accepted for presentation in the 14th International Conference on Wireless Communications and Signal Processing (WCSP 2022), Nanjing, China, November, 2022. (Jiyuan Yang and An-An Lu are co-first authors.) (Corresponding author: Xiqi Gao.)

Jiyuan Yang, An-An Lu, Yan Chen and Xiqi Gao are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China and are also with Purple Mountain Laboratories, Nanjing 211111, China (e-mail: {jyyang, aalu, 213160372, xqgao}@seu.edu.cn).

Xiang-Gen Xia is with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA (e-mail: xxia@ee.udel.edu).

Dirk Slock is with the Department of Communication Systems, EURECOM, 06410 Biot, France (e-mail: Dirk.Slock@eurecom.fr).

differentiable manifolds. It has been applied in multisensor estimation fusion [18], false alarm rate detection [19] and generalized Bayesian prediction [20], recently.

The main advantages of information geometry are as follows. Information geometry can provide a unified framework for analyzing the existing algorithms theoretically. The geometrical perspective provides an intuitive understanding of the statistical model and promotes an intrinsic study of the existing problems. The work [21] analyzes the turbo and low-density parity-check codes from the perspective of information geometry. The error between the solution of belief propagation (BP) [6] and the true marginals is regarded as the discrepancy between two manifolds and is analyzed from the geometrical view in [17]. Moreover, information geometry could also improve algorithms from a more general and intrinsic standpoint. Many facts have indicated that the intrinsic analysis usually yields a more general result. The work [21] proposes an explicit algorithm for the improvement of the turbo and low-density parity-check codes from the geometrical view. An acceleration method of the well-known concave-convex procedure (CCCP) algorithms for minimizing the Bethe free energies is given by [16] from the view of the natural gradient. The work [16] also proposes several new stochastic reasoning algorithms, which are called the e -constraint and m -constraint algorithms (or variants of them), respectively.

In this paper, we propose an information geometry approach for channel estimation for massive MIMO-OFDM systems. We first derive a space-frequency (SF) beam based statistical channel model by using the sampled steering vectors in the space and frequency domain. The accuracy of the SF beam based channel model is guaranteed by a sufficiently large number of sampling steering vectors. The channel estimation is formulated as obtaining the *a posteriori* information of the beam domain channel. We solve this problem by calculating an approximation of the *a posteriori* distribution's marginals within the information geometry framework. By viewing the set of Gaussian distributions and the set of the marginals as a manifold and its e -flat submanifold, we turn the calculation of the marginals into an iterative projection process between submanifolds with different constraints. By calculating the solution of m -projections, we derive the information geometry approach (IGA) for the channel estimation. We improve the stability of the proposed approach by damped updating and give a theoretical analysis from the information geometry view. It is shown that the mean of the approximate *a posteriori* marginals at the equilibrium of the algorithm is equal to that of the *a posteriori* distribution.

The rest of the article is organized as follows. The channel model is presented in Section II. Preliminaries of information geometry are introduced in Section III. The information geometry approach for channel estimation is proposed in Section IV. Simulation results are provided in Section V. The conclusion is drawn in Section VI.

Notations: We adopt the following notations in this paper. Upper (lower) case boldface letters denote matrices (column vectors). $\delta(\cdot)$ denotes the delta function. $\bar{j} = \sqrt{-1}$ is the imaginary unit. We use $\lceil x \rceil$ to denote the largest integer not larger than x . $\langle \cdot \rangle_n$ is modulo n . $*$ denotes the convolution

operator. $|\mathcal{B}|$ denotes the cardinality of the set \mathcal{B} . \setminus denotes the set subtraction operation. The notation \triangleq is used for definitions. The superscript (\cdot) , $(\cdot)^T$ and $(\cdot)^H$ denote the conjugate, transpose and conjugate-transpose operator, respectively. $\text{diag}(\mathbf{x})$ denotes the diagonal matrix with \mathbf{x} along its main diagonal and $\text{diag}(\mathbf{X})$ denotes a vector consisting of the diagonal elements of \mathbf{X} . $\text{Bdiag}(\mathbf{A}_1, \mathbf{A}_2, \dots)$ denote a block diagonal matrix with the elements \mathbf{A}_i located along the main diagonal. We use y_n or $[\mathbf{y}]_n$, a_{ij} and $[\mathbf{F}]_{:,i}$ to denote the n -th element of the vector \mathbf{y} , the (i, j) -th element of the matrix \mathbf{A} and the i -th row of the matrix \mathbf{F} , respectively, where the element indices start with 1. $\mathbb{C}^{N \times M}$ ($\mathbb{R}^{N \times M}$) denotes the $N \times M$ dimensional complex (real) vector space. $\mathbb{E}_p\{\cdot\}$ denotes the expectation operation w.r.t. the distribution $p(\mathbf{x})$. \odot and \otimes denote the Hadamard product and Kronecker product, respectively. $\text{vec}(\cdot)$ denotes the vectorization operation. $\mathbf{0}$, $\mathbf{1}$ and \mathbf{O} are the zero vector, all one vector and zero matrix with proper dimension, respectively. \circ is an operator of complex vectors or matrices, $\mathbf{a} \circ \mathbf{b} = \frac{1}{2}(\mathbf{b}^H \mathbf{a} + \mathbf{a}^H \mathbf{b})$ and $\mathbf{A} \circ \mathbf{B} = \frac{1}{2} \text{tr}\{\mathbf{B}^H \mathbf{A} + \mathbf{A}^H \mathbf{B}\}$, where $\mathbf{a}, \mathbf{b} \in \mathbb{C}^{P \times 1}$ and $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{Q \times Q}$. $\mathbf{f}(\mathbf{a}, \mathbf{A})$ is a vector function, which is defined as $\mathbf{f}(\mathbf{a}, \mathbf{A}) = [\mathbf{a}^T, \text{vec}^T(\mathbf{A})]^T \in \mathbb{C}^{(P+Q^2) \times 1}$ with $\mathbf{a} \in \mathbb{C}^{P \times 1}$ and $\mathbf{A} \in \mathbb{C}^{Q \times Q}$. It is not difficult to show that $\mathbf{f}(\mathbf{a}_1, \mathbf{A}_1) \circ \mathbf{f}(\mathbf{a}_2, \mathbf{A}_2) = \mathbf{a}_1 \circ \mathbf{a}_2 + \mathbf{A}_1 \circ \mathbf{A}_2$.

II. CHANNEL MODEL AND PROBLEM FORMULATION

In this section, we derive the SF beam based statistical channel model for massive MIMO-OFDM systems by using sampled steering vectors in the space and frequency domain. Then, the problem of the channel estimation is formulated.

A. System Model

Consider a massive MIMO-OFDM system working in time division duplexing (TDD) mode, where the base station (BS) equipped with UPA of $N_r = N_{r,v} \times N_{r,h}$ antenna elements communicates with K single-antenna users within a cell, where $N_{r,v}$ and $N_{r,h}$ are the numbers of the antennas at each vertical column and horizontal row, respectively. In TDD mode, due to the channel reciprocity, channel state information (CSI) can be obtained from uplink (UL) training and then used for UL signal detection and DL precoding. We focus on the uplink massive MIMO-OFDM channel estimation.

In the OFDM modulation, there are N_c subcarriers. The system sampling interval and the length of the cyclic prefix (CP) are denoted as T_s and N_g , respectively. Let the set of subcarriers be $\mathcal{N} = \{n | n = 0, 1, \dots, N_c - 1\}$. We assume the channel is quasi-static. Thus, we consider the UL training within one OFDM symbol. The set of the training subcarriers is denoted as $\mathcal{N}^d = \{n | n = N_1, N_1 + 1, \dots, N_2\}$, and $|\mathcal{N}^d| = N_p \leq N_c - N_v$, where N_1 and N_2 are the start and end indices of the training subcarriers, respectively, N_v is the number of virtual subcarriers. Let $\{x_k[n], n \in \mathcal{N}^d\}$ be the frequency-domain sequence transmitted by user k at an OFDM symbol.

During the UL training, the continuous-time baseband signal received by BS can be expressed as

$$\mathbf{y}^t(t) = \sum_{k=1}^K \int \mathbf{c}_k(\tau) s_k(t - \tau) d\tau + \mathbf{z}(t), \quad (1)$$

where $s_k(t)$ is the continuous time baseband signal transmitted by user k , $\mathbf{c}_k(\tau) \in \mathbb{C}^{N_r}$ is the equivalent channel impulse response with the transceiver pulse shaping considered between user k and the BS, and $\mathbf{z}(t) \in \mathbb{C}^{N_r}$ is the white Gaussian noise vector, for fixed τ and t . Assume that the range of the equivalent channel impulse response $\mathbf{c}_k(\tau)$ is not larger than $N_g T_s$. Then, the frequency-domain received vector at the n -th subcarrier can be expressed as [22]–[24]

$$\mathbf{y}_n = \sum_{k=1}^K \mathbf{g}_{n,k} x_k[n] + \mathbf{z}_n, \quad (2)$$

where $\mathbf{g}_{n,k} = \tilde{\mathbf{c}}_k(f) \Big|_{f=n\Delta_f}$, $\tilde{\mathbf{c}}_k(f)$ is the Fourier transform of $\mathbf{c}_k(\tau)$, $\Delta_f = \frac{1}{N_c T_s}$ is the subcarrier interval and $\mathbf{z}_n \in \mathbb{C}^{N_r \times 1}$ is the circularly symmetric Gaussian noise with $\mathbb{E}\{\mathbf{z}_n \mathbf{z}_n^H\} = \sigma_z^2 \mathbf{I}$. We denote the channel of user k over all training subcarriers as

$$\mathbf{G}_k = [\mathbf{g}_{N_1,k} \ \cdots \ \mathbf{g}_{N_2,k}] \in \mathbb{C}^{N_r \times N_p}. \quad (3)$$

\mathbf{G}_k is the SF domain channel coefficient matrix of user k . Let $\mathbf{Y} = [\mathbf{y}_{N_1} \ \cdots \ \mathbf{y}_{N_2}] \in \mathbb{C}^{N_r \times N_p}$, $\mathbf{X}_k = \text{diag}(\mathbf{x}_k) \in \mathbb{C}^{N_p \times N_p}$ with $\mathbf{x}_k = [x_k[N_1], \dots, x_k[N_2]]^T$ and $\mathbf{Z} = [\mathbf{z}_{N_1} \ \cdots \ \mathbf{z}_{N_2}] \in \mathbb{C}^{N_r \times N_p}$. Then, we have

$$\mathbf{Y} = \sum_{k=1}^K \mathbf{G}_k \mathbf{X}_k + \mathbf{Z}. \quad (4)$$

B. SF Beam Based Statistical Channel Model

The equivalent channel impulse response $\mathbf{c}_k(\tau)$ can be expressed as [22], [23], [25], [26],

$$\mathbf{c}_k(\tau) = \sum_{p=1}^{P_k} \alpha_{p,k} p_{tc}(\tau) * \tilde{\mathbf{s}}_{p,k}(\tau), \quad (5)$$

where P_k is the number of paths of user k , $\alpha_{p,k}$ is the complex-valued channel gain of the p -th path of user k , $p_{tc}(\tau)$ is the equivalent transceiver filter and $[\tilde{\mathbf{s}}_{p,k}(\tau)]_{n_R} = \exp(-j\bar{2}\pi f_c \tau_{n_R,p,k}) \delta(\tau - \tau_{n_R,p,k})$, with f_c and $\tau_{n_R,p,k}$ being the carrier frequency and the delay of the path p between the user k and the n_R -th antenna, respectively. With the far filed assumption, the delay of the n_R -th antenna at BS can be expressed as [27]–[29]:

$$\tau_{n_R,p,k} = \tau_{p,k} + \frac{d_{n_R,v}}{c} u_{p,k} + \frac{d_{n_R,h}}{c} v_{p,k}, \quad (6)$$

where $\tau_{p,k} = \tau_{1,p,k}$ is the delay of the first antenna, $d_{n_R,v}$ and $d_{n_R,h}$ are the vertical and the horizontal distances between the n_R -th antenna and the first antenna, respectively, $u_{p,k} = \sin \theta_{p,k}$ and $v_{p,k} = \cos \theta_{p,k} \sin \phi_{p,k}$ are the directional cosines [29], and $\theta_{p,k}, \phi_{p,k} \in [-\pi/2, \pi/2]$ are the vertical and the horizontal angles of arrival (AoA) at BS, respectively. Then, the space steering vector $\mathbf{v}(u, v)$ is defined as

$$\mathbf{v}(u, v) = \mathbf{v}_v(u) \otimes \mathbf{v}_h(v) \in \mathbb{C}^{N_r \times 1}, \quad (7a)$$

$$\mathbf{v}_v(u) = [p(1) \ p(2) \ \cdots \ p(N_{r,v})]^T \in \mathbb{C}^{N_{r,v} \times 1}, \quad (7b)$$

$$\mathbf{v}_h(v) = [q(1) \ q(2) \ \cdots \ q(N_{r,h})]^T \in \mathbb{C}^{N_{r,h} \times 1}, \quad (7c)$$

where $p(n) = \exp\left\{-j\bar{2}\pi \frac{(n-1)\Delta_v}{\lambda_c} u\right\}$, $q(n) = \exp\left\{-j\bar{2}\pi \frac{(n-1)\Delta_h}{\lambda_c} v\right\}$, Δ_v and Δ_h are the vertical and horizontal antenna spacings of the UPA, respectively, and λ_c is the carrier wavelength. We assume that $\Delta_v = \Delta_h = 0.5\lambda_c$. Then, $\mathbf{g}_{n,k} = \tilde{\mathbf{c}}_k(n\Delta_f)$, $n \in \mathcal{N}$ can be expressed as

$$\mathbf{g}_{n,k} = \beta[n] \sum_{p=1}^{P_k} \alpha_{p,k} \exp\{-j\bar{2}\pi f_c \tau_{p,k}\} \mathbf{v}(u_{p,k}, v_{p,k}) \odot \tilde{\mathbf{w}}_{p,k}(n), \quad (8)$$

where $\beta[n] = P_{tc}(2\pi f) \Big|_{f=n\Delta_f}$ is the sampled sequence of the Fourier transform of the $p_{tc}(\tau)$ and $[\tilde{\mathbf{w}}_{p,k}(n)]_{n_R} = \exp(-j\bar{2}\pi n \Delta_f \tau_{n_R,p,k})$. When the effect of beam squint [30] can be ignored, all the components of $\tilde{\mathbf{w}}_{p,k}(n)$ can be approximated by $[\tilde{\mathbf{w}}_{p,k}(n)]_1 = \exp(-j\bar{2}\pi n \Delta_f \tau_{p,k})$ with $\tau_{p,k}$ being the delay of the first antenna. We also assume that the guard band is not larger than $\frac{N_v}{N_c} B$ and $\beta[n] = 1, n \in \mathcal{N}^d$ [25]. Hence, the space-domain channel coefficient vector $\mathbf{g}_{n,k}$ can be expressed as

$$\mathbf{g}_{n,k} = \sum_p \alpha_{p,k} \exp\{-j\bar{2}\pi \tau_{p,k} (f_c + n\Delta_f)\} \mathbf{v}(u_{p,k}, v_{p,k}). \quad (9)$$

