Channel Estimation for Densifying MIMO via Two-Dimensional Ice Filling

Zijian Zhang and Mingyao Cui

***Abstract*—In recent years, densifying multiple-input multiple-output (MIMO) has attracted much attention from the communication community. Thanks to the subwavelength antenna spacing, the strong correlations among densifying antennas provide sufficient prior knowledge about channel state information (CSI). This inspires the careful design of observation matrices (e.g., transmit precoders and receive combiners), that exploits the CSI prior knowledge, to boost channel estimation performance. Aligned with this vision, this work proposes to jointly design the combiners and precoders by maximizing the mutual information between the received pilots and densifying MIMO channels. A two-dimensional ice-filling (2DIF) algorithm is proposed to efficiently accomplish this objective. The algorithm is motivated by the observation that the eigenspace of MIMO channel covariance can be decoupled into two sub-eigenspaces, which are associated with the correlations of transmitter antennas and receiver antennas, respectively. By properly setting the precoder and the combiner as the eigenvectors from these two sub-eigenspaces, the 2DIF promises to generate near-optimal observation matrices. Moreover, we further extend the 2DIF method to the popular hybrid combining systems, where a two-stage 2DIF (TS-2DIF) algorithm is developed to handle the analog combining circuits realized by phase shifters. Simulation results demonstrate that, compared to the state-of-the-art schemes, the proposed 2DIF method and TS-2DIF method can achieve superior channel estimation accuracy.**

***Index Terms*—Channel estimation, densifying MIMO, dense array systems (DAS), observation matrix design.**

I. Introduction

In recent years, densifying multiple-input multiple-output (MIMO) have attracted considerable attention from the wireless communication community [1]–[6]. Different from the conventional MIMO whose antennas are usually spaced of half wavelength $\lambda/2$, the antenna spacing of densifying MIMO is much smaller, such as $\lambda/6$ [7], $\lambda/8$ [8], $\lambda/10$ [9], or even $\lambda/23$ [10]. By densely arranging massive subwavelength-spaced antennas in a compact space, densifying MIMO promises to realize the ultimate control of the radiated/received electromagnetic waves on limited apertures. To this end, many dense-antenna transceiver architectures have emerged, such as holographic MIMO (H-MIMO) [1], holographic reconfigurable surfaces (RHSs) [2], continuous-aperture MIMO (CAP-MIMO) [3], superdirective antenna arrays [4], reconfigurable intelligent surfaces (RISs) [5], fluid antenna systems (FASs) [6], and so on. Utilizing the extensive channel observations facilitated by a multitude of antennas, densifying MIMO is anticipated to achieve significant array gains and multiplexing-diversity gains [11]–[14]. Furthermore, densifying MIMO can mitigate the effects of grating lobes and offer enhanced performance for large oblique angles of incidence [15]. Some studies have also highlighted their capabilities to realize super-directivity [4], [16] or super-bandwidth [17] in wireless transmissions.

Enabled by their phase shifters and radio frequency (RF) chains, the transmission performance of MIMO is determined by the constructive precoders/combiners at transceivers [18]. To implement effective precoding/combining, an indispensable technology for MIMO systems is the acquisition of channel state information (CSI) knowledge [19], [20]. To date, numerous technologies have been proposed to estimate the channels of classical MIMO systems. For example, when the available pilot length exceeds the number of antennas, some classical estimators [21], such as the least squares (LS) estimator and the minimum mean square error (MMSE) estimator, can be used to recover MIMO channels in a non-parametric way. By exploiting the channel sparsity in the angular domain, compressed sensing (CS)-based channel estimators can enhance the estimation accuracy and reduce the pilot overhead [22], [23]. Relevant techniques include the orthogonal matching pursuit (OMP)-based estimator [24] and the approximate message passing (AMP)-based estimator [22], [25]. Additionally, some deep learning (DL) approaches, which involve training neural networks based on channel datasets, are utilized to realize data-driven channel estimation in MIMO systems [26]–[28].

Although many channel estimators in the literature can be adopted in densifying MIMO systems, they often exhibit a non-negligible performance gap compared to the optimal estimator [21]. This is because most existing estimators overlook the strong correlations among densifying MIMO antennas. Specifically, since the antenna spacing of densifying MIMO is very small, the channels associated with close-by antennas are spatially similar [14]. Besides, the circuit mutual coupling induces signal interactions between adjacent antennas, further enhancing the channel correlation in densifying MIMO systems [7]–[10]. These facts lead to the highly structured and/or underdetermined covariance matrices of densifying MIMO channels. Existing works have revealed that, such an informative covariance matrix can provide appreciable prior knowledge for the featured design of observation matrices (e.g., combiners and precoders) in estimations, thus improving the accuracy of CSI acquisition [29]–[31].

To exploit the strong channel correlations for improved CSI acquisition, our prior work [32] proposes an ice filling (IF) based observation matrix design in dense array systems (DASs), which is inspired by the idea of Gaussian Process Regression (GPR). By maximizing the mutual information

Zijian Zhang is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, as well as the Beijing National Research Center for Information Science and Technology (BNRist), Beijing 100084, China (e-mail: zhangzj20@mails.tsinghua.edu.cn).

Mingyao Cui is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: mycui@eee.hku.hk). *(Corresponding author: Mingyao Cui.)*

(MI) characterized by the correlation matrix, the IF algorithm can sequentially produce the observation vectors of receivers in a pilot-by-pilot manner. Through optimizing pilot allocation to the channel covariance eigenvectors, this method works like filling ice blocks onto different orthogonal channels. Then, the designed observation matrix is shown to have near-optimal channel estimation performance in DASs, which achieves much higher estimation accuracy compared to the state-of-the-art schemes [32]. Despite its ability to exploit the channel covariance, IF method is only feasible to design the *vector-form receive combiner* in single-input multiple-output (SIMO) system with a single-antenna transmitter and a single-RF-chain receiver [32]. For a general densifying MIMO system with multiple antennas and multiple RF chains at both transceivers, *the receive and transmit channel covariance pair*, as well as the *matrix-form receive combiner and transmit precoder pair* are coupled together. As the IF scheme fails to tackle these couplings, it is far from optimal for desifying MIMO. To the best of our knowledge, the full exploitation of densifying MIMO's channel covariance for designing observation matrices is still an unaddressed challenge.

To fill in this gap, this work generalizes the IF scheme to a two-dimensional ice filling (2DIF) scheme, whose core idea is to design precoders and combiners by decoupling the coupled channel covariances in their eigenspace. Our key contributions and findings are summarized as follows.

- **Generalized framework of observation matrix design**: Inspired by the IF algorithm, we apply the technique of GPR into desifying MIMO channel estimation. Our key idea is to maximize the mutual information between the received pilots and the MIMO channel by jointly optimizing the receive combiners and transmit precoders. The formulated observation matrix design problem is shown to be a generalization of that discussed in IF [32] because of the consideration of practical MIMO systems with multi-RF-chain receivers. To be specific, the receiver-side channel covariance considered by IF is generalized to a MIMO channel covariance, which relies on the correlations at both sides of the transceivers. Moreover, the observation matrix is no longer a vector-form combiner, but the Kronecker product of the matrix-form combiner and precoder. These properties fundamentally distinguish our design from the IF scheme.
- **2DIF based observation matrix design**: To overcome the design challenges imposed by the coupling of matrix-form combiners and precoders in observation matrices, a 2DIF based observation matrix design is developed. The proposed design employs a greedy method to jointly produce the combiners and precoders in a block-by-block way. Concretely, we first prove that the eigenspace of the channel covariance can be decoupled into two sub-eigenspaces, which are associated with the correlations of transmitter antennas and receiver antennas, respectively. Then, utilizing the eigenspace invariance, we show that the near-optimal observation matrix can be obtained by properly setting the precoder and the combiner as the eigenvectors from these two sub-eigenspaces, which can be realized by a linear search algorithm. Besides, we also provide an intuitive and insightful explanation for 2DIF to clarify its physical significance. Similar to the water-filling precoding which maximizes the MIMO capacity, the implementation of 2DIF can be viewed as a two-dimensional ice-filling process.
- **TS-2DIF based observation matrix design**: The proposed 2DIF method requires that the amplitude of each receiver antenna can be controlled independently. However, for many hybrid MIMO structures, only the phase shifts of analogy combiners can be reconfigured, thus the proposed 2DIF cannot be directly adopted in these scenarios. To address this issue, we propose the two-stage 2DIF (TS-2DIF) method. Specifically, at the first stage, the developed 2DIF is used to acquire an ideal observation matrix. Then, at the second stage, the analog combiner, digital combiner, and the precoder at the transceivers are alternatingly optimized to approach the ideal observation matrix, subject to the hardware constraints of the considered hybrid MIMO structure. Numerical results show that, the observation matrices designed by the TS-2DIF method can achieve almost the same performance as the ideal observation matrix designed by the 2DIF method.

The rest of this paper is organized as follows. In Section II, the system model and problem formulation are introduced. In Section III, the proposed 2DIF based observation matrix design for channel estimation is illustrated. In Section IV, the proposed TS-2DIF based observation matrix design is provided. In Section V, the computational complexities of the proposed methods are analyzed, and the kernel selection is discussed. In Section VI, simulations are carried out to verify the effectiveness of the proposed schemes. In Section VII, conclusions are drawn and future works are discussed.

Notation: $[\cdot]^T$, $[\cdot]^H$, $[\cdot]^*$, and $[\cdot]^{-1}$ denote the transpose, conjugate-transpose, conjugate, and inverse operations, respectively; $\|\cdot\|$ denotes the l_2-norm operation; $\|\cdot\|_F$ denotes the Frobenius-norm operation; $z(i)$ denotes the i-th entry of vector $\mathbf{z}$; $\mathbf{Z}(i,j)$, $\mathbf{Z}(j,:)$ and $\mathbf{Z}(:,j)$ denote the (i,j)-th entry, the j-th row, and the j-th column of matrix $\mathbf{Z}$, respectively; $\mathrm{Tr}(\cdot)$ denotes the trace of its argument; $\mathrm{diag}(\cdot)$ and $\mathrm{blkdiag}(\cdot)$ are the diagonal and the block-diagonal operations, respectively; $\mathsf{E}(\cdot)$ is the expectation operator; $\Re\{\cdot\}$ denotes the real part of the argument; $\ln(\cdot)$ denotes the natural logarithm of its argument; $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$; $\mathcal{U}(a,b)$ denotes the uniform distribution between a and b; $\mathbf{I}_L$ is an $L \times L$ identity matrix; $\mathbf{1}_L$ is an all-one vector or matrix with dimension L; and $\mathbf{0}_L$ is a zero vector or matrix with dimension L.

II. System Model and Problem Formulation

A. Transceiver Model

This paper considers the uplink channel estimation of an densifying MIMO system, consisting of an N_{R}-antenna base station (BS) equipped with N_{RF} RF chains and an N_{T}-antenna user [32]. The antennas at transceivers are densely arranged with sub-wavelength antenna spacing d. We define $\mathbf{H} \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{T}}}$ as the wireless channel and Q as the number

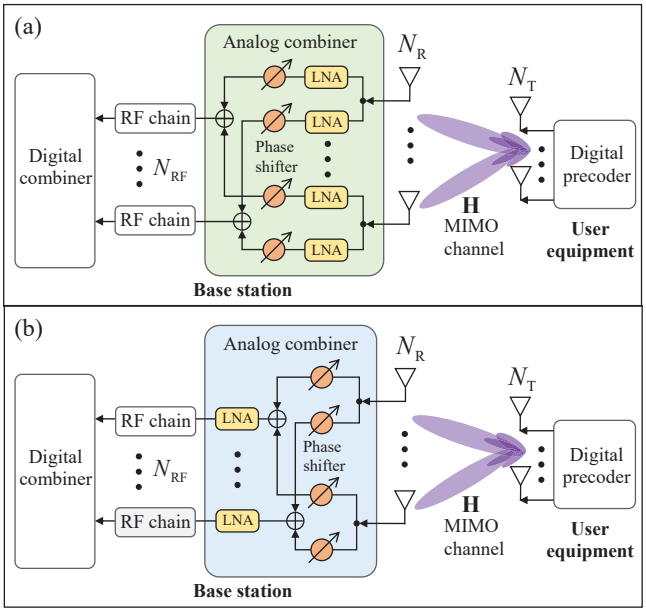

Fig. 1. An illustration of hybrid analog and digital MIMO architectures, where the structure of BS is built on (a) an amplitude-and-phase controllable analog combiner and (b) a phase-only controllable analog combiner, respectively.