Moreover, we define the frequency steering vector $\mathbf{u}(\tau)$ as

$$\mathbf{u}(\tau) = [r(N_1) \ \cdots \ r(N_2)]^T \in \mathbb{C}^{N_p \times 1}, \quad (10)$$

where $r(n) = \exp\{-j\bar{2}\pi \Delta_f n \tau\}$. Then, the SF domain channel coefficient matrix \mathbf{G}_k (3) can be expressed as

$$\mathbf{G}_k = \sum_{p=1}^{P_k} h_{p,k} \mathbf{v}(u_{p,k}, v_{p,k}) \mathbf{u}^T(\tau_{p,k}), \quad (11)$$

where $h_{p,k} = \alpha_{p,k} \exp(-j\bar{2}\pi f_c \tau_{p,k})$. Then, the SF domain channel covariance matrix of user k is given by $\mathbf{R}_k \in \mathbb{C}^{N_r \times N_r} \triangleq \mathbb{E}_{p_n} \{\text{vec}(\mathbf{G}_k) \text{vec}^H(\mathbf{G}_k)\}$. The above expression provides a physically-motivated model for massive MIMO-OFDM systems [31]–[33].

In practice, obtaining the large dimensional $\mathbf{R}_k, \forall k$, for massive MIMO systems is resource-intensive. The problem can be significantly simplified as we can turn to calculate the statistics of paths $h_{p,k}$ in (11) instead due to the fact that the number of scatterers in the propagation environment is usually limited. Nevertheless, the acquisition of the model parameters $\{P_k, u_{p,k}, v_{p,k}, \tau_{p,k}\}$ is still complicated since there are infinitely many possible values for them. In order to make the channel model usable for practical systems, we derive the SF beam based statistical channel model of the massive MIMO systems by discretizing directional cosines u, v and the delay τ . Define $h_k(u, v, \tau) \triangleq \sum_{p=1}^{P_k} h_{p,k} \delta(u - u_{p,k}) \delta(v - v_{p,k}) \delta(\tau - \tau_{p,k})$, where the parameters $u_{p,k}, v_{p,k}$ and $\tau_{p,k}$ are the same as those in (11). The ranges of u, v and τ are denoted as $\mathcal{B}_u = \{u | u \in [-1, 1]\}$,

$\mathcal{B}_v = \{v|v \in [-1, 1)\}$ and $\mathcal{B}_\tau = \{\tau|\tau \in [0, N_g T_s)\}$, respectively. Then, the SF domain channel \mathbf{G}_k in (11) can be rewritten as

$$\mathbf{G}_k = \iiint_{u \in \mathcal{B}_u, v \in \mathcal{B}_v, \tau \in \mathcal{B}_\tau} h_k(u, v, \tau) \mathbf{v}(u, v) \mathbf{u}^T(\tau) dudvd\tau. \quad (12)$$

We then define sets $\mathcal{B}_{u,i}$, $\mathcal{B}_{v,j}$, and $\mathcal{B}_{\tau,\ell}$ as follows:

$$\mathcal{B}_{u,i} \triangleq [u_i, u_{i+1}), i \in \mathcal{Z}_{N_v}^+, \quad (13a)$$

$$\mathcal{B}_{v,j} \triangleq [v_j, v_{j+1}), j \in \mathcal{Z}_{N_h}^+, \quad (13b)$$

$$\mathcal{B}_{\tau,\ell} \triangleq [\tau_\ell, \tau_{\ell+1}), \ell \in \mathcal{Z}_{N_\tau}^+, \quad (13c)$$

where $\mathcal{Z}_N^+ \triangleq \{1, 2, \dots, N\}$, u_i , v_j and τ_ℓ are the sampled directional cosines and delays with $u_i = \frac{2(i-1)-N_v}{N_v}$, $i \in \mathcal{Z}_{N_v}^+$, $v_j = \frac{2(j-1)-N_h}{N_h}$, $j \in \mathcal{Z}_{N_h}^+$, and $\tau_\ell = \frac{(\ell-1)N_f}{N_\tau N_p \Delta_f}$, $\ell \in \mathcal{Z}_{N_\tau}^+$, and $N_f = \lceil N_p N_g / N_c \rceil$. $N_v \triangleq F_v N_{r,v}$, $N_h \triangleq F_h N_{r,h}$ and $N_\tau \triangleq F_\tau N_f$, where F_v , F_h , and F_τ are the fine factors (FFs). We assume N_v , N_h and N_τ are not less than $N_{r,v}$, $N_{r,h}$ and N_f , respectively. Then, (12) can be rewritten as

$$\mathbf{G}_k = \sum_{\substack{i,j,\ell \\ u \in \mathcal{B}_{u,i}, v \in \\ \mathcal{B}_{v,j}, \tau \in \mathcal{B}_{\tau,\ell}}} \iiint h_k(u, v, \tau) \mathbf{v}(u, v) \mathbf{u}^T(\tau) dudvd\tau. \quad (14)$$

When N_v, N_h and N_τ are sufficient large, the sizes of the subsets $\mathcal{B}_{u,i}$, $\mathcal{B}_{v,j}$ and $\mathcal{B}_{\tau,\ell}$ are small, and thus, the steering vectors $\mathbf{v}_v(u)$ in $\mathcal{B}_{u,i}$, $\mathbf{v}_h(v)$ in $\mathcal{B}_{v,j}$ and $\mathbf{u}(\tau)$ in $\mathcal{B}_{\tau,\ell}$ can be well approximated by the sampled space steering vectors $\mathbf{v}_v(u_i)$, $\mathbf{v}_h(v_j)$ and the sampled frequency steering vectors $\mathbf{u}(\tau_\ell)$, respectively. Hence, the SF domain channel matrix (14) can be approximated as

$$\mathbf{G}_k \approx \sum_{i,j,\ell} \tilde{h}_k(u_i, v_j, \tau_\ell) \mathbf{v}(u_i, v_j) \mathbf{u}^T(\tau_\ell), \quad (15)$$

where

$$\tilde{h}_k(u_i, v_j, \tau_\ell) = \iiint_{u \in \mathcal{B}_{u,i}, v \in \\ \mathcal{B}_{v,j}, \tau \in \mathcal{B}_{\tau,\ell}} h_k(u, v, \tau) dudvd\tau. \quad (16)$$

Let $\mathbf{V} \in \mathbb{C}^{N_\tau \times N_v N_h} = \mathbf{V}_v \otimes \mathbf{V}_h$ and $\mathbf{F} = [\mathbf{u}(\tau_1) \dots \mathbf{u}(\tau_{N_\tau})] \in \mathbb{C}^{N_p \times N_\tau}$, where $\mathbf{V}_v = [\mathbf{v}_v(u_1) \dots \mathbf{v}_v(u_{N_v})] \in \mathbb{C}^{N_{r,v} \times N_v}$ and $\mathbf{V}_h = [\mathbf{v}_h(v_1) \dots \mathbf{v}_h(v_{N_h})] \in \mathbb{C}^{N_{r,h} \times N_h}$. Then, the SF domain channel matrix \mathbf{G}_k can be expressed as,

$$\mathbf{G}_k = \mathbf{V} \mathbf{H}_k \mathbf{F}^T, \quad (17)$$

where $\mathbf{H}_k \in \mathbb{C}^{N_v N_h \times N_\tau}$ and $[\mathbf{H}_k]_{m,n} = \tilde{h}_k(u_i, v_j, \tau_n)$ with $i = \lfloor \frac{m-1}{N_h} + 1 \rfloor$ and $j = m - (i-1)N_h$. In the channel representation of (17), all the users share the same sets of sampled space and frequency steering vectors. Each sampled space/frequency steering vector corresponds to a physical beam in space/frequency domain. Therefore, we refer to it as SF beam based channel model. \mathbf{H}_k is the SF beam domain channel matrix. With the assumption of wide-sense stationary uncorrelated scattering Rayleigh fading channel, the elements in \mathbf{H}_k follow the independent complex Gaussian distributions

with zero mean and different variances, and we define the beam domain channel power matrix as:

$$\mathbf{\Omega}_k = \mathbb{E} \{ \mathbf{H}_k \odot \overline{\mathbf{H}_k} \}. \quad (18)$$

Compared with the SF domain channel covariance matrix \mathbf{R}_k , the dimension of the beam domain channel power matrix $\mathbf{\Omega}_k$ is substantially smaller. Meanwhile, due to the channel sparsity, most of the elements in $\mathbf{\Omega}_k$ are close to zero, and the non-zero elements usually gather in clusters, where each cluster corresponds to a physical scatterer [24], [34]. Therefore, there are sufficient resources to acquire $\mathbf{\Omega}_k, \forall k$. For instance, the authors of [29] propose a method that can obtain the estimate of $\mathbf{\Omega}_k$ with guaranteed accuracy and low complexity. Thus, we assume that $\mathbf{\Omega}_k$ of all users are known at the BS in the rest of the paper. When the virtual subcarriers are eliminated, the angle-delay domain channel matrix [24], [35] is equivalent to \mathbf{H}_k with the fine factors set to be 1. The key difference between the SF beam based channel model and the channel model in [24] is that the fine factors can be set to be greater than 1 in the SF beam based channel model, which allows to sample the AoAs as well as the delay more intensively in the SF beam based channel model. Fine sampling is necessary to accurately model the channels of massive MIMO-OFDM systems. It is not difficult to show that both \mathbf{V} and \mathbf{F} are (partial) discrete Fourier transform (DFT) matrices.

C. Problem Formulation

During the UL training, the task of channel estimation is to obtain the *a posteriori* information of the SF domain channel matrix $\mathbf{G}_k, \forall k$. The *a posteriori* information of \mathbf{G}_k can be calculated from that of the beam domain channel matrix \mathbf{H}_k through (17). Thus, we focus on the estimation of $\mathbf{H}_k, \forall k$. By substituting (17) into the UL received signal model (4), we have

$$\mathbf{Y} = \mathbf{V} \mathbf{H}_a \mathbf{M} + \mathbf{Z}, \quad (19)$$

where $\mathbf{H}_a = [\mathbf{H}_1 \mathbf{H}_2 \dots \mathbf{H}_K] \in \mathbb{C}^{N_v N_h \times K N_\tau}$ and $\mathbf{M} = [\mathbf{X}_1 \mathbf{F} \mathbf{X}_2 \mathbf{F} \dots \mathbf{X}_K \mathbf{F}]^T \in \mathbb{C}^{K N_\tau \times N_p}$. After the vectorization of (19) and removing the elements in $\text{vec}(\mathbf{H}_a)$ with zero variance, denoted as \mathbf{h} , and the corresponding column in $\mathbf{M}^T \otimes \mathbf{V}$, we have,

$$\mathbf{y} = \mathbf{A} \mathbf{h} + \mathbf{z}, \quad (20)$$

where $\mathbf{A} \in \mathbb{C}^{N \times M}$ is a deterministic matrix extracted from $\mathbf{M}^T \otimes \mathbf{V}$, $N = N_r N_p$, M is the number of elements in \mathbf{H}_a with non-zero variance, \mathbf{y} and \mathbf{z} are the vectorizations of \mathbf{Y} and \mathbf{Z} , respectively, $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \mathbf{D})$ with positive definite and diagonal \mathbf{D} , and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \sigma_z^2 \mathbf{I})$. \mathbf{h} and \mathbf{z} are assumed to be independent with each other. M can be further interpreted as the number of physical beams in space/frequency domain for all users. Then, the *a posteriori* distribution is also Gaussian, i.e., $p(\mathbf{h}|\mathbf{y}) = p_G(\mathbf{h}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, where $p_G(\mathbf{h}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the PDF of a complex Gaussian distribution $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The *a posteriori* mean $\tilde{\boldsymbol{\mu}}$ and covariance $\tilde{\boldsymbol{\Sigma}}$ are given by [36]

$$\tilde{\boldsymbol{\mu}} = \mathbf{D} (\mathbf{A}^H \mathbf{A} \mathbf{D} + \sigma_z^2 \mathbf{I})^{-1} \mathbf{A}^H \mathbf{y}, \quad (21a)$$

$$\tilde{\boldsymbol{\Sigma}} = \left(\mathbf{D}^{-1} + \frac{1}{\sigma_z^2} \mathbf{A}^H \mathbf{A} \right)^{-1}. \quad (21b)$$

It should be noted that the MMSE estimate of \mathbf{h} is equivalent to the *a posteriori* mean (21a) [36].