of transmit pilots within a coherence-time frame. The received signal $\mathbf{y}_q \in \mathbb{C}^{N_{\mathrm{RF}}}$ at the BS in timeslot q is modeled as

$$\mathbf{y}_q = \mathbf{W}_q^H \mathbf{H} \mathbf{v}_q s_q + \mathbf{W}_q^H \mathbf{z}_q = \left(\mathbf{v}_q^T \otimes \mathbf{W}_q^H\right) \mathbf{h} s_q + \mathbf{W}_q^H \mathbf{z}_q,$$

where $\mathbf{h} \equiv \mathrm{vec}(\mathbf{H})$, s_q is the pilot symbol, and $\mathbf{z}_q \sim \mathcal{CN}\left(\mathbf{0}_M, \sigma^2 \mathbf{I}_M\right)$ is the additive white Gaussian noise (AWGN). Vector $\mathbf{v}_q \in \mathbb{C}^{N_{\mathrm{T}}}$ denotes the precoder at the user. For the transmitter, the user equipment typically employs a fully digital precoder with a moderate number of antennas, N_T. Thereby, the coefficient of $\mathbf{v}_q$ can be freely configured as long as the total power constraint is satisfied: $\|\mathbf{v}_q\|^2 = P$, where P is the maximum transmit power per pilot. For the receiver, $\mathbf{W}_q := \mathbf{A}_q \mathbf{D}_q \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{RF}}}$ is the hybrid combiner at the BS, with $\mathbf{A}_q \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{RF}}}$ and $\mathbf{D}_q \in \mathbb{C}^{N_{\mathrm{RF}} \times N_{\mathrm{RF}}}$ being the analog and digital combiners, respectively. As presented in Fig. 1, we focus on two typical implementations of the analog combiners: the amplitude-and-phase controllable combiner and the phase-only controllable combiner. The former deploys one phase shifter and one low-noise-amplifier (LNA) between each antenna-to-RF chain link. In this architecture, both the amplitude and phase of the elements of $\mathbf{A}_q$ are adjustable thus the coefficients of $\mathbf{W}_q$ can be freely controlled. The latter, however, employs one phase shifter to connect each antenna-to-RF chain link, while the signals aggregated on each RF chain are jointly processed by a global LNA. In this context, only the phase of $\mathbf{A}_q$ is adjustable, which imposes an inherent structural constraint on the feasible set of $\mathbf{W}_q = \mathbf{A}_q \mathbf{D}_q$.

Without loss of generality, we assume that $s_q = 1, \forall q \in \{1, \cdots, Q\}$. Considering the total Q timeslots for pilot transmission, we arrive at

$$\mathbf{y} = \mathbf{X}^H \mathbf{h} + \mathbf{z}, \tag{1}$$

where $\mathbf{y} := \left[\mathbf{y}_1^T, \cdots, \mathbf{y}_Q^T\right]^T$, $\mathbf{z} := \left[\mathbf{z}_1^H \mathbf{W}_1, \cdots, \mathbf{z}_Q^H \mathbf{W}_Q\right]^H$, $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_Q]$, and $\mathbf{X}_q := \mathbf{v}_q^* \otimes \mathbf{W}_q$ is defined as the observation matrix for each pilot. This paper aims at accurately estimating $\mathbf{h}$ from $\mathbf{y}$ by jointly designing combiners $\{\mathbf{W}_q\}_{q=1}^Q$ and precoders $\{\mathbf{v}_q\}_{q=1}^Q$.

B. Channel Model

We consider the Salen-Valenzula (SV) multi-cluster channel model [33]. The uniform linear arrays (ULAs) are deployed at the transceivers for the ease of discussion, while the derivations and results in this paper can be extended to the uniform planar array (UPA) case without difficulty. Let $\mathbf{a}(\theta) \in \mathbb{C}^{N_{\mathrm{R}}}$ and $\mathbf{b}(\varphi) \in \mathbb{C}^{N_{\mathrm{T}}}$ denote the steering vectors at the receiver and transmitter, respectively, given by

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{N_{\mathrm{R}}}} \left[1, e^{\mathrm{j}\frac{2\pi}{\lambda} d \cos(\theta)}, \cdots, e^{\mathrm{j}\frac{2\pi}{\lambda}(N_{\mathrm{R}}-1) d \cos(\theta)}\right]^{\mathrm{T}}, \tag{2}$$

$$\mathbf{b}(\varphi) = \frac{1}{\sqrt{N_{\mathrm{T}}}} \left[1, e^{\mathrm{j}\frac{2\pi}{\lambda} d \cos(\varphi)}, \cdots, e^{\mathrm{j}\frac{2\pi}{\lambda}(N_{\mathrm{T}}-1) d \cos(\varphi)}\right]^{\mathrm{T}}, \tag{3}$$

where θ and ϕ respectively refer to the angle-of-arrival (AoA) and the angle-of-departure (AoD), and λ is the wavelength. Assuming that the number of clusters is C each consisting of R rays, the SV channel $\mathbf{H}$ is represented as

$$\mathbf{H} = \sqrt{\frac{N_{\mathrm{T}} N_{\mathrm{R}}}{CR}} \sum_{c=1}^{C} \sum_{r=1}^{R} g_{c,r} \mathbf{a}(\theta_{c,r}) \mathbf{b}^H(\varphi_{c,r}), \tag{4}$$

where $g_{c,r}$, $\theta_{c,r}$, $\varphi_{c,r}$ are the complex path gain, AoA, and AoD associated with the r-th ray in the c-th cluster, respectively. According to the standard 3GPP setting [34], the path gains $\{g_{c,r}\}_{c=1,r=1}^{C,R}$ are usually modeled as independently and identically distributed (i.i.d.) with zero mean and normalized power, i.e., $\mathsf{E}(g_{c,r}) = 0$ and $\mathsf{E}(|g_{c,r}|^2) = 1$; and the AoAs $\{\theta_{c,r}\}_{c=1,r=1}^{C,R}$ and the AoDs $\{\varphi_{c,r}\}_{c=1,r=1}^{C,R}$ associated with the same cluster are correlated (depending on the angle spread), while those associated with the different clusters are i.i.d [33].

C. Problem Formulation

As the antennas of densifying MIMO are packed with a small spacing, the channels across close-by antennas are strongly correlated. Define the covariance of channel $\mathbf{h}$ as $\mathsf{E}\left(\mathbf{h}\mathbf{h}^H\right) = \mathbf{\Sigma}_{\mathbf{h}} \in \mathbb{C}^{N_{\mathrm{R}} N_{\mathrm{T}} \times N_{\mathrm{R}} N_{\mathrm{T}}}$, which also called the *kernel* of channel. The high-correlation property of wireless channel indicates that the kernel $\mathbf{\Sigma}_{\mathbf{h}}$ is structural and underdetermined, which can provide prior knowledge to achieve high-accuracy channel estimation [35]–[38]. To materialize this vision, we follow the idea of GPR to design the optimal estimator and the optimal observation matrix. Specifically, the channel is assumed to be sampled from the Gaussian process $\mathcal{CN}(\mathbf{0}_{N_{\mathrm{R}} N_{\mathrm{T}}}, \mathbf{\Sigma}_{\mathbf{h}})$. The joint probability distribution of $\mathbf{h}$ and $\mathbf{y}$ then satisfies

$$\begin{bmatrix} \mathbf{h} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{CN}\left(\begin{bmatrix} \mathbf{0}_{N_{\mathrm{R}} N_{\mathrm{T}}} \\ \mathbf{0}_{N_{\mathrm{RF}} Q} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{\mathbf{h}} & \mathbf{\Sigma}_{\mathbf{h}} \mathbf{X} \\ \mathbf{X}^H \mathbf{\Sigma}_{\mathbf{h}} & \mathbf{X}^H \mathbf{\Sigma}_{\mathbf{h}} \mathbf{X} + \mathbf{\Xi} \end{bmatrix}\right), \tag{5}$$

where $\mathbf{\Xi} = \sigma^2 \mathrm{blkdiag}\left(\mathbf{W}_1^H \mathbf{W}_1, \cdots, \mathbf{W}_Q^H \mathbf{W}_Q\right)$ represents the covariance matrix of the noise $\mathbf{z}$. Thereby, the posterior mean and the posterior covariance of $\mathbf{h}$ are expressed as

$$\boldsymbol{\mu}_{\mathbf{h}|\mathbf{y}} = \mathbf{\Sigma}_{\mathbf{h}} \mathbf{X} \left(\mathbf{X}^H \mathbf{\Sigma}_{\mathbf{h}} \mathbf{X} + \mathbf{\Xi}\right)^{-1} \mathbf{y}, \tag{6}$$

$$\boldsymbol{\Sigma}_{\mathbf{h}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{h}} - \boldsymbol{\Sigma}_{\mathbf{h}}\mathbf{X}\left(\mathbf{X}^H\boldsymbol{\Sigma}_{\mathbf{h}}\mathbf{X} + \boldsymbol{\Xi}\right)^{-1}\mathbf{X}^H\boldsymbol{\Sigma}_{\mathbf{h}}, \quad (7)$$

which give rise to the optimal channel estimator $\boldsymbol{\mu}_{\mathbf{h}|\mathbf{y}}$ and the associated estimation error $\boldsymbol{\Sigma}_{\mathbf{h}|\mathbf{y}}$, respectively. Notably, the posterior covariance $\boldsymbol{\Sigma}_{\mathbf{h}|\mathbf{y}}$ is largely dependent on the observation matrix $\mathbf{X}$. Thereby, well-designed combiners and precoders, $\{\mathbf{W}_q\}_{q=1}^Q$ and $\{\mathbf{v}_q\}_{q=1}^Q$, can substantially reduce the channel estimation error. Motivated by this fact, GPR attempts to produce the observation matrices to gain as much information of $\mathbf{h}$ as possible from the received signal $\mathbf{y}$. Following this idea, our objective is to find $\{\mathbf{W}_q\}_{q=1}^Q$ and $\{\mathbf{v}_q\}_{q=1}^Q$ that maximize the MI between $\mathbf{y}$ and $\mathbf{h}$, which is formulated as:

$$\max_{\mathbf{W}\in\mathcal{W},\|\mathbf{v}_q\|_2^2=P} \; I(\mathbf{y};\mathbf{h}) = \log_2\det\left(\mathbf{I}_{N_{\mathrm{RF}}Q} + \boldsymbol{\Xi}^{-1}\mathbf{X}^H\boldsymbol{\Sigma}_{\mathbf{h}}\mathbf{X}\right). \quad (8)$$

where $\mathcal{W}$ stands for the feasible set of hybrid combiners depending on the receiver hardware. In the subsequent sections, we will first elaborate on the observation matrix design while considering the amplitude-and-phase controllable analog combiners in Section III. Then, our design will be extended to the case of phase-only controllable analog combiners in Section IV.

III. Proposed Two-Dimensional Ice Filling (2DIF) Based Observation Matrix Design

In this section, we consider the ideal case when the amplitudes and phases of all precoder and combiner coefficients are adjustable, as shown in Fig. 1 (a). In this context, $\{\mathbf{W}_q\}_{q=1}^Q$ can be freely configured and $\{\mathbf{v}_q\}_{q=1}^Q$ should satisfy the transmit power constraints $\|\mathbf{v}_q\|^2 = P$ for all $q \in \{1,\cdots,Q\}$.