If the fine factors are set to be 1 and the adjustable phase shift pilots in [24] are exploited, the columns of \mathbf{A} can be orthogonal to each other, i.e., $\mathbf{A}^H \mathbf{A} = \mathbf{I}$. In this case, the computational complexity of the *a posteriori* mean and covariance in (21) will be $\mathcal{O}(NM)$ and $\mathcal{O}(M)$, respectively. However, since the corresponding channel model is inaccurate when the number of antennas or training subcarriers is limited, the performance of the channel estimation will degrade. On the other hand, the proposed SF beam based channel model can accurately model the physical channel by increasing the fine factors greater than 1 and, therefore, can significantly improve the channel estimation performance. Under this circumstance, however, the columns of \mathbf{A} are not orthogonal to each other anymore. The computational complexity of the *a posteriori* information in (21) will become $\mathcal{O}(M^3 + M^2N)$. When the number of users is relatively large, M can be comparable to N even though channel sparsity exists. In this case, it is unaffordable to apply (21) in practice when both N and M are large. In this work, we propose the information geometry approach (IGA) for channel estimation which aims to obtain an approximation of the marginals, $p(h_i|\mathbf{y})$, $i = 1, 2, \dots, M$, of the *a posteriori* distribution $p(\mathbf{h}|\mathbf{y})$. Note that the proposed IGA can serve as a generic Bayesian inference technique, although we focus on the channel estimation for massive MIMO systems.

III. PRELIMINARIES OF INFORMATION GEOMETRY

In this section, we give preliminaries of information geometry. Its basic concepts and application to statistical inference are briefly introduced. More details can be found in [15], [16].

An exponential family is defined as [15]–[17]

$$p(\mathbf{x}; \boldsymbol{\vartheta}) = \exp \{ \boldsymbol{\vartheta} \circ \mathbf{t} - \psi(\boldsymbol{\vartheta}) \}, \quad (22)$$

where \mathbf{t} is the sufficient statistic of random vector \mathbf{x} , $\boldsymbol{\vartheta}$ is the natural parameter of the exponential family, and $\psi(\boldsymbol{\vartheta})$ is the normalization factor, which is called the partition function or the (Helmholtz) free energy and makes $\int p(\mathbf{x}; \boldsymbol{\vartheta}) d\mathbf{x} = 1$. The notation \circ is an operator of complex vectors or matrices, $\mathbf{a} \circ \mathbf{b} = \frac{1}{2}(\mathbf{b}^H \mathbf{a} + \mathbf{a}^H \mathbf{b})$ and $\mathbf{A} \circ \mathbf{B} = \frac{1}{2} \text{tr} \{ \mathbf{B}^H \mathbf{A} + \mathbf{A}^H \mathbf{B} \}$, where $\mathbf{a}, \mathbf{b} \in \mathbb{C}^{P \times 1}$ and $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{Q \times Q}$. Many important distributions are included in the exponential family, such as the Poisson distribution, the multinomial distribution and the Gaussian distribution.

A (sub)manifold U is said to be e -flat if for all $a \in [0, 1]$, $q(\mathbf{x}), p(\mathbf{x}) \in U$, the following $r(\mathbf{x}; a)$ belongs to U [16]

$$\ln r(\mathbf{x}; a) = (1 - a) \ln q(\mathbf{x}) + a \ln p(\mathbf{x}) + c(a), \quad (23)$$

where $c(a)$ is the normalization factor. Intuitively, a submanifold is e -flat, when all the distributions in it are linear in logarithm. From the definition, any exponential family is e -flat [16].

Let S be a manifold and $U \subset S$ be an e -flat submanifold. Let $p(\mathbf{x}; \boldsymbol{\vartheta}_q) \in S$. The point in U that minimizes

the Kullback-Leibler (K-L) divergence from $p(\mathbf{x}; \boldsymbol{\vartheta}_q)$ to U , denoted by

$$\pi_U \{p(\mathbf{x}; \boldsymbol{\vartheta}_q)\} = \arg \min_{p(\mathbf{x}; \boldsymbol{\vartheta}_p) \in U} D_{KL} \{p(\mathbf{x}; \boldsymbol{\vartheta}_q); p(\mathbf{x}; \boldsymbol{\vartheta}_p)\}, \quad (24)$$

is called the m -projection of $p(\mathbf{x}; \boldsymbol{\vartheta}_q)$ to U , where the K-L divergence is given by

$$D_{KL} \{p(\mathbf{x}; \boldsymbol{\vartheta}_q); p(\mathbf{x}; \boldsymbol{\vartheta}_p)\} = \mathbb{E}_{p(\mathbf{x}; \boldsymbol{\vartheta}_q)} \left\{ \ln \frac{p(\mathbf{x}; \boldsymbol{\vartheta}_q)}{p(\mathbf{x}; \boldsymbol{\vartheta}_p)} \right\}. \quad (25)$$

Let $\mathbf{x}_h \in \mathbb{R}^{M \times 1}$ and $\mathbf{y}_o \in \mathbb{R}^{N \times 1}$ be hidden and observed random variables, respectively. Denote the *a posteriori* distribution as $p(\mathbf{x}_h|\mathbf{y}_o)$, and we want to calculate its marginals, i.e., $p(x_i|\mathbf{y}_o)$, $i = 1, 2, \dots, M$, with x_i being the i -th element of \mathbf{x}_h . We focus on the cases where all the components of \mathbf{x}_h are independent and all the components of \mathbf{y}_o given \mathbf{x}_h are independent as well, i.e.,

$$p(\mathbf{x}_h|\mathbf{y}_o) \propto p(\mathbf{x}_h) p(\mathbf{y}_o|\mathbf{x}_h) = \prod_{i=1}^M p_i(x_i) \prod_{n=1}^N p_n(y_{o,n}|\mathbf{x}_h), \quad (26)$$

where $y_{o,n}$ is the n -th element of \mathbf{y}_o , $p_i(x_i)$ are the marginals of $p(\mathbf{x}_h)$ and $p_n(y_{o,n}|\mathbf{x}_h)$ are the marginals of $p(\mathbf{y}_o|\mathbf{x}_h)$. Assume that each marginal distribution of $p(\mathbf{x}_h)$ belongs to the exponential family, i.e.,

$$p_i(x_i) = p(x_i; \mathbf{d}_i) = \exp \{ \mathbf{d}_i \circ \mathbf{t}_i - \psi_i \}, \quad (27)$$

where $\mathbf{d}_i = [d_{i,1}, d_{i,2}, \dots, d_{i,M_i}]^T$ is the natural parameter, $\mathbf{t}_i = [t_{i,1}, t_{i,2}, \dots, t_{i,M_i}]^T \in \mathbb{R}^{M_i \times 1}$ is the sufficient statistic of the single random variable x_i , e.g., x_i and $|x_i|^2$, and ψ_i is the free energy. The prior distribution can be expressed as

$$p(\mathbf{x}_h) = \prod_{i=1}^M p(x_i; \mathbf{d}_i) = \exp \{ \mathbf{d}_h \circ \mathbf{t}_h - \psi_h \}, \quad (28)$$

where $\mathbf{d}_h = [\mathbf{d}_1^T, \mathbf{d}_2^T, \dots, \mathbf{d}_M^T]^T \in \mathbb{R}^{M_a \times 1}$, $\mathbf{t}_h = [\mathbf{t}_1^T, \mathbf{t}_2^T, \dots, \mathbf{t}_M^T]^T \in \mathbb{R}^{M_a \times 1}$, and $M_a = \sum_i M_i$. For the marginals of the conditional distribution, we also assume they belong to the exponential family and can be expressed as

$$p_n(y_{o,n}|\mathbf{x}_h) = \exp \{ c_n(\mathbf{x}_h) - \psi_n \}, \quad (29)$$

where $c_n(\mathbf{x}_h)$ is a polynomial of \mathbf{x}_h parameterized by variables including $y_{o,n}$, and ψ_n is the free energy. $c_n(\mathbf{x}_h)$ is called the interaction item in information geometry since it usually contains interactions between the random variables. We will see an example of c_n in the next section. Then, the *a posteriori* distribution can be expressed as

$$p(\mathbf{x}_h|\mathbf{y}_o) = \exp \left\{ \mathbf{d}_h \circ \mathbf{t}_h + \sum_{n=1}^N c_n(\mathbf{x}_h) - \psi_q \right\}, \quad (30)$$

where ψ_q is the normalization factor.

Let us then consider a set of distributions, which are parameterized by $\boldsymbol{\vartheta}_0$,

$$M_0 = \{p_0(\mathbf{x}_h; \boldsymbol{\vartheta}_0) | \boldsymbol{\vartheta}_0 \in \mathbb{R}^{M_a \times 1}\}, \quad (31a)$$

$$p_0(\mathbf{x}_h; \boldsymbol{\vartheta}_0) = \exp \{ \mathbf{d}_h \circ \mathbf{t}_h + \boldsymbol{\vartheta}_0 \circ \mathbf{t}_h - \psi_0(\boldsymbol{\vartheta}_0) \}, \quad (31b)$$

where \mathbf{d}_h and \mathbf{t}_h are defined the same as before. Distributions of M_0 contain no interaction between x_i and x_j , $i \neq j$, since each component of \mathbf{t}_h only includes the sufficient statistic of a single random variable x_i . Thus, their marginal distributions can be obtained easily. We call M_0 the objective manifold (OBM). Any distribution in M_0 belongs to the exponential family, and M_0 is e -flat. Comparing (30) with (31b), we can obtain the approximations of the marginals if we can approximate $\sum_n c_n(\mathbf{x}_h)$ as $\vartheta_0 \circ \mathbf{t}_h$.

We m -project $p(\mathbf{x}_h|\mathbf{y}_o)$ onto M_0 and denote the resultant parameter as

$$\vartheta_0^* = \arg \min_{\vartheta_0} D_{KL} \{p(\mathbf{x}_h|\mathbf{y}_o); p_0(\mathbf{x}_h; \vartheta_0)\}. \quad (32)$$

According to [15, Theorem 11.6] and [16, Theorem 1], the m -projection $p_0(\mathbf{x}_h; \vartheta_0^*)$ is unique, and $p_0(\mathbf{x}_h; \vartheta_0^*)$ keeps the expectation of \mathbf{t}_h invariant, i.e., $\mathbb{E}_{p(\mathbf{x}_h|\mathbf{y}_o)} \{\mathbf{t}_h\} = \mathbb{E}_{p_0(\mathbf{x}_h; \vartheta_0^*)} \{\mathbf{t}_h\}$. The later property is of great importance in practical applications since we need the *a posteriori* marginal information and the m -projection keeps it invariant. Nevertheless, the calculation of the direct m -projection in (32) may be unacceptable with large M since the direct m -projection may be too complicated.

To solve the problem mentioned above, N auxiliary manifolds (AMs) parameterized by $\vartheta_n, n \in \mathcal{Z}_N^+$ are introduced,

$$M_n = \{p_n(\mathbf{x}_h; \vartheta_n) | \vartheta_n \in \mathbb{R}^{M_a \times 1}\}, \quad (33a)$$

$$p_n(\mathbf{x}_h; \vartheta_n) = \exp \{(\mathbf{d}_h + \vartheta_n) \circ \mathbf{t}_h + c_n(\mathbf{x}_h) - \psi_n(\vartheta_n)\}, \quad (33b)$$

Only the n -th interaction item $c_n(\mathbf{x}_h)$ is retained in M_n , and all others, i.e., $\sum_{n' \neq n} c_{n'}(\mathbf{x}_h)$ in (30) are replaced by $\vartheta_n \circ \mathbf{t}_h$. Compared with $\sum_{n=1}^N c_n$, c_n may have some distinctive properties, e.g., the quadratic form with a full-rank matrix in $\sum_{n=1}^N c_n$ will become the one with a rank-1 matrix in c_n . As we shall see in the next section, the term c_n we use will have such a property so that the computation of the m -projection of $p_n(\mathbf{x}_h; \vartheta_n)$ is much simpler than that of the original m -projection in (32) of $p(\mathbf{x}_h|\mathbf{y}_o)$. By m -projecting $p_n(\mathbf{x}_h; \vartheta_n)$ of M_n onto OBM, we can obtain an approximation of single interaction polynomial $c_n(\mathbf{x}_h)$. For $p_n(\mathbf{x}_h; \vartheta_n)$, we denote the parameter of its m -projection as ϑ_{0n} , then the m -projection can be expressed as

$$\begin{aligned} p_0(\mathbf{x}_h; \vartheta_{0n}) &= \exp(\mathbf{d}_h \circ \mathbf{t}_h + \vartheta_{0n} \circ \mathbf{t}_h - \psi_0) \\ &= \exp(\mathbf{d}_h \circ \mathbf{t}_h + \vartheta_n \circ \mathbf{t}_h + \xi_n \circ \mathbf{t}_h - \psi_0), \end{aligned} \quad (34)$$

where the resultant parameters ϑ_{0n} can be regarded as a sum of two components: the parameter of p_n and the approximation item of the interaction item $c_n(\mathbf{x}_h)$, which is denoted as ξ_n . Thus, the parameter of the approximation of the interaction item $c_n(\mathbf{x}_h)$ is calculated as

$$\xi_n = \vartheta_{0n} - \vartheta_n, \forall n. \quad (35)$$

Then, the parameter ϑ_0 of $p_0(\mathbf{x}_h; \vartheta_0)$ in OBM can be approximated as $\vartheta_0 = \sum_{n=1}^N \xi_n$, since $\vartheta_0 \circ \mathbf{t}_h$ in (31b) can be viewed as the approximation of $\sum_n c_n(\mathbf{x}_h)$ in (30). $p_0(\mathbf{x}_h; \vartheta_0)$ is used as an approximation of the m -projection $p_0(\mathbf{x}_h; \vartheta_0^*)$ of $p(\mathbf{x}_h|\mathbf{y}_o)$. Therefore, the approximation of the

marginals of the *a posteriori* distribution can be obtained through p_0 .

From the above description, one can see that the m -projection of $p(\mathbf{x}_h|\mathbf{y}_o)$ to OBM is converted to the m -projections of a sequence of $p_n(\mathbf{x}_h; \vartheta_n)$ to OBM. We will see later why this conversion may simplify the computations. Note that the complete algorithm is an iterative procedure [16]. Specifically, we first initialize ϑ_n with any value. Then, calculate the m -projection ϑ_{0n} and the approximation item ξ_n . The parameter of M_n is then updated as $\vartheta_n = \sum_{n' \neq n} \xi_{n'}$. Note that $\vartheta_n \circ \mathbf{t}_h$ replaces $\sum_{n' \neq n} c_{n'}(\mathbf{x}_h)$ in p_n and each interaction item $c_n(\mathbf{x}_h)$ is approximated as $\xi_n \circ \mathbf{t}_h$ after m -projection. Thus, we update ϑ_n with the sum of the approximate items. Similarly, the parameter of M_0 is updated as $\vartheta_0 = \sum_{n=1}^N \xi_n$. Then, repeat the m -projections, calculate the approximation items and the updates until convergence. Discussions on the accuracy of the approximation and the convergence with a fixed point of the above algorithm can be found in [17].