A. Precoder/Combiner Design Using Greedy Method

Observing problem (8), one can find that the MI $I(\mathbf{y};\mathbf{h})$ is non-concave with respect to (w.r.t) the overall observation matrix $\mathbf{X}$. Besides, due to the coupled term $\mathbf{v}_q^T \otimes \mathbf{W}_q^H$ in $\mathbf{X}$ and the colored-noise covariance matrix $\boldsymbol{\Xi}$ introduced by combiners $\{\mathbf{W}_q\}_{q=1}^Q$, the global optimal solution to (8) is hard to obtain. To address this issue, we adopt a greedy method to generate $\{\mathbf{W}_q\}_{q=1}^Q$ and $\{\mathbf{v}_q\}_{q=1}^Q$ in a pilot-by-pilot manner. Specifically, we define $\bar{\mathbf{X}}_t = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_t]$ as the overall observation matrix for timeslots $1 \sim t$, where $t \le Q$. Let $\bar{\mathbf{y}}_t = \bar{\mathbf{X}}_t^H\mathbf{h} + \bar{\mathbf{z}}_t$ denote the corresponding received signal, wherein $\bar{\mathbf{y}}_t = [\mathbf{y}_1^T, \mathbf{y}_2^T, \cdots, \mathbf{y}_t^T]^T$ and $\bar{\mathbf{z}}_t := \left[\mathbf{z}_1^H\mathbf{W}_1, \cdots, \mathbf{z}_t^H\mathbf{W}_t\right]^H$. Given the current observation matrices $\{\mathbf{W}_q\}_{q=1}^t$ and vectors $\{\mathbf{v}_q\}_{q=1}^t$ in the first t timeslots, our greedy strategy aims to find the combiner $\mathbf{W}_{t+1}$ and the precoder $\mathbf{v}_{t+1}$ in the next timeslot, which maximize the MI increment from timeslot t to $t+1$, i.e.,

$$\max_{\mathbf{W}_{t+1}\in\mathcal{W},\mathbf{v}_{t+1}\in\mathcal{V}} \; \Delta I_{t+1} := I(\bar{\mathbf{y}}_{t+1};\mathbf{h}) - I(\bar{\mathbf{y}}_t;\mathbf{h}). \quad (9)$$

For clarity, we summarize the proposed design strategy in **Algorithm 1**, and the sequential designs of $\{\mathbf{W}_q\}_{q=1}^Q$ and $\{\mathbf{v}_q\}_{q=1}^Q$ are illustrated as follows.

Algorithm 1 2DIF Based Combiner and Precoder Design

Input: Number of pilots Q, kernel $\boldsymbol{\Sigma}_{\mathbf{h}}$.
Output: Designed precoders $\{\mathbf{v}_q^{\mathrm{opt}}\}_{q=1}^Q$ and combiners $\{\mathbf{W}_q^{\mathrm{opt}}\}_{q=1}^Q$.

Rewrite kernel as $\boldsymbol{\Sigma}_{\mathbf{h}} = \boldsymbol{\Sigma}_{\mathrm{T}} \otimes \boldsymbol{\Sigma}_{\mathrm{R}}$
Find the eigenvectors $[\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{N_{\mathrm{T}}}]$ and the corresponding eigenvalues $[\alpha_1, \alpha_2, \cdots, \alpha_{N_{\mathrm{T}}}]$ of $\boldsymbol{\Sigma}_{\mathrm{T}}$
Find the eigenvectors $[\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_{N_{\mathrm{R}}}]$ and the corresponding eigenvalues $[\beta_1, \beta_2, \cdots, \beta_{N_{\mathrm{R}}}]$ of $\boldsymbol{\Sigma}_{\mathrm{R}}$
Initialize: $[\lambda_{1,1}^0, \lambda_{1,2}^0, \cdots, \lambda_{N_{\mathrm{T}},N_{\mathrm{R}}}^0] = [\alpha_1\beta_1, \alpha_1\beta_2, \cdots, \alpha_{N_{\mathrm{T}}}\beta_{N_{\mathrm{R}}}]$
for $t = 0, \cdots, Q-1$ **do**
  Find the optimal $n_{\mathrm{T}}^{\mathrm{opt}}$ and $\{n_{\mathrm{R},k}^{\mathrm{opt}}\}_{k=1}^{N_{\mathrm{RF}}}$ via **Algorithm 2**
  Eigenvector-assignment: $\mathbf{v}_{t+1}^{\mathrm{opt}} = \sqrt{P}\mathbf{a}_{n_{\mathrm{T}}^{\mathrm{opt}}}^*$ and $\mathbf{W}_{t+1}^{\mathrm{opt}} = [\mathbf{b}_{n_{\mathrm{R},1}^{\mathrm{opt}}}, \cdots, \mathbf{b}_{n_{\mathrm{R},N_{\mathrm{RF}}}^{\mathrm{opt}}}]$
  Eigenvalue-update for all $n_{\mathrm{T}} \in \{1,\cdots,N_{\mathrm{T}}\}$ and $n_{\mathrm{R}} \in \{1,\cdots,N_{\mathrm{R}}\}$ via
  $\lambda_{n_{\mathrm{T}},n_{\mathrm{R}}}^{t+1} = \begin{cases} \frac{\lambda_{n_{\mathrm{T}},n_{\mathrm{R}}}^t \sigma^2}{P\lambda_{n_{\mathrm{T}},n_{\mathrm{R}}}^t + \sigma^2}, & n_{\mathrm{T}} = n_{\mathrm{T}}^{\mathrm{opt}} \;\&\; n_{\mathrm{R}} \in \{n_{\mathrm{R},k}^{\mathrm{opt}}\}_{k=1}^{N_{\mathrm{RF}}}, \\ \lambda_{n_{\mathrm{T}},n_{\mathrm{R}}}^t, & \text{else.} \end{cases}$
end for
return Designed precoders $\{\mathbf{v}_q^{\mathrm{opt}}\}_{q=1}^Q$ and combiners $\{\mathbf{W}_q^{\mathrm{opt}}\}_{q=1}^Q$ for channel estimation.

1) When $t = 1$: For ease of understanding, we begin with handling the first timeslot, i.e., $t = 1$. In this context, problem (9) can be rewritten as

$$\max_{\|\mathbf{v}_1\|^2=P,\,\mathbf{W}_1} \; I(\mathbf{y}_1;\mathbf{h}) \quad (10)$$

where the mutual information $I(\mathbf{y}_1;\mathbf{h})$ is given by

$$I(\mathbf{y}_1;\mathbf{h}) = \log_2\det\left(\mathbf{I}_{N_{\mathrm{RF}}} + \frac{1}{\sigma^2}\left(\mathbf{W}_1^H\mathbf{W}_1\right)^{-1} \times \left(\mathbf{v}_1^T \otimes \mathbf{W}_1^H\right)\boldsymbol{\Sigma}_{\mathbf{h}}\left(\mathbf{v}_1^* \otimes \mathbf{W}_1\right)\right). \quad (11)$$

Since the reformulated problem (10) is still intricate, we seek to simplify it using the following lemmas.

Lemma 1: In MIMO systems, the covariance of the vectored channel $\boldsymbol{\Sigma}_{\mathbf{h}}$ can be rewritten as the form of a Kronecker product of two kernels, i.e.,

$$\boldsymbol{\Sigma}_{\mathbf{h}} = \boldsymbol{\Sigma}_{\mathrm{T}} \otimes \boldsymbol{\Sigma}_{\mathrm{R}}, \quad (12)$$

where $\boldsymbol{\Sigma}_{\mathrm{T}} \in \mathbb{C}^{N_{\mathrm{T}}\times N_{\mathrm{T}}}$ and $\boldsymbol{\Sigma}_{\mathrm{R}} \in \mathbb{C}^{N_{\mathrm{R}}\times N_{\mathrm{R}}}$ characterize the channel correlation among the transmit antennas and that among the receive antennas, respectively.

Proof: See Appendix A. ∎

Lemma 2: Introducing the orthogonality constraints $\mathbf{W}_q^H\mathbf{W}_q = \mathbf{I}_{N_{\mathrm{RF}}}$ for all $q \in \{1,\cdots,Q\}$ does not influence the optimal value of MI $I(\mathbf{y};\mathbf{h})$ in (8).

Proof: See Appendix B. ∎

Utilizing **Lemma 1** and **Lemma 2** and letting $t = 1$, problem (10) can be equivalently rewritten as

$$\max_{\mathbf{v}_1,\mathbf{W}_1} \; I(\mathbf{y}_1;\mathbf{h}) = \log_2\det\left(\mathbf{I}_{N_{\mathrm{RF}}} + \frac{\mathbf{v}_1^T\boldsymbol{\Sigma}_{\mathrm{T}}\mathbf{v}_1^*}{\sigma^2}\mathbf{W}_1^H\boldsymbol{\Sigma}_{\mathrm{R}}\mathbf{W}_1\right)$$

$$\text{s.t.} \quad \|\mathbf{v}_1\|^2 = P, \\ \mathbf{W}_1^H \mathbf{W}_1 = \mathbf{I}_{N_{\mathrm{RF}}}, \tag{13}$$

where the reorganized objective function $I(\mathbf{y}_1; \mathbf{h})$ is obtained by substituting $\mathbf{\Sigma_h} = \mathbf{\Sigma}_{\mathrm{T}} \otimes \mathbf{\Sigma}_{\mathrm{R}}$ and $\mathbf{W}_1^H \mathbf{W}_1 = \mathbf{I}_{N_{\mathrm{RF}}}$ into (10). Observing (13), one can prove with ease that the optimal $\mathbf{v}_1^*$ is the eigenvector of $\mathbf{\Sigma}_{\mathrm{T}}$ associated with its largest eigenvalue, and the optimal $\mathbf{W}_1$ is composed of the eigenvectors of $\mathbf{\Sigma}_{\mathrm{R}}$ associated with its top N_{RF} eigenvalues. Define the eigenvalue decompositions (EVDs) $\mathbf{\Sigma}_{\mathrm{T}} = \mathbf{U}_{\mathrm{T}} \mathbf{\Lambda}_{\mathrm{T}} \mathbf{U}_{\mathrm{T}}^H$ and $\mathbf{\Sigma}_{\mathrm{R}} = \mathbf{U}_{\mathrm{R}} \mathbf{\Lambda}_{\mathrm{R}} \mathbf{U}_{\mathrm{R}}^H$, wherein the eigenvalues in $\mathbf{\Lambda}_{\mathrm{T}} = \mathrm{diag}\left(\alpha_1, \cdots, \alpha_{N_{\mathrm{T}}}\right)$ and $\mathbf{\Lambda}_{\mathrm{R}} = \mathrm{diag}\left(\beta_1, \cdots, \beta_{N_{\mathrm{R}}}\right)$ are arranged in descending order. In this way, the optimal MI in (13) can be derived as

$$\max_{\mathbf{v}_1, \mathbf{W}_1} \; I(\mathbf{y}_1; \mathbf{h}) = \sum_{n=1}^{N_{\mathrm{RF}}} \log_2 \left(1 + \frac{P \alpha_1 \beta_n}{\sigma^2}\right), \tag{14}$$

and its achievable solution are expressed as

$$\mathbf{v}_1^{\mathrm{opt}} = \sqrt{P} \mathbf{U}_{\mathrm{T}}^* \left(:, 1\right) \text{ and } \mathbf{W}_1^{\mathrm{opt}} = \mathbf{U}_{\mathrm{R}} \left(:, [1, \cdots, N_{\mathrm{RF}}]\right). \tag{15}$$

Equation (15) can be viewed as the optimal initialization settings for our proposed observation matrix design.

2) From t to $t+1$: Given the optimal precoder $\mathbf{v}_1$ and combiner $\mathbf{W}_1$ at timeslot $t = 1$, we then consider designing $\{\mathbf{v}_q\}_{q=2}^{Q}$ and $\{\mathbf{W}_q\}_{q=2}^{Q}$ in a sequential manner. Invoking the principle of recursion, we only need to address the design of precoder $\mathbf{v}_{t+1}$ and combiner $\mathbf{W}_{t+1}$ for given $\{\mathbf{W}_q\}_{q=1}^{t}$ and $\{\mathbf{v}_q\}_{q=1}^{t}$. As proved in Appendix C, the MI increment, ΔI_{t+1}, can be equivalently rewritten as

$$\Delta I_{t+1} = \log_2 \det \left(\mathbf{I}_{N_{\mathrm{RF}}} + \frac{1}{\sigma^2} \mathbf{X}_{t+1}^H \mathbf{\Sigma}_t \mathbf{X}_{t+1}\right), \tag{16}$$

wherein

$$\mathbf{\Sigma}_t = \mathbf{\Sigma_h} - \mathbf{\Sigma_h} \bar{\mathbf{X}}_t \left(\bar{\mathbf{X}}_t^H \mathbf{\Sigma_h} \bar{\mathbf{X}}_t + \sigma^2 \mathbf{I}_{N_{\mathrm{RF}} t}\right)^{-1} \bar{\mathbf{X}}_t^H \mathbf{\Sigma_h} \tag{17}$$