IV. INFORMATION GEOMETRY APPROACH FOR CHANNEL ESTIMATION

This section applies information geometry into the beam domain channel estimation. We present a detailed description of the proposed information geometry approach (IGA), and analyze the stability of the IGA at equilibrium from the viewpoint of information geometry.

A. IGA for Beam Domain Channel Estimation

Recall the received signal model (20), with prior distributions of \mathbf{h} and \mathbf{z} being independent Gaussian distributions, i.e., $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \mathbf{D})$ with positive definite and diagonal \mathbf{D} , and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \sigma_z^2 \mathbf{I})$, the *a posteriori* distribution is also Gaussian, of which the PDF is expressed as

$$\begin{aligned} p(\mathbf{h}|\mathbf{y}) &= C \prod_{i=1}^M p_i(h_i) \prod_{n=1}^N p_n(y_n|\mathbf{h}) \\ &= C' \exp \{-\mathbf{h}^H \mathbf{D}^{-1} \mathbf{h}\} \prod_{n=1}^N \exp \left\{ -\frac{|y_n - \gamma_n^H \mathbf{h}|^2}{\sigma_z^2} \right\} \\ &= \exp \left\{ \mathbf{d}_h \circ \mathbf{t}_h + \sum_{n=1}^N c_n(\mathbf{h}) - \psi_q \right\}, \end{aligned} \quad (36)$$

where y_n is the n -th element of \mathbf{y} , $\mathbf{d}_h = \mathbf{f}(\mathbf{0}, -\mathbf{D}^{-1})$, $\mathbf{t}_h = \mathbf{f}(\mathbf{h}, \mathbf{I} \odot (\mathbf{h}\mathbf{h}^H))$, C, C' and ψ_q are the normalization factors and $c_n(\mathbf{h})$ is given by

$$c_n(\mathbf{h}) = -\mathbf{h}^H \frac{\gamma_n \gamma_n^H}{\sigma_z^2} \mathbf{h} + \mathbf{h}^H \frac{\gamma_n y_n}{\sigma_z^2} + \frac{y_n^H \gamma_n^H}{\sigma_z^2} \mathbf{h}, \quad (37)$$

$$\gamma_n = [\mathbf{A}^H]_{:,n} = [\bar{a}_{n1} \ \cdots \ \bar{a}_{nM}]^T \in \mathbb{C}^{M \times 1}. \quad (38)$$

$\mathbf{f}(\mathbf{a}, \mathbf{A})$ above is defined as $\mathbf{f}(\mathbf{a}, \mathbf{A}) = [\mathbf{a}^T, \text{vec}^T(\mathbf{A})]^T \in \mathbb{C}^{(P+Q^2) \times 1}$ with $\mathbf{a} \in \mathbb{C}^{P \times 1}$ and $\mathbf{A} \in \mathbb{C}^{Q \times Q}$. It can be shown that $\mathbf{f}(\mathbf{a}_1, \mathbf{A}_1) \circ \mathbf{f}(\mathbf{a}_2, \mathbf{A}_2) = \mathbf{a}_1 \circ \mathbf{a}_2 + \mathbf{A}_1 \circ \mathbf{A}_2$. From (36), we can find that \mathbf{t}_h only contains the sufficient statistics of single random variables, i.e., h_i as well as $|h_i|^2$,

and no interactions terms of $h_i, i = 1, 2, \dots, M$, while $c_n(\mathbf{h})$ contains the interactions, i.e., the off-diagonal elements of $\mathbf{h}\mathbf{h}^H$.

According to (36), we can define OBM and AMs directly as in the previous section,

$$M_0 : p_0(\mathbf{h}; \boldsymbol{\vartheta}_0) = \exp\{\mathbf{d}_h \circ \mathbf{t}_h + \boldsymbol{\vartheta}_0 \circ \mathbf{t}_h - \psi_0(\boldsymbol{\vartheta}_0)\}, \quad (39a)$$

$$M_n : p_n(\mathbf{h}; \boldsymbol{\vartheta}_n) = \exp\{\mathbf{d}_h \circ \mathbf{t}_h + \boldsymbol{\vartheta}_n \circ \mathbf{t}_h + c_n(\mathbf{h}) - \psi_n(\boldsymbol{\vartheta}_n)\}, n = 1, 2, \dots, N, \quad (39b)$$

where $\boldsymbol{\vartheta}_n = \mathbf{f}(\boldsymbol{\theta}_n, \boldsymbol{\Theta}_n)$ is the parameter of p_n , $\boldsymbol{\theta}_n \in \mathbb{C}^{M \times 1}$, $\boldsymbol{\Theta}_n \in \mathbb{D}^M$ is a diagonal parameter matrix for p_n , \mathbb{D}^M is the set of $M \times M$ real diagonal matrices, $\psi_n(\boldsymbol{\vartheta}_n)$ is the free energy, the subscript $n \in \mathcal{Z}_N$, $\mathcal{Z}_N \triangleq \{0, 1, \dots, N\}$, and $\mathbf{d}_h, \mathbf{t}_h$ as well as $c_n(\mathbf{h})$ are the same as before. The free energy $\psi_n(\boldsymbol{\vartheta}_n)$, $n \in \mathcal{Z}_N$ can be calculated through Gaussian integration [37, Equation 3.18, pp 104] and is given by (40), where $\mathcal{Z}_N^+ = \{1, 2, \dots, N\}$. From the definition, both OBM and AMs are e-flat.

We can now calculate the m -projection of the *a posteriori* distribution $p(\mathbf{h}|\mathbf{y})$ and p_n in AM onto OBM by minimizing the KL divergence, respectively. Nevertheless, the calculation of these two results is repetitive. In fact, the m -projections of all Gaussian distributions onto OBM have a unified form. To show it, we first define the set of Gaussian distributions as a manifold, which is called original manifold (OM). Then, we show that both OBM and AMs are submanifolds of OM. Finally, from the definition of the m -projection in the previous section, we can obtain the m -projection of any Gaussian distribution onto OBM. OM M_{or} is defined as the set of PDFs of M dimensional complex Gaussian random vectors

$$M_{or} = \{p(\mathbf{h}) | p(\mathbf{h}) = p_G(\mathbf{h}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{\mu} \in \mathbb{C}^{M \times 1}, \boldsymbol{\Sigma} \in \mathbb{H}_+^M\}, \quad (41)$$

where $p_G(\mathbf{h}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the PDF of the complex Gaussian distribution $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and \mathbb{H}_+^M is the set of $M \times M$ positive definite matrices. Any M dimensional complex Gaussian distribution can be viewed as a point in M_{or} . Then, $p(\mathbf{h}|\mathbf{y}) \in M_{or}$ can be obtained immediately since the *a posteriori* distribution $p(\mathbf{h}|\mathbf{y})$ is Gaussian, of which the mean $\tilde{\boldsymbol{\mu}}$ and covariance $\tilde{\boldsymbol{\Sigma}}$ are given by (21a) and (21b), respectively. The distributions in OBM also belong to M_{or} since $p_0(\mathbf{h}; \boldsymbol{\vartheta}_0)$ in (39a) can be rewritten as

$$\begin{aligned} & p_0(\mathbf{h}; \boldsymbol{\vartheta}_0) \\ &= \exp\left\{-\mathbf{h}^H (\mathbf{D}^{-1} - \boldsymbol{\Theta}_0) \mathbf{h} + \frac{1}{2} (\boldsymbol{\theta}_0^H \mathbf{h} + \mathbf{h}^H \boldsymbol{\theta}_0) - \psi_0\right\} \\ &= \frac{\exp\left\{-\mathbf{h}^H \boldsymbol{\Sigma}_0^{-1} \mathbf{h} + \boldsymbol{\mu}_0^H \boldsymbol{\Sigma}_0^{-1} \mathbf{h} + \mathbf{h}^H \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_0^H \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0\right\}}{\pi^M \det \boldsymbol{\Sigma}_0} \\ &= p_G(\mathbf{h}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \end{aligned} \quad (42)$$

where ψ_0 is given by (40a), the mean $\boldsymbol{\mu}_0$ and the covariance matrix $\boldsymbol{\Sigma}_0$ of p_0 are given by

$$\boldsymbol{\mu}_0 = \frac{1}{2} (\mathbf{D}^{-1} - \boldsymbol{\Theta}_0)^{-1} \boldsymbol{\theta}_0, \quad (43a)$$

$$\boldsymbol{\Sigma}_0 = (\mathbf{D}^{-1} - \boldsymbol{\Theta}_0)^{-1}, \quad (43b)$$

respectively. Thus, OBM M_0 is a submanifold of OM M_{or} . Similarly, distributions in AMs are Gaussian since p_n in (39b) can be rewritten as $p_n(\mathbf{h}; \boldsymbol{\vartheta}_n) = p_G(\mathbf{h}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, $n \in \mathcal{Z}_N^+$ where ψ_n is given by (40b), the mean $\boldsymbol{\mu}_n$ and the covariance $\boldsymbol{\Sigma}_n$ of p_n are given by (44a) and (44b), respectively, and (a) is from Sherman-Morrison formula. AMs are also submanifolds of OM. Finally, as proved above, M_0 in (39a) is an e -flat submanifold of M_{or} , thus, given a Gaussian distribution $p \in M_{or}$, we can find its m -projection on M_0 .

Theorem 1. For a given Gaussian distribution $p \in M_{or}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, the m -projection of p onto OBM is given by $p_0(\mathbf{h}; \boldsymbol{\vartheta}_0^*)$, where $\boldsymbol{\vartheta}_0^* = \mathbf{f}(\boldsymbol{\theta}_0^*, \boldsymbol{\Theta}_0^*)$ with

$$\boldsymbol{\theta}_0^* = 2(\mathbf{I} \odot \boldsymbol{\Sigma})^{-1} \boldsymbol{\mu}, \quad (45a)$$

$$\boldsymbol{\Theta}_0^* = \mathbf{D}^{-1} - (\mathbf{I} \odot \boldsymbol{\Sigma})^{-1}. \quad (45b)$$

Proof. See in Appendix A. \square

From Theorem 1, the m -projection on OBM is determined only by the mean and the diagonal of the covariance matrix of the original distribution p . To calculate the m -projection of the *a posteriori* distribution $p(\mathbf{h}|\mathbf{y})$ onto OBM, we need the *a posteriori* mean in (21a) and the diagonal of the *a posteriori* covariance in (21b). The computational complexity of such a process is still $\mathcal{O}(M^3 + M^2N)$ since (21) involves matrix inversions. Nevertheless, the computational complexity will be significantly reduced when we turn to m -project $p_n(\mathbf{h}; \boldsymbol{\vartheta}_n)$ in M_n onto OBM by using Theorem 1 and (44)-(45) as follows. In the covariance matrix of p_n , i.e., (44b), $(\boldsymbol{\gamma}_n \boldsymbol{\gamma}_n^H) / \sigma_z^2$ is a rank-1 matrix. Thus, Sherman-Morrison formula can be used for the inversion in (44b). Then, according to [38, Equation 11.42, pp 252], the diagonal of $\boldsymbol{\Sigma}_n$ can be calculated as

$$\begin{aligned} \mathbf{I} \odot \boldsymbol{\Sigma}_n &= (\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} - \\ &= \frac{(\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} \mathbf{I} \odot (\boldsymbol{\gamma}_n \boldsymbol{\gamma}_n^H) (\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1}}{\sigma_z^2 + \boldsymbol{\gamma}_n^H (\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} \boldsymbol{\gamma}_n}, \end{aligned} \quad (46)$$

of which the computational complexity is $\mathcal{O}(M)$ since $(\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)$ is diagonal. Furthermore, it can be checked that the computational complexity of the mean $\boldsymbol{\mu}_n$ in (44a) is also $\mathcal{O}(M)$ since $(\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)$ is diagonal. These results finally lead to that the computational complexity of the m -projection of p_n onto OBM is $\mathcal{O}(M)$. Denote the m -projection of $p_n(\mathbf{h}; \boldsymbol{\vartheta}_n)$, $n \in \mathcal{Z}_N^+$ onto OBM as $p_0(\mathbf{h}; \boldsymbol{\vartheta}_{0n})$, where $\boldsymbol{\vartheta}_n = \mathbf{f}(\boldsymbol{\theta}_n, \boldsymbol{\Theta}_n)$ and $\boldsymbol{\vartheta}_{0n} = \mathbf{f}(\boldsymbol{\theta}_{0n}, \boldsymbol{\Theta}_{0n})$. $\boldsymbol{\theta}_{0n}$ as well as $\boldsymbol{\Theta}_{0n}$ can be calculated through substituting (44a) and (44b) into (45). After some calculations, $\boldsymbol{\vartheta}_{0n}$ is given by (47). The matrices to be inverted in (47) are all diagonal with dimension M . Thus, the computational complexities of both (47a) and (47b) are $\mathcal{O}(M)$.