is the posterior kernel of channel $\mathbf{h}$ given the observation $\bar{\mathbf{y}}_t$. In particular, we have $\mathbf{\Sigma}_0 := \mathbf{\Sigma_h}$. Utilizing **Lemma 1** and **Lemma 2**, the optimal precoder $\mathbf{v}_{t+1}$ and combiner $\mathbf{W}_{t+1}$ at the $(t+1)$-th timeslot can be obtained by solving

$$\max_{\mathbf{v}_{t+1}, \mathbf{W}_{t+1}} \quad \log_2 \det \left(\mathbf{I}_{N_{\mathrm{RF}}} + \frac{1}{\sigma^2} \mathbf{X}_{t+1}^H \mathbf{\Sigma}_t \mathbf{X}_{t+1}\right) \\ \text{s.t.} \quad \|\mathbf{v}_{t+1}\|^2 = P, \\ \mathbf{W}_{t+1}^H \mathbf{W}_{t+1} = \mathbf{I}_{N_{\mathrm{RF}}}, \tag{18}$$

where $\mathbf{X}_{t+1} := \mathbf{v}_{t+1}^* \otimes \mathbf{W}_{t+1}$ refers to the Kronecker constraint.

Note that, in problem (13), the kernel $\mathbf{\Sigma_h}$ is decoupled into $\mathbf{\Sigma}_{\mathrm{T}} \otimes \mathbf{\Sigma}_{\mathrm{R}}$ such that $\mathbf{v}_1$ and $\mathbf{W}_1$ can be obtained by selecting the appropriate eigenvectors of $\mathbf{\Sigma}_{\mathrm{T}}$ and $\mathbf{\Sigma}_{\mathrm{R}}$, as shown in (15). Inspired by this fact, we attempt to use the similar idea to solve for $\mathbf{v}_{t+1}$ and $\mathbf{W}_{t+1}$.

Specifically, define the full eigenvalue decomposition $\mathbf{\Sigma}_t = \mathbf{U}_t \mathbf{\Lambda}_t \mathbf{U}_t^H$ where $\mathbf{U}_t \in \mathbb{C}^{N_{\mathrm{T}} N_{\mathrm{R}} \times N_{\mathrm{T}} N_{\mathrm{R}}}$. Notice that the constraints $\|\mathbf{v}_{t+1}\|^2 = P$ and $\mathbf{W}_{t+1}^H \mathbf{W}_{t+1} = \mathbf{I}_{N_{\mathrm{RF}}}$ make the overall matrix observation $\mathbf{X}_{t+1}$ orthogonal, i.e., $\mathbf{X}_{t+1}^H \mathbf{X}_{t+1} = P \mathbf{I}_{\mathrm{RF}}$. If we temporarily omit the Kronecker constraint $\mathbf{X}_{t+1} = \mathbf{v}_{t+1}^* \otimes \mathbf{W}_{t+1}$ and try to solve (18) by considering the orthogonal constraint $\mathbf{X}_{t+1}^H \mathbf{X}_{t+1} = P \mathbf{I}_{\mathrm{RF}}$ only, it becomes evident that the global optimal solution to (18) is $\mathbf{X}_{t+1} = \sqrt{P} \mathbf{U}_t \left(:, [1, \cdots, N_{\mathrm{RF}}]\right)$. Motivated by this discovery, a natural question arises: *when the Kronecker constraint holds, is it possible to set $\mathbf{X}_{t+1}$ as the principal eigenvectors of $\mathbf{\Sigma}_t$ by properly designing $\mathbf{v}_{t+1}$ and $\mathbf{W}_{t+1}$?* Addressing this question is crucial for generating near-optimal observation matrices. We would like to investigate it by analyzing the impacts and feasibility of setting $\mathbf{X}_{t+1}$ as the principal eigenvectors of $\mathbf{\Sigma}_t$.

i) Impacts of setting $\mathbf{X}_{t+1}$ as the principal eigenvectors of $\mathbf{\Sigma}_t$. Before evaluating the feasibility of setting $\mathbf{X}_{t+1} = \sqrt{P} \mathbf{U}_t \left(:, [1, \cdots, N_{\mathrm{RF}}]\right)$, we first need to exploit its influence on the evolution rule of the posterior kernel $\mathbf{\Sigma}_{t+1}$. Specifically, the following lemma characterizes the relationship between $\mathbf{\Sigma}_{t+1}$ and $\mathbf{\Sigma}_t$.

Lemma 3: Let $\lambda_n\left(\cdot\right)$ denote the n-th largest eigenvalue of the matrix in its argument, e.g., $\lambda_n\left(\mathbf{\Sigma}_t\right) = \mathbf{\Lambda}_t\left(n, n\right)$ for $\mathbf{\Sigma}_t = \mathbf{U}_t \mathbf{\Lambda}_t \mathbf{U}_t^H$. If $\mathbf{X}_{t+1} = \sqrt{P} \mathbf{U}_t \left(:, [1, \cdots, N_{\mathrm{RF}}]\right)$, then the EVD of $\mathbf{\Sigma}_{t+1}$ can be derived from $\mathbf{\Sigma}_t = \mathbf{U}_t \mathbf{\Lambda}_t \mathbf{U}_t^H$ by

$$\mathbf{\Sigma}_{t+1} = \mathbf{U}_t \mathrm{diag}\Bigg(\frac{\lambda_1\left(\mathbf{\Sigma}_t\right) \sigma^2}{P \lambda_1\left(\mathbf{\Sigma}_t\right) + \sigma^2}, \frac{\lambda_2\left(\mathbf{\Sigma}_t\right) \sigma^2}{P \lambda_2\left(\mathbf{\Sigma}_t\right) + \sigma^2}, \cdots, \\ \frac{\lambda_{N_{\mathrm{RF}}}\left(\mathbf{\Sigma}_t\right) \sigma^2}{P \lambda_{N_{\mathrm{RF}}}\left(\mathbf{\Sigma}_t\right) + \sigma^2}, \lambda_{N_{\mathrm{RF}}+1}\left(\mathbf{\Sigma}_t\right), \lambda_{N_{\mathrm{RF}}+2}\left(\mathbf{\Sigma}_t\right), \\ \cdots, \lambda_{N_{\mathrm{R}} N_{\mathrm{T}}}\left(\mathbf{\Sigma}_t\right) \Bigg) \mathbf{U}_t^H. \tag{19}$$