After the m -projection of p_n on OBM is obtained, the approximation item of $c_n(\mathbf{h})$ can be calculated as

$$\boldsymbol{\xi}_n = \boldsymbol{\vartheta}_{0n} - \boldsymbol{\vartheta}_n, n \in \mathcal{Z}_N^+, \quad (48)$$

where $\boldsymbol{\vartheta}_n$ is the parameter of the $p_n(\mathbf{h}; \boldsymbol{\vartheta}_n)$ in M_n and $\mathcal{Z}_N^+ = \{1, 2, \dots, N\}$. The parameters of AMs and OBM can be updated iteratively as mentioned in the previous section.

$$\psi_0(\boldsymbol{\vartheta}_0) = \frac{\boldsymbol{\theta}_0^H (\mathbf{D}^{-1} - \boldsymbol{\Theta}_0)^{-1} \boldsymbol{\theta}_0}{4} - \ln \det (\mathbf{D}^{-1} - \boldsymbol{\Theta}_0) + M \ln \pi \quad (40a)$$

$$\begin{aligned} \psi_n(\boldsymbol{\vartheta}_n) = & \frac{1}{4} \left(\boldsymbol{\theta}_n + \frac{2\gamma_n y_n}{\sigma_z^2} \right)^H \left(\mathbf{D}^{-1} - \boldsymbol{\Theta}_n + \frac{\gamma_n \gamma_n^H}{\sigma_z^2} \right)^{-1} \left(\boldsymbol{\theta}_n + \frac{2\gamma_n y_n}{\sigma_z^2} \right) - \\ & \ln \det \left(\mathbf{D}^{-1} - \boldsymbol{\Theta}_n + \frac{\gamma_n \gamma_n^H}{\sigma_z^2} \right) + M \ln \pi, n \in \mathcal{Z}_N^+ \end{aligned} \quad (40b)$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_n \left(\frac{\gamma_n y_n}{\sigma_z^2} + \frac{\boldsymbol{\theta}_n}{2} \right) \quad (44a)$$

$$\boldsymbol{\Sigma}_n = \left(\mathbf{D}^{-1} - \boldsymbol{\Theta}_n + \frac{\gamma_n \gamma_n^H}{\sigma_z^2} \right)^{-1} \stackrel{(a)}{=} (\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} - \frac{(\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} \gamma_n \gamma_n^H (\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1}}{\sigma_z^2 + \gamma_n^H (\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} \gamma_n} \quad (44b)$$

$$\boldsymbol{\theta}_{0n} = \left(\mathbf{I} - \frac{(\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} \mathbf{I} \odot (\gamma_n \gamma_n^H)}{\sigma_z^2 + \gamma_n^H (\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} \gamma_n} \right)^{-1} \left(\frac{2y_n - \gamma_n^H (\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} \boldsymbol{\theta}_n}{\sigma_z^2 + \gamma_n^H (\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} \gamma_n} \gamma_n + \boldsymbol{\theta}_n \right) \quad (47a)$$

$$\boldsymbol{\Theta}_{0n} = \mathbf{D}^{-1} - \left[(\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} - \frac{((\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1})^2 \mathbf{I} \odot (\gamma_n \gamma_n^H)}{\sigma_z^2 + \gamma_n^H (\mathbf{D}^{-1} - \boldsymbol{\Theta}_n)^{-1} \gamma_n} \right]^{-1} \quad (47b)$$

Nevertheless, the algorithm updated in this way might be unstable and even divergent. Inspired by the analysis of the equilibrium in [16], [17], we update the parameters of AMs and OBM as

$$\boldsymbol{\vartheta}_n^{t+1} = \alpha \sum_{n' \neq n} \boldsymbol{\xi}_{n'}^t + (1 - \alpha) \boldsymbol{\vartheta}_n^t, n \in \mathcal{Z}_N^+, \quad (49a)$$

$$\boldsymbol{\vartheta}_0^{t+1} = \alpha \sum_{n=1}^N \boldsymbol{\xi}_n^t + (1 - \alpha) \boldsymbol{\vartheta}_0^t, \quad (49b)$$

$$\boldsymbol{\xi}_n^t = \boldsymbol{\vartheta}_{0n}^t - \boldsymbol{\vartheta}_n^t, n \in \mathcal{Z}_N^+ \quad (49c)$$

where $\alpha \leq 1$ is a positive coefficient and $n \in \mathcal{Z}_N^+$ is the same as above. For the case with $\alpha = 1$, the above update process is equivalent to the original algorithm. We will discuss the role of α on the stability of the IGA in the next subsection under the framework of information geometry.

We summarize the IGA in Algorithm 1. The computational complexity of the IGA is $\mathcal{O}(TNM)$, where T is the number of the iterations, $N = N_r N_p$ is the product of the number of antennas at the BS and the number of transmitting subcarriers, and M is the number of variables to be estimated for the beam domain channel. Comparing to the original computational complexity $\mathcal{O}(M^3 + M^2 N)$, when T is small and M is comparable to N , $\mathcal{O}(TNM)$ is about the order of M^2 .

Given $p(\mathbf{h}) \in M_{or}$ and its m -projection $p_0(\mathbf{h}; \boldsymbol{\vartheta}_0) \in M_0$. Denote the means and covariance matrices of those two distributions as $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, respectively. Then, from Theorem 1, the following relationship holds

$$\boldsymbol{\mu}_0 = \boldsymbol{\mu}, \boldsymbol{\Sigma}_0 = \mathbf{I} \odot \boldsymbol{\Sigma}. \quad (50)$$

Recall that the m -projection keeps the expectation of \mathbf{t}_h invariant. We have defined $\mathbf{t}_h = \mathbf{f}(\mathbf{h}, \mathbf{I} \odot (\mathbf{h}\mathbf{h}^H))$ in the

Algorithm 1: IGA for Channel Estimation

Input: The covariance \mathbf{D} of the priori distribution $p(\mathbf{h})$, the received signal \mathbf{y} , the noise power σ_z^2 and the maximal iteration number t_{\max} .

Initialization: set $t = 0$, set α , initialize the parameters of $p_n(\mathbf{h}; \boldsymbol{\vartheta}_n^t)$, $n = 0, 1, \dots, N$;

repeat

 Calculate the m -projection as (47);

 Update the parameters of AMs and OBM as (49a) and (49b), respectively;

$t = t + 1$;

until Convergence or $t > t_{\max}$;

Output: The mean and variance of the approximate marginal, $p(h_i|\mathbf{y})$, $i = 1, 2, \dots, M$, are given by the i -th component of $\boldsymbol{\mu}_0$ and $\text{diag}(\boldsymbol{\Sigma}_0)$, respectively, where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ of $p_0(\mathbf{h}; \boldsymbol{\vartheta}_0^t)$ are given by (43a) and (43b), respectively.

above. The expectation of \mathbf{t}_h w.r.t. $p_0(\mathbf{h}; \boldsymbol{\vartheta}_0)$ is the mean and variance of \mathbf{h} , respectively. Then, from the properties of the m -projection mentioned above, the expectation of \mathbf{t}_h w.r.t. $p(\mathbf{h})$ should be equal to that of p_0 , which is equivalent to (50).

B. Stability Analysis

In this subsection, we present an analysis for the equilibrium, i.e., fixed point or limit point, of the IGA. We first prove that the mean of $p_0(\mathbf{h}; \boldsymbol{\vartheta}_0^*)$ at the equilibrium is equivalent to the *a posteriori* mean in (21a). Then, we give the update relation of the perturbations at the equilibrium and stability condition for the equilibrium through the update relation.

Define a dynamical system as $\omega^{t+1} = \boldsymbol{\varrho}(\omega^t)$, where $\omega \in \mathbb{C}^n$ and $\boldsymbol{\varrho} : \mathbb{C}^n \rightarrow \mathbb{C}^n$. Then, ω^* is an equilibrium (fixed point) of the system if $\omega^* = \boldsymbol{\varrho}(\omega^*)$. The equilibrium ω^* is stable if for every $\epsilon > 0$, there exists a $\zeta = \zeta(\epsilon) > 0$ such that $\|\omega^{t_0+t} - \omega^*\| < \epsilon, \forall t > 0$ whenever the initial point ω^{t_0} satisfies $\|\omega^{t_0} - \omega^*\| < \zeta$. ω^* is asymptotically stable if it is stable and there exists an $\varepsilon > 0$ such that if $\|\omega^{t_0} - \omega^*\| < \varepsilon$, then $\lim_{t \rightarrow \infty} \|\omega^{t_0+t} - \omega^*\| \rightarrow 0$ [39]. Briefly speaking, a (asymptotically) stable equilibrium attracts its nearby orbits. One of the common methods for stability analysis is adding perturbations at the equilibrium and find the update relation of the perturbations through $\Delta\omega^{t+1} = \boldsymbol{\varrho}(\omega^* + \Delta\omega^t) - \boldsymbol{\varrho}(\omega^*) = \boldsymbol{\nu}(\Delta\omega^t)$. We can then determine the stability through the partial derivative of the $\boldsymbol{\nu}$. Moreover, if the perturbation and its update satisfy a linear relation, i.e., $\Delta\omega^{t+1} = \mathbf{J}\Delta\omega^t$, the equilibrium is asymptotically stable if and only if $\rho(\mathbf{J}) < 1$, where $\rho(\mathbf{J})$ is the spectral radius of \mathbf{J} .

By solving the equilibrium equation of (49), i.e., letting $\boldsymbol{\vartheta}_n^{t+1} = \boldsymbol{\vartheta}_n^t = \boldsymbol{\vartheta}_n^*$, $n \in \mathcal{Z}_N$, and $\boldsymbol{\xi}_n^t = \boldsymbol{\xi}_n^*$, $n \in \mathcal{Z}_N^+$, we can obtain

$$\boldsymbol{\vartheta}_0^* = \sum_{n=1}^N \boldsymbol{\xi}_n^* = \boldsymbol{\vartheta}_n^* + \boldsymbol{\xi}_n^*, n \in \mathcal{Z}_N^+, \quad (51a)$$

$$\boldsymbol{\vartheta}_n^* = \sum_{n' \neq n} \boldsymbol{\xi}_{n'}^*, n \in \mathcal{Z}_N^+, \quad (51b)$$

$$\boldsymbol{\xi}_n^* = \boldsymbol{\vartheta}_{0n}^* - \boldsymbol{\vartheta}_n^*, n \in \mathcal{Z}_N^+, \quad (51c)$$

where \mathcal{Z}_N and \mathcal{Z}_N^+ are the same as before. From (51a) and (51b), we have $\sum_{n=1}^N \boldsymbol{\vartheta}_n^* = (N-1)\boldsymbol{\vartheta}_0^*$. Substituting (51c) into the second equation of (51a), we can obtain $\boldsymbol{\vartheta}_0^* = \boldsymbol{\vartheta}_{0n}^*$, $n \in \mathcal{Z}^+$. Thus, we can obtain two conditions of the equilibrium,

$$\begin{cases} \boldsymbol{\vartheta}_0^* = \boldsymbol{\vartheta}_{0n}^*, n \in \mathcal{Z}^+, \\ \boldsymbol{\vartheta}_0^* = \frac{1}{N-1} \sum_{n=1}^N \boldsymbol{\vartheta}_n^*, \end{cases} \quad (52)$$

which are referred as m -condition and e -condition, respectively. Through the two conditions, we have the following theorem.

Theorem 2. *At the equilibrium of IGA, the mean of $p_0(\mathbf{h}; \boldsymbol{\vartheta}_0^*)$ is equal to that of the a posteriori distribution $p(\mathbf{h}|\mathbf{y})$.*

Proof. See in Appendix B. \square

It should be noted that the above relationship holds for arbitrary matrix \mathbf{A} . For the most popular AMP algorithms, the same conclusion can be obtained when \mathbf{A} is a zero-mean i.i.d. sub-Gaussian matrix satisfying the large-system limit, and the fixed point (equilibrium) of AMP is unique [40].

We then perform the stability analysis of IGA through the update relation $\{\Delta\boldsymbol{\vartheta}_n^{t+1}\}_{n=1}^N = \boldsymbol{\nu}(\{\Delta\boldsymbol{\vartheta}_n^t\}_{n=1}^N)$. Since the forms of (47a) and (47b) in the IGA are complicated, it is difficult to obtain the update relationships directly. We solve this problem by using the properties of the defined manifolds and the m -projections. We find that the perturbations in $\{\boldsymbol{\vartheta}_n^t\}$ and its updates in $\{\boldsymbol{\vartheta}_n^{t+1}\}$ satisfy a linear relationship. We first give the definition of the expectation parameter of OBM and AMs. Let $\boldsymbol{\chi}_n = \left[\boldsymbol{\theta}_n^T \bar{\boldsymbol{\theta}}_n^T \text{diag}(\boldsymbol{\Theta}_n)^T \right]^T \in \mathbb{C}^{3M \times 1}$, $n \in \mathcal{Z}_N$,

where $\mathcal{Z}_N = \{0, 1, \dots, N\}$. We then rewrite the free energy $\psi_n(\boldsymbol{\vartheta}_n)$, $n \in \mathcal{Z}_N$, as $\psi_n(\boldsymbol{\chi}_n)$ since the free variables of the free energy are $\boldsymbol{\theta}_n$, $\text{diag}(\boldsymbol{\Theta}_n)$, and $\boldsymbol{\theta}_n$ is a complex vector. The expectation parameter $\boldsymbol{\eta}_n(\boldsymbol{\chi}_n)$ of $p_n(\mathbf{h}; \boldsymbol{\vartheta}_n)$, $n \in \mathcal{Z}_N$, are then defined as follows,

$$\boldsymbol{\eta}_n(\boldsymbol{\chi}_n) \triangleq \left[\begin{array}{c} \frac{1}{2} \mathbb{E}_{p_n(\mathbf{h}; \boldsymbol{\vartheta}_n)} \{\mathbf{h}\} \\ \frac{1}{2} \mathbb{E}_{p_n(\mathbf{h}; \boldsymbol{\vartheta}_n)} \{\bar{\mathbf{h}}\} \\ \mathbb{E}_{p_n(\mathbf{h}; \boldsymbol{\vartheta}_n)} \{\mathbf{h} \odot \bar{\mathbf{h}}\} \end{array} \right] \stackrel{(a)}{=} \frac{\partial \psi_n(\boldsymbol{\chi}_n)}{\partial \boldsymbol{\chi}_n} = \left[\begin{array}{c} \frac{\partial \psi_n(\boldsymbol{\chi}_n)}{\partial \boldsymbol{\theta}_n} \\ \frac{\partial \psi_n(\boldsymbol{\chi}_n)}{\partial \text{diag}(\boldsymbol{\Theta}_n)} \\ \frac{\partial \psi_n(\boldsymbol{\chi}_n)}{\partial \text{diag}(\boldsymbol{\Theta}_n)^T} \end{array} \right], \quad (53)$$

where the process of (a) is given in Appendix C. Then, the Fisher information (FI) \mathcal{I} is defined as $\mathcal{I}_n(\boldsymbol{\chi}_n) \triangleq \frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \boldsymbol{\chi}_n \partial \boldsymbol{\chi}_n^T} = \frac{\partial \boldsymbol{\eta}_n(\boldsymbol{\chi}_n)}{\partial \boldsymbol{\chi}_n^T}$, $n \in \mathcal{Z}_N$, and is given by

$$\mathcal{I}_n(\boldsymbol{\chi}_n) = \left[\begin{array}{ccc} \frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \boldsymbol{\theta}_n \partial \boldsymbol{\theta}_n^T} & \frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \boldsymbol{\theta}_n \partial \text{diag}(\boldsymbol{\Theta}_n)^T} & \frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \text{diag}(\boldsymbol{\Theta}_n) \partial \text{diag}(\boldsymbol{\Theta}_n)^T} \\ \frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \boldsymbol{\theta}_n \partial \text{diag}(\boldsymbol{\Theta}_n)^T} & \frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \text{diag}(\boldsymbol{\Theta}_n) \partial \boldsymbol{\theta}_n^H} & \frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \text{diag}(\boldsymbol{\Theta}_n) \partial \text{diag}(\boldsymbol{\Theta}_n)^T} \\ \frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \text{diag}(\boldsymbol{\Theta}_n) \partial \boldsymbol{\theta}_n^H} & \frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \text{diag}(\boldsymbol{\Theta}_n) \partial \text{diag}(\boldsymbol{\Theta}_n)^T} & \frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \text{diag}(\boldsymbol{\Theta}_n) \partial \text{diag}(\boldsymbol{\Theta}_n)^T} \end{array} \right]. \quad (54)$$

For notational convenience, we define $\tilde{\boldsymbol{\theta}}_n \triangleq \text{diag}(\boldsymbol{\Theta}_n)$. Then, the sub-matrices of $\mathcal{I}_n(\boldsymbol{\chi}_n)$ can be calculated through the higher-order moments of the complex Gaussian distribution [41] and are given by (55). Also, the process of the equivalence between the Hessian and expectations can be found in Appendix C. Substituting (55) into (54) and after some calculation, $\mathcal{I}_n(\boldsymbol{\chi}_n)$, $n \in \mathcal{Z}_N$ is given by

$$\mathcal{I}_n(\boldsymbol{\chi}_n) = \mathbf{P}_n^H \mathbf{B} \text{diag} \left(\frac{\boldsymbol{\Sigma}_n}{4}, \frac{\bar{\boldsymbol{\Sigma}}_n}{4}, \boldsymbol{\Sigma}_n \odot \bar{\boldsymbol{\Sigma}}_n \right) \mathbf{P}_n, \quad (56)$$

where

$$\mathbf{P}_n = \left[\begin{array}{ccc} \mathbf{I} & \mathbf{0} & 2 \text{diag}(\boldsymbol{\mu}_n) \\ \mathbf{0} & \mathbf{I} & 2 \text{diag}(\bar{\boldsymbol{\mu}}_n) \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{array} \right]. \quad (57)$$

From [38, 11.46 (e)], if $\mathbf{A} \in \mathbb{C}^{n \times n}$ and $\mathbf{B} \in \mathbb{C}^{n \times n}$ are positive definite, $\mathbf{A} \odot \mathbf{B}$ is positive definite. Then, $\boldsymbol{\Sigma}_n \odot \bar{\boldsymbol{\Sigma}}_n$ is invertible. Thus, it can be shown that $\mathcal{I}_n(\boldsymbol{\chi}_n)$ is invertible.

At the equilibrium of IGA, let a small perturbation in $\boldsymbol{\chi}_n$ and its corresponding change in $\boldsymbol{\chi}_{0n}$ be denoted by $\Delta\boldsymbol{\chi}_n$ and $\Delta\boldsymbol{\chi}_{0n}$, respectively. Then, we have the following result.

Theorem 3. *For the m -projection, the following relationship holds*

$$\Delta\boldsymbol{\chi}_{0n} = \mathbf{K}_n \Delta\boldsymbol{\chi}_n, n \in \mathcal{Z}^+, \quad (58)$$

where $\mathcal{Z}^+ = \{1, 2, \dots, N\}$, $\boldsymbol{\chi}_{0n} = \left[\boldsymbol{\theta}_{0n}^T \bar{\boldsymbol{\theta}}_{0n}^T \text{diag}(\boldsymbol{\Theta}_{0n})^T \right]^T$, $\boldsymbol{\chi}_n = \left[\boldsymbol{\theta}_n^T \bar{\boldsymbol{\theta}}_n^T \text{diag}(\boldsymbol{\Theta}_n)^T \right]^T$, $\mathbf{K}_n = \mathbf{P}_n^{-1} \mathbf{Q}_n \mathbf{P}_n$, and

$$\mathbf{Q}_n = \mathbf{B} \text{diag}(\boldsymbol{\Sigma}_{0n}^{-1} \boldsymbol{\Sigma}_n, \boldsymbol{\Sigma}_{0n}^{-1} \bar{\boldsymbol{\Sigma}}_n, (\boldsymbol{\Sigma}_{0n} \odot \bar{\boldsymbol{\Sigma}}_{0n})^{-1} \boldsymbol{\Sigma}_n \odot \bar{\boldsymbol{\Sigma}}_n). \quad (59)$$

Proof. See in Appendix D. \square

Review the update of $\boldsymbol{\vartheta}_n$,

$$\boldsymbol{\vartheta}_n^{t+1} = \alpha \sum_{n' \neq n} (\boldsymbol{\vartheta}_{0n'}^t - \boldsymbol{\vartheta}_{n'}^t) + (1 - \alpha) \boldsymbol{\vartheta}_n^t, \quad (60)$$

$$\frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \bar{\boldsymbol{\theta}}_n \partial \bar{\boldsymbol{\theta}}_n^T} = \overline{\left(\frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \boldsymbol{\theta}_n \partial \boldsymbol{\theta}_n^H} \right)} = \mathbb{E}_{p_n} \left\{ \left(\frac{1}{2} \mathbf{h} - \frac{1}{2} \mathbb{E}_{p_n} \{ \mathbf{h} \} \right) \left(\frac{1}{2} \mathbf{h} - \frac{1}{2} \mathbb{E}_{p_n} \{ \mathbf{h} \} \right)^H \right\} = \frac{1}{4} \boldsymbol{\Sigma}_n \quad (55a)$$

$$\frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \bar{\boldsymbol{\theta}}_n \partial \bar{\boldsymbol{\theta}}_n^H} = \overline{\left(\frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \boldsymbol{\theta}_n \partial \boldsymbol{\theta}_n^T} \right)} = \mathbb{E}_{p_n} \left\{ \left(\frac{1}{2} \mathbf{h} - \frac{1}{2} \mathbb{E}_{p_n} \{ \mathbf{h} \} \right) \left(\frac{1}{2} \mathbf{h} - \frac{1}{2} \mathbb{E}_{p_n} \{ \mathbf{h} \} \right)^T \right\} = \mathbf{O} \quad (55b)$$

$$\begin{aligned} \frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \bar{\boldsymbol{\theta}}_n \partial \bar{\boldsymbol{\theta}}_n^T} &= \overline{\left(\frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \boldsymbol{\theta}_n \partial \bar{\boldsymbol{\theta}}_n^T} \right)} = \left(\frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \bar{\boldsymbol{\theta}}_n \partial \boldsymbol{\theta}_n^T} \right)^H = \left(\frac{\partial^2 \psi_n(\boldsymbol{\chi}_n)}{\partial \bar{\boldsymbol{\theta}}_n \partial \boldsymbol{\theta}_n^H} \right)^T \\ &= \frac{1}{2} \mathbb{E}_{p_n} \left\{ (\mathbf{h} - \mathbb{E}_{p_n} \{ \mathbf{h} \}) (\mathbf{h} \odot \bar{\mathbf{h}} - \mathbb{E}_{p_n} \{ \mathbf{h} \odot \bar{\mathbf{h}} \})^T \right\} = \frac{1}{2} \boldsymbol{\Sigma}_n \text{diag}(\boldsymbol{\mu}_n) \end{aligned} \quad (55c)$$

$$\begin{aligned} \frac{\partial \psi_n(\boldsymbol{\chi}_n)}{\partial \bar{\boldsymbol{\theta}}_n \partial \bar{\boldsymbol{\theta}}_n^T} &= \mathbb{E}_{p_n} \left\{ (\mathbf{h} \odot \bar{\mathbf{h}} - \mathbb{E}_{p_n} \{ \mathbf{h} \odot \bar{\mathbf{h}} \}) (\mathbf{h} \odot \bar{\mathbf{h}} - \mathbb{E}_{p_n} \{ \mathbf{h} \odot \bar{\mathbf{h}} \})^T \right\} \\ &= \boldsymbol{\Sigma}_n \odot \bar{\boldsymbol{\Sigma}}_n + \text{diag}(\boldsymbol{\mu}_n)^H \boldsymbol{\Sigma}_n \text{diag}(\boldsymbol{\mu}_n) + \text{diag}(\boldsymbol{\mu}_n) \boldsymbol{\Sigma}_n^T \text{diag}(\boldsymbol{\mu}_n)^H \end{aligned} \quad (55d)$$

where $\boldsymbol{\vartheta}_{0n}^t$ is the m -projection of the $p_n(\mathbf{h}; \boldsymbol{\vartheta}_{0n}^t)$. From Theorem 3, a small perturbation at the equilibrium is updated in $\boldsymbol{\vartheta}_n^{t+1}$ as

$$\Delta \boldsymbol{\chi}_n^{t+1} = \alpha \sum_{n' \neq n} (\mathbf{K}_{n'} - \mathbf{I}) \Delta \boldsymbol{\chi}_{n'}^t + (1 - \alpha) \Delta \boldsymbol{\chi}_n^t, \quad n \in \mathcal{Z}^+, \quad (61)$$

which can be organized as

$$\begin{aligned} & [(\Delta \boldsymbol{\chi}_1^{t+1})^T \ (\Delta \boldsymbol{\chi}_2^{t+1})^T \ \dots \ (\Delta \boldsymbol{\chi}_N^{t+1})^T]^T \\ &= \mathbf{M} [(\Delta \boldsymbol{\chi}_1^t)^T \ (\Delta \boldsymbol{\chi}_2^t)^T \ \dots \ (\Delta \boldsymbol{\chi}_N^t)^T]^T, \end{aligned} \quad (62)$$

where $\mathbf{M} = \alpha \mathbf{N} + (1 - \alpha) \mathbf{I}$ and

$$\mathbf{N} = \begin{bmatrix} \mathbf{O} & \mathbf{K}_2 - \mathbf{I} & \dots & \mathbf{K}_N - \mathbf{I} \\ \mathbf{K}_1 - \mathbf{I} & \mathbf{O} & \dots & \mathbf{K}_N - \mathbf{I} \\ \vdots & \dots & \ddots & \vdots \\ \mathbf{K}_1 - \mathbf{I} & \dots & \dots & \mathbf{O} \end{bmatrix}. \quad (63)$$

The spectral radius of \mathbf{M} depends on the eigenvalues of \mathbf{N} and the size of α . The equilibrium of the IGA is asymptotically stable when all the eigenvalues of \mathbf{M} , $\lambda_i, i = 1, 2, \dots, 3NM$, satisfy $|\lambda_i| < 1$. From the simulation results, we will see that a small α can improve the stability of the IGA, and the number of iterations required for the convergence is still relatively small. We notice that in the Gaussian belief propagation algorithm proposed in [42], [43], a similar approach to have the algorithm convergence is adopted, and its effectiveness is confirmed through numerical simulations.

V. SIMULATION RESULTS

This section provides some simulation results to illustrate the performance of the proposed information geometry approach for massive MIMO-OFDM channel estimation. We adopt the widely used QuaDRiGa channel model [44]. The main parameters for the simulations are summarized in Table I. The simulation scenario is set to "3GPP_38.901_UMa_NLOS"¹.

The layout of the massive MIMO-OFDM system is plotted in Fig. 1. The BS is located at $(0, 0, 25)$. The users are

¹Note that channels under either LOS or NLOS conditions can be modeled with SF beam based channel model.

TABLE I
PARAMETER SETTINGS OF THE QUADRiGA

Parameter	Value
Number of BS antenna $N_{r,v} \times N_{r,h}$	8×16
UT number K	48
Center frequency f_c	4.8GHz
Number of training subcarriers N_p	360
Subcarrier spacing Δ_f	15kHz
Number of subcarriers N_C	2048
CP length N_g	144
Fine factors F_h, F_v, F_τ	1, 2 or 4

randomly generated in a 120° sector with radius $r = 200m$ around $(0, 0, 1.5)$. We normalize the channel as $\mathbb{E} \{ \|\mathbf{G}_k\|_F^2 \} = N_r N_p$. We adopt the adjustable phase shift pilots [24] as the training signal. The transmit power of the training signal for each user is set to 1. It should be noted that any other training signal can be applied. The SNR is set as $\text{SNR} = \frac{1}{\sigma_z^2}$. Furthermore, we use the algorithm proposed in [29] to obtain the channel power matrices $\boldsymbol{\Omega}_k, \forall k$. The normalized mean-squared error (NMSE) is used as the performance metric for the channel estimation,

$$\text{NMSE} = \frac{1}{KN_{sam}} \sum_{k=1}^K \sum_{n=1}^{N_{sam}} \frac{\|\mathbf{G}_k^{(n)} - \hat{\mathbf{G}}_k^{(n)}\|_F^2}{\|\mathbf{G}_k^{(n)}\|_F^2} \quad (64)$$

where N_{sam} is the number of the channel samples, $\mathbf{G}_k^{(n)}$ is the n -th channel sample of user k , $\hat{\mathbf{G}}_k^{(n)}$ is the estimate of the $\mathbf{G}_k^{(n)}$ and $\|\cdot\|_F$ is the F-norm. We set $N_{sam} = 200$ in our simulations. We compare the proposed IGA with the following algorithms.

GAMP: Generalized approximate message passing algorithm proposed in [8].

VEP: A low-complexity variant of the EP algorithm proposed in [45].

MMSE: The MMSE estimation of the beam domain channels based on (21a).

The computational complexities of both GAMP and VEP are $\mathcal{O}(NM)$ for each iteration [45], which is the same as that of IGA.

A. Effect of Fine Factors

Fig. 2 shows the NMSE performance comparisons between different fine factor settings. The performance of the MMSE estimation is shown. It can be found that the performance of channel estimation improves gradually as the fine factors increase. Setting fine factors to 2 brings a substantial performance gain compared to setting them to 1, which illustrates that increasing the fine factors is necessary for obtaining accurate channel estimates. Meanwhile, $F_v = F_h = F_\tau = 4$ brings a performance gain of only around 1 dB compared to $F_v = F_h = F_\tau = 2$ when SNR = 30dB.

B. IGA Performance

Fig. 3 shows the NMSE performance of IGA channel estimation compared with GAMP, VEP and MMSE with $F_v = F_h = F_\tau = 2$. The maximal iteration number of IGA, GAMP and VEP is set as 100. It can be found that the IGA can obtain almost the same NMSE performance as the MMSE estimation at all SNRs. The SNR gain of the IGA compared to GAMP and VEP is about 5dB when the NMSE performance is -29dB.

Fig. 4 and Fig. 5 shows the convergence performance of IGA channel estimation compared with GAMP and VEP, where the SNR is set as 10dB and 20dB respectively, and $F_v = F_h = F_\tau = 2$. We can find that in the case with SNR = 10dB, IGA requires about 150 iterations to converge and achieves the optimal solution as that by the MMSE estimation, while the GAMP and VEP converge in more than 450 iterations. In the case with SNR = 20dB, IGA converge in about 200 iterations, while GAMP and VEP take more than 550 iterations to converge. We can also find that VEP and GAMP show almost identical convergence behavior in all the simulations. This might be caused by the similarity in the processes of VEP and GAMP. The key difference between VEP and GAMP lies in the computation of one set of intermediate variables, which is denoted as $\{\tilde{\tau}_{m,n}\}$ in [45]. Other than that, they are nearly identical to each other [45]. Apart from the simulations in this work, VEP and GAMP also show almost identical convergence behavior in the simulations of [45], see, e.g., Fig. 3 (b) therein. Compared with GAMP and VEP, IGA has a faster convergence rate. The proposed IGA is derived based on the structure of the *a posteriori* distribution $p(\mathbf{h}|\mathbf{y})$ within the information geometry framework. The geometrical perspective provides an intuitive understanding of the statistical model, and thus allows to solve the statistical inference problem from an intrinsic and general standpoint. This might be a key reason for the improved convergence behavior of IGA compared with GAMP and VEP in massive MIMO channel estimation.

Fig.6 shows the effect of α on the IGA performance. We set the number of users to 4 or 48, and the remaining parameters are the same as above. The SNR is set as 20dB. From the figure, we can find that when $K = 4$, the IGA converges and obtains the same NMSE performance for both $\alpha = 1$ and $\alpha = 0.1$. When $K = 48$, the IGA diverges for $\alpha = 1$, while for $\alpha = 0.025$, the stability of the IGA is significantly improved.

VI. CONCLUSION

We have proposed an information geometry approach for channel estimation in massive MIMO-OFDM systems. We first derive the SF beam based statistical channel model for massive MIMO-OFDM systems by using sampled steering vectors in space and frequency domain. The accuracy of the beam based channel model is guaranteed by sufficiently large number of sampled steering vectors. With the established channel model, the channel estimation is formulated as calculating the *a posteriori* information of the beam domain channel. We calculate approximate marginals of the *a posteriori* distribution within the information geometry framework. Specifically, the calculation of the marginals is formulated as an iterative *m*-projection process. We derive the IGA for channel estimation by finding the solution of the *m*-projection. We show that the mean of the obtained marginals at the IGA's equilibrium equals the *a posteriori* mean. Simulation results verify that the proposed IGA can obtain high channel estimation accuracy with much less number of iterations compared with the existing approaches. This demonstrates the superiority of our proposed channel estimation approach for massive MIMO-OFDM systems.

APPENDIX A PROOF OF THEOREM 1

The *m*-projection is calculated by

$$\boldsymbol{\vartheta}_0^* = \arg \min_{\boldsymbol{\vartheta}_0} D_{KL} \{p(\mathbf{h}; \boldsymbol{\vartheta}) : p_0(\mathbf{h}; \boldsymbol{\vartheta}_0)\}. \quad (65)$$

Given $p(\mathbf{h}; \boldsymbol{\vartheta})$, its entropy is a constant, and hence, we can express the K-L divergence as $D_{KL} \{p : p_0\} = c_p - \mathbb{E}_p \{\ln p_0\}$. Then, $\mathbb{E}_p \{\ln p_0\}$ can be expressed as

$$\begin{aligned} & \mathbb{E}_p \{\ln p_0\} \\ &= \mathbb{E}_p \left\{ \mathbf{h}^H (-\mathbf{D}^{-1} + \boldsymbol{\Theta}_0) \mathbf{h} + \frac{1}{2} \mathbf{h}^H \boldsymbol{\theta}_0 + \frac{1}{2} \boldsymbol{\theta}_0^H \mathbf{h} - \psi_0 \right\} \\ &= \mathbb{E}_p \left\{ \text{tr} \{ (-\mathbf{D}^{-1} + \boldsymbol{\Theta}_0) \mathbf{h} \mathbf{h}^H \} \right\} + \frac{\boldsymbol{\mu}^H \boldsymbol{\theta}_0 + \boldsymbol{\theta}_0^H \boldsymbol{\mu}}{2} - \psi_0 \\ &= -\text{tr} \{ (\mathbf{D}^{-1} - \boldsymbol{\Theta}_0) (\boldsymbol{\mu} \boldsymbol{\mu}^H + \boldsymbol{\Sigma}) \} + \frac{\boldsymbol{\mu}^H \boldsymbol{\theta}_0 + \boldsymbol{\theta}_0^H \boldsymbol{\mu}}{2} - \\ & \quad \frac{\boldsymbol{\theta}_0^H (\mathbf{D}^{-1} - \boldsymbol{\Theta}_0)^{-1} \boldsymbol{\theta}_0}{4} + \ln \det (\mathbf{D}^{-1} - \boldsymbol{\Theta}_0) - M \ln \pi, \end{aligned} \quad (66)$$

where ψ_0 is given by (40a). Hence, the partial derivatives of $\mathbb{E}_p \{\ln p_0\}$ are given by (67). By setting the partial derivatives equal to zero, we have the solution of *m*-projection,

$$\begin{cases} \boldsymbol{\theta}_0^* = 2 (\mathbf{I} \odot \boldsymbol{\Sigma})^{-1} \boldsymbol{\mu}, \\ \boldsymbol{\Theta}_0^* = \mathbf{D}^{-1} - (\mathbf{I} \odot \boldsymbol{\Sigma})^{-1}. \end{cases} \quad (68)$$

This completes the proof.

APPENDIX B PROOF OF THEOREM 2

From the *m*-condition, we have

$$\boldsymbol{\vartheta}_0^* = \boldsymbol{\vartheta}_{0n}^* = \pi_{M_0} \{p_n(\mathbf{h}; \boldsymbol{\vartheta}_n^*)\}, n \in \mathcal{Z}_N^+, \quad (69)$$

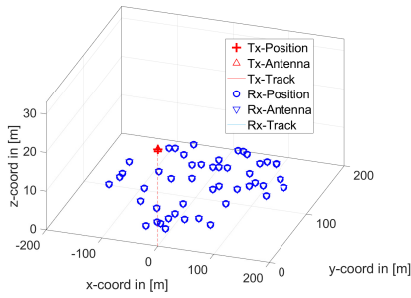


Fig. 1. The layout of the massive MIMO-OFDM system.

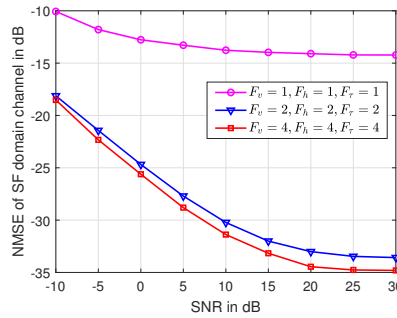


Fig. 2. NMSE versus SNR for different fine factors.

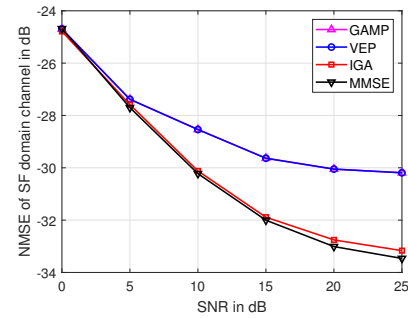


Fig. 3. NMSE performance of IGA channel estimation compared with GAMP, VEP and MMSE.

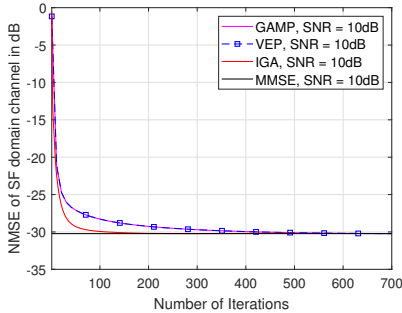


Fig. 4. Convergence performance of IGA channel estimation compared with GAMP and VEP at SNR = 10 dB.

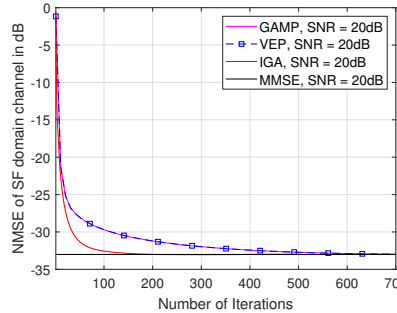


Fig. 5. Convergence performance of IGA channel estimation compared with GAMP and VEP at SNR = 20 dB.

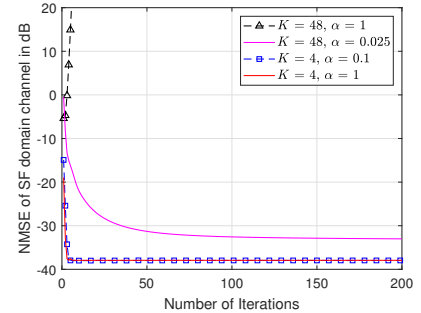


Fig. 6. Effect of α on the performance of IGA channel estimation.

$$\frac{\partial \mathbb{E}_{p_G} \{\ln p_0\}}{\partial \boldsymbol{\theta}_0} = \frac{\boldsymbol{\mu}}{2} - \frac{(\mathbf{D}^{-1} - \boldsymbol{\Theta}_0)^{-1} \boldsymbol{\theta}_0}{4} \quad (67a)$$

$$\frac{\partial \mathbb{E}_{p_G} \{\ln p_0\}}{\partial \boldsymbol{\Theta}_0} = \mathbf{I} \odot \left[\boldsymbol{\mu} \boldsymbol{\mu}^H + \boldsymbol{\Sigma} - (\mathbf{D}^{-1} - \boldsymbol{\Theta}_0)^{-1} - \frac{(\mathbf{D}^{-1} - \boldsymbol{\Theta}_0)^{-1} \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^H (\mathbf{D}^{-1} - \boldsymbol{\Theta}_0)^{-1}}{4} \right]^T \quad (67b)$$

where $\pi_{M_0} \{\cdot\}$ denotes the m -projection onto OBM. Then, from Theorem 1 and (50), we can obtain

$$\boldsymbol{\mu}_0^* = \boldsymbol{\mu}_{0n}^* = \boldsymbol{\mu}_n^*, \quad \boldsymbol{\Sigma}_0^* = \boldsymbol{\Sigma}_{0n}^* = \mathbf{I} \odot \boldsymbol{\Sigma}_n^*. \quad (70)$$

Then, from the e -condition we have

$$\begin{aligned} & \boldsymbol{\theta}_0^* \\ &= \frac{1}{N-1} \sum_n \boldsymbol{\theta}_n^* \stackrel{(a)}{=} \frac{2}{N-1} \sum_n \left((\boldsymbol{\Sigma}_n^*)^{-1} \boldsymbol{\mu}_n^* - \frac{\boldsymbol{\gamma}_n y_n}{\sigma_z^2} \right) \\ & \stackrel{(b)}{=} \frac{2}{N-1} \left(\sum_n (\boldsymbol{\Sigma}_n^*)^{-1} \right) \boldsymbol{\mu}_0^* - \frac{2 \mathbf{A}^H \mathbf{y}}{(N-1) \sigma_z^2} \\ & \stackrel{(c)}{=} \frac{2}{N-1} \left(N \mathbf{D}^{-1} - \sum_n \boldsymbol{\Theta}_n^* + \frac{\mathbf{A}^H \mathbf{A}}{\sigma_z^2} \right) \boldsymbol{\mu}_0^* - \frac{2 \mathbf{A}^H \mathbf{y}}{(N-1) \sigma_z^2} \\ & \stackrel{(d)}{=} \frac{2}{N-1} \left[\left(N \mathbf{D}^{-1} - (N-1) \boldsymbol{\Theta}_0^* + \frac{\mathbf{A}^H \mathbf{A}}{\sigma_z^2} \right) \boldsymbol{\mu}_0^* - \frac{\mathbf{A}^H \mathbf{y}}{\sigma_z^2} \right] \\ & \stackrel{(e)}{=} \frac{2}{N-1} \left[\left(\mathbf{D}^{-1} + (N-1) (\boldsymbol{\Sigma}_0^*)^{-1} + \frac{\mathbf{A}^H \mathbf{A}}{\sigma_z^2} \right) \boldsymbol{\mu}_0^* - \frac{\mathbf{A}^H \mathbf{y}}{\sigma_z^2} \right] \end{aligned}$$

$$\stackrel{(f)}{=} \boldsymbol{\theta}_0^* + \frac{2}{N-1} \left(\mathbf{D}^{-1} + \frac{\mathbf{A}^H \mathbf{A}}{\sigma_z^2} \right) \boldsymbol{\mu}_0^* - \frac{2 \mathbf{A}^H \mathbf{y}}{(N-1) \sigma_z^2}, \quad (71)$$

where (a) is from (44a), (b) is from (70) and the definition of $\boldsymbol{\gamma}_n$, (c) is from (44b), (d) is from e -condition, (e) is from (43b) and (f) is from (43a). Thus, we can obtain

$$\begin{aligned} \boldsymbol{\mu}_0^* &= (\mathbf{A}^H \mathbf{A} + \sigma_z^2 \mathbf{D}^{-1})^{-1} \mathbf{A}^H \mathbf{y} \\ &= \mathbf{D} (\mathbf{A}^H \mathbf{A} \mathbf{D} + \sigma_z^2 \mathbf{I})^{-1} \mathbf{A}^H \mathbf{y}. \end{aligned} \quad (72)$$

This completes the proof.

APPENDIX C

EXPECTATION PARAMETER AND FISHER INFORMATION

For OBM and AMs, the distributions can be expressed as

$$p(\mathbf{h}; \boldsymbol{\vartheta}) = \exp \{ \boldsymbol{\vartheta} \circ \mathbf{t}_h + m(\mathbf{h}) - \psi(\boldsymbol{\vartheta}) \}, \quad (73)$$

where $\boldsymbol{\vartheta} = \mathbf{f}(\boldsymbol{\theta}, \boldsymbol{\Theta})$ with $\boldsymbol{\theta} \in \mathbb{C}^M$ and $\boldsymbol{\Theta} \in \mathbb{D}^M$, \mathbb{D}^M is the set of $M \times M$ real diagonal matrices, $m(\mathbf{h})$ is a function

independent of ϑ , $\mathbf{t}_h = \mathbf{f}(\mathbf{h}, \mathbf{I} \odot (\mathbf{h}\mathbf{h}^H))$ and $\psi(\vartheta)$ is the free energy,

$$\psi(\vartheta) = \ln \int \exp \{ \vartheta \circ \mathbf{t}_h + m(\mathbf{h}) \} d\mathbf{h}. \quad (74)$$

Specially, we have $m(\mathbf{h}) = \mathbf{d}_h \circ \mathbf{t}_h$ for OBM and $m(\mathbf{h}) = \mathbf{d}_h \circ \mathbf{t}_h + c_n(\mathbf{h})$ for AMs. From $\chi = [\boldsymbol{\theta}^T, \boldsymbol{\theta}^H, \text{diag}(\boldsymbol{\Theta})^T]^T$, we have $\vartheta \circ \mathbf{t}_h = \chi^H \mathbf{t}'_h$, where ϑ is the same as above and $\mathbf{t}'_h = [\frac{1}{2}\mathbf{h}^T, \frac{1}{2}\mathbf{h}^H, (\mathbf{h} \circ \bar{\mathbf{h}})^T]^T$. We express ψ as $\psi(\chi) = \ln \int \exp \{ \chi^H \mathbf{t}'_h + m(\mathbf{h}) \} d\mathbf{h}$. The derivative of $\psi(\chi)$ is given by

$$\begin{aligned} \frac{\partial \psi(\chi)}{\partial \chi} &= \exp \{ -\psi(\chi) \} \int \mathbf{t}'_h \exp \{ \chi^H \mathbf{t}'_h + m(\mathbf{h}) \} d\mathbf{h} \\ &= \mathbb{E}_p \{ \mathbf{t}'_h \} = \boldsymbol{\eta}(\chi). \end{aligned} \quad (75)$$

Thus, $\boldsymbol{\eta}(\chi)$ is also called the expectation parameter. Similarly, the Hessian is given by

$$\begin{aligned} \mathcal{I}(\chi) &\triangleq \frac{\partial^2 \psi(\chi)}{\partial \chi \partial \chi^T} = \frac{\partial \boldsymbol{\eta}}{\partial \chi^T} = \int \mathbf{t}'_h \frac{\partial p(\mathbf{h}; \vartheta)}{\partial \chi^T} d\mathbf{h} \\ &= \int \mathbf{t}'_h \left(\mathbf{t}'_h{}^H - \frac{\partial \psi}{\partial \chi^T} \right) p(\mathbf{h}; \vartheta) d\mathbf{h} \\ &= \mathbb{E}_p \left\{ \mathbf{t}'_h \left(\mathbf{t}'_h{}^H - \boldsymbol{\eta}^H(\chi) \right) \right\} \\ &= \mathbb{E}_p \left\{ (\mathbf{t}'_h - \boldsymbol{\eta}(\chi)) (\mathbf{t}'_h - \boldsymbol{\eta}(\chi))^H \right\}. \end{aligned} \quad (76)$$

$\mathcal{I}(\chi)$ is also called the Fisher information.

APPENDIX D PROOF OF THEOREM 3

From Theorem 1, we can obtain the following relationship:

$$\boldsymbol{\mu}_{0n} = \boldsymbol{\mu}_n, \quad \boldsymbol{\Sigma}_{0n} = \mathbf{I} \odot \boldsymbol{\Sigma}_n, \quad (77)$$

where $n \in \mathcal{Z}_N^+$ and $\mathcal{Z}_N^+ = \{1, 2, \dots, N\}$. From the definition of expectation parameter, we have

$$\boldsymbol{\eta}_0(\chi_{0n}) = \boldsymbol{\eta}_n(\chi_n), n \in \mathcal{Z}_N^+. \quad (78)$$

With the first-order Taylor series of (78), we have the following relationship [16, (4.8)],

$$\begin{aligned} \boldsymbol{\eta}_0(\chi_{0n}) + \mathcal{I}_0(\chi_{0n}) \Delta \chi_{0n} &\simeq \boldsymbol{\eta}_0(\chi_{0n} + \Delta \chi_{0n}) = \\ \boldsymbol{\eta}_n(\chi_n + \Delta \chi_n) &\simeq \boldsymbol{\eta}_n(\chi_n) + \mathcal{I}_n(\chi_n) \Delta \chi_n, \end{aligned} \quad (79a)$$

$$\mathcal{I}_0(\chi_{0n}) \Delta \chi_{0n} = \mathcal{I}_n(\chi_n) \Delta \chi_n, \quad (79b)$$

where $n \in \mathcal{Z}_N^+$ and \mathcal{Z}_N^+ is the same as above. Thus, we have

$$\Delta \chi_{0n} = \mathbf{K}_n \Delta \chi_n, \quad (80)$$

where $\mathbf{K}_n = \mathcal{I}_0(\chi_{0n})^{-1} \mathcal{I}_n(\chi_n)$, $n \in \mathcal{Z}_N^+$, can be decomposed as (81). This completes the proof.

REFERENCES

- [1] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.
- [2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.
- [3] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
- [4] L. Cimini, "Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing," *IEEE Trans. Commun.*, vol. 33, no. 7, pp. 665–675, Jul. 1985.
- [5] L. Tong, B. Sadler, and M. Dong, "Pilot-assisted wireless transmissions: general model, design criteria, and signal processing," *IEEE Signal Processing Magazine*, vol. 21, no. 6, pp. 12–25, Nov. 2004.
- [6] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.
- [7] T. P. Minka, "Expectation propagation for approximate bayesian inference," *arXiv preprint arXiv:1301.2294*, 2013.
- [8] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *IEEE ISIT, St. Petersburg, Russia, July 31 - August 5, 2011*, pp. 2168–2172.
- [9] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.
- [10] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, Oct. 2019.
- [11] J. P. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4658–4672, Oct. 2013.
- [12] C.-K. Wen, S. Jin, K.-K. Wong, J.-C. Chen, and P. Ting, "Channel estimation for massive MIMO using Gaussian-mixture bayesian learning," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1356–1368, Mar. 2015.
- [13] A. Liu, L. Lian, V. K. N. Lau, and X. Yuan, "Downlink channel estimation in multiuser massive MIMO with hidden markovian sparsity," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4796–4810, Sep. 2018.
- [14] S. Wu, Z. Ni, X. Meng, and L. Kuang, "Block expectation propagation for downlink channel estimation in massive MIMO systems," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2225–2228, Nov 2016.
- [15] S.-i. Amari, *Information Geometry and Its Applications*. Tokyo, Japan: Springer, 2016.
- [16] S. Ikeda, T. Tanaka, and S. Amari, "Stochastic reasoning, free energy, and information geometry," *Neural Computation*, vol. 16, no. 9, pp. 1779–1810, Sep. 2004.
- [17] S. Ikeda, "Information geometrical framework to analyze the belief propagation algorithm," in *Workshop on Mathematics of Statistical Inference, December 2004, Tohoku University, Sendai, Japan*.
- [18] M. Tang, Y. Rong, J. Zhou, and X. R. Li, "Information geometric approach to multisensor estimation fusion," *IEEE Trans. Signal Process.*, vol. 67, no. 2, pp. 279–292, Jan. 2019.
- [19] L. Ye, Q. Yang, Q. Chen, and W. Deng, "Multidimensional joint domain localized matrix constant false alarm rate detector based on information geometry method with applications to high frequency surface wave radar," *IEEE Access*, vol. 7, pp. 28 080–28 088, 2019.
- [20] F. Zhang, Y. Shi, H. K. T. Ng, and R. Wang, "Information geometry of generalized bayesian prediction using α -divergences as loss functions," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1812–1824, Mar. 2018.
- [21] S. Ikeda, T. Tanaka, and S. Amari, "Information geometry of turbo and low-density parity-check codes," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1097–1114, Jun. 2004.
- [22] M. Wang, F. Gao, N. Shlezinger, M. F. Flanagan, and Y. C. Eldar, "A block sparsity based estimator for mmwave massive MIMO channels with beam squint," *IEEE Trans. Signal Process.*, vol. 68, pp. 49–64, 2020.
- [23] M. Jian, F. Gao, Z. Tian, S. Jin, and S. Ma, "Angle-domain aided UL/DL channel estimation for wideband mmwave massive MIMO systems with beam squint," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3515–3527, Jul. 2019.
- [24] L. You, X. Q. Gao, A. L. Swindlehurst, and W. Zhong, "Channel acquisition for massive MIMO-OFDM with adjustable phase shift pilots," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1461–1476, Mar. 2016.
- [25] O. Simeone, Y. Bar-Ness, and U. Spagnolini, "Pilot-based channel estimation for OFDM systems by tracking the delay-subspace," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 315–325, Jan. 2004.

$$\mathbf{K}_n = \begin{bmatrix} \mathbf{I} & \mathbf{0} & 2\text{diag}(\boldsymbol{\mu}_n) \\ \mathbf{0} & \mathbf{I} & 2\text{diag}(\overline{\boldsymbol{\mu}}_n) \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\Sigma}_{0n}^{-1} \boldsymbol{\Sigma}_n & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{0n}^{-1} \overline{\boldsymbol{\Sigma}}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (\boldsymbol{\Sigma}_{0n} \odot \overline{\boldsymbol{\Sigma}}_{0n})^{-1} \boldsymbol{\Sigma}_n \odot \overline{\boldsymbol{\Sigma}}_n \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} & 2\text{diag}(\boldsymbol{\mu}_n) \\ \mathbf{0} & \mathbf{I} & 2\text{diag}(\overline{\boldsymbol{\mu}}_n) \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}. \quad (81)$$

- [26] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.
- [27] A.-A. Lu, X. Q. Gao, X. Meng, and X.-G. Xia, "Omnidirectional precoding for 3D massive MIMO with uniform planar arrays," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2628–2642, Apr. 2020.
- [28] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge, UK: Cambridge university press, 2005.
- [29] A.-A. Lu, X. Q. Gao, and C. Xiao, "Robust precoder design for 3D massive MIMO downlink with a posteriori channel model," *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7274–7286, 2022.
- [30] M. Jian, F. Gao, Z. Tian, S. Jin, and S. Ma, "Angle-domain aided UL/DL channel estimation for wideband mmwave massive MIMO systems with beam squint," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3515–3527, Jul. 2019.
- [31] G. Auer, "3D MIMO-OFDM channel estimation," *IEEE Trans. Commun.*, vol. 60, no. 4, pp. 972–985, Apr. 2012.
- [32] B. H. Fleury, "First- and second-order characterization of direction dispersion and space selectivity in the radio channel," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 2027–2044, Sep. 2000.
- [33] Ke Liu, V. Raghavan, and A. M. Sayeed, "Capacity scaling and spectral efficiency in wide-band correlated MIMO channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2504–2526, Oct. 2003.
- [34] X. Kuai, X. Yuan, W. Yan, H. Liu, and Y. J. Zhang, "Double-sparsity learning-based channel-and-signal estimation in massive MIMO with generalized spatial modulation," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2863–2877, May 2020.
- [35] Y. Zhang, D. Wang, J. Wang, and X. You, "Channel estimation for massive MIMO-OFDM systems by tracking the joint angle-delay subspace," *IEEE Access*, vol. 4, pp. 10 166–10 179, 2016.
- [36] S. M. Kay, *Fundamentals of statistical signal processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
- [37] A. Altland and B. D. Simons, *Condensed Matter Field Theory*. Cambridge, UK: Cambridge university press, 2010.
- [38] G. A. Seber, *A Matrix Handbook for Statisticians*. Hoboken, NJ, USA: Wiley, 2008.
- [39] P. J. Antsaklis and A. N. Michel, *Linear Systems*. Boston, MA: Birkhauser, 2006.
- [40] S. Rangan, P. Schniter, A. K. Fletcher, and S. Sarkar, "On the convergence of approximate message passing with arbitrary matrices," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5339–5351, 2019.
- [41] S. Campese, "Fourth moment theorems for complex gaussian approximation," *arXiv preprint arXiv:1511.00547v1*, 2015.
- [42] D. Bickson, "Gaussian belief propagation: Theory and application," *arXiv preprint arXiv:0811.2518*, 2008.
- [43] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, "Walk-sums and belief propagation in gaussian graphical models," *The Journal of Machine Learning Research*, vol. 7, pp. 2031–2064, 2006.
- [44] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "Quadriga: A 3-d multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, 2014.
- [45] D. Zhang, X. Song, W. Wang, G. Fettweis, and X. Q. Gao, "Unifying message passing algorithms under the framework of constrained bethe free energy minimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4144–4158, Jul. 2021.