Proof: See Appendix D ∎

Lemma 3 reveals that, when $\mathbf{X}_{t+1} = \sqrt{P} \mathbf{U}_t \left(:, [1, \cdots, N_{\mathrm{RF}}]\right)$, the posterior covariance matrix $\mathbf{\Sigma}_{t+1}$ shares the same eigenvector space, $\mathbf{U}_t$, as $\mathbf{\Sigma}_t$. The only difference on their EVDs is that the N_{RF}-largest eigenvalues of $\mathbf{\Sigma}_t$, i.e., $\{\lambda_n(\mathbf{\Sigma}_t)\}_{n=1}^{N_{\mathrm{RF}}}$, are replaced by $\{\frac{\lambda_n(\mathbf{\Sigma}_t)\sigma^2}{P\lambda_n(\mathbf{\Sigma}_t)+\sigma^2}\}_{n=1}^{N_{\mathrm{RF}}}$ in $\mathbf{\Sigma}_{t+1}$. Considering the generality of t, we can conclude that, the eigenvectors of channel covariance $\mathbf{\Sigma}_0 := \mathbf{\Sigma_h}$ are preserved by all subsequent posterior kernels $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$, $\cdots$, and $\mathbf{\Sigma}_{Q-1}$. In other words, the only difference among $\mathbf{U}_0$, $\mathbf{U}_1$, $\cdots$, $\mathbf{U}_{Q-1}$ is their column arrangement orders.

ii) Feasibility of setting $\mathbf{X}_{t+1}$ as the principal eigenvectors of $\mathbf{\Sigma}_t$. Given the eigenvector-preserving property in **Lemma 3**, the feasibility of setting $\mathbf{X}_{t+1}$ as the principal eigenvectors of $\mathbf{\Sigma}_t$ lies in the feasibility of setting $\mathbf{X}_{t+1}$ as the eigenvectors of $\mathbf{\Sigma}_0 = \mathbf{\Sigma_h}$. To assess this feasibility, we examine the structure of the eigenvector space of $\mathbf{\Sigma_h}$ below.

Corollary 1: The EVD of the prior covariance $\mathbf{\Sigma_h}$ can be written as

$$\begin{aligned} \mathbf{\Sigma_h} &= \mathbf{U}_0 \mathbf{\Lambda}_0 \mathbf{U}_0^H \\ &= \underbrace{\left(\mathbf{U}_{\mathrm{T}} \otimes \mathbf{U}_{\mathrm{R}}\right)}_{\text{Orthogonal matrix}} \underbrace{\left(\mathbf{\Lambda}_{\mathrm{T}} \otimes \mathbf{\Lambda}_{\mathrm{R}}\right)}_{\text{Eigenvalue matrix}} \left(\mathbf{U}_{\mathrm{T}}^H \otimes \mathbf{U}_{\mathrm{R}}^H\right) \\ &= \sum_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}} \sum_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}} \alpha_{n_{\mathrm{T}}} \beta_{n_{\mathrm{R}}} \left(\mathbf{a}_{n_{\mathrm{T}}} \otimes \mathbf{b}_{n_{\mathrm{R}}}\right) \left(\mathbf{a}_{n_{\mathrm{T}}}^H \otimes \mathbf{b}_{n_{\mathrm{R}}}^H\right), \end{aligned} \tag{20}$$

where $\mathbf{a}_{n_{\mathrm{T}}} := \mathbf{U}_{\mathrm{T}}\left(:, n_{\mathrm{T}}\right)$ denotes the n_{T}-th eigenvector of $\mathbf{\Sigma}_{\mathrm{T}}$ and $\mathbf{b}_{n_{\mathrm{R}}} := \mathbf{U}_{\mathrm{R}}\left(:, n_{\mathrm{R}}\right)$ denotes the n_{R}-th eigenvector of $\mathbf{\Sigma}_{\mathrm{R}}$. In particular, $\{\alpha_{n_{\mathrm{T}}} \beta_{n_{\mathrm{R}}}\}_{n_{\mathrm{T}}=1, n_{\mathrm{R}}=1}^{N_{\mathrm{T}}, N_{\mathrm{R}}}$ and

$\{\mathbf{a}_{n_\mathrm{T}} \otimes \mathbf{b}_{n_\mathrm{R}}\}_{n_\mathrm{T}=1,n_\mathrm{R}=1}^{N_\mathrm{T},N_\mathrm{R}}$ are the eigenvalues and the corresponding eigenvectors of $\mathbf{\Sigma_h}$, respectively.

Proof: See Appendix E. ∎

Recalling that $\mathbf{X}_{t+1} := \mathbf{v}_{t+1}^* \otimes \mathbf{W}_{t+1}$, we find that the analytical form of $\mathbf{X}_{t+1}$ perfectly matches that of the eigenvectors of $\mathbf{\Sigma_h}$, i.e., $\{\mathbf{a}_{n_\mathrm{T}} \otimes \mathbf{b}_{n_\mathrm{R}}\}_{n_\mathrm{T}=1,n_\mathrm{R}=1}^{N_\mathrm{T},N_\mathrm{R}}$. This encouraging fact inspires us that, the desired optimal $\mathbf{X}_{t+1}$ to solve (18) can be achieved by setting $\mathbf{v}_{t+1}$ and $\mathbf{W}_{t+1}$ to the appropriate eigenvectors from $\left\{\sqrt{P}\mathbf{a}_{n_\mathrm{T}}^*\right\}_{n_\mathrm{T}=1}^{N_\mathrm{T}}$ and $\{\mathbf{b}_{n_\mathrm{R}}\}_{n_\mathrm{R}=1}^{N_\mathrm{R}}$, respectively.

We provide an example to show the implementation process. Firstly, to achieve $\mathbf{X}_1 = \sqrt{P}\mathbf{U}_0(:,[1,\cdots,N_\mathrm{RF}])$ at timeslot $t=1$, we can set $\mathbf{v}_1 = \sqrt{P}\mathbf{a}_1^* = \sqrt{P}\mathbf{U}_\mathrm{T}^*(:,1)$ and $\mathbf{W}_1 = [\mathbf{b}_1,\mathbf{b}_2,\cdots,\mathbf{b}_{N_\mathrm{RF}}] = \mathbf{U}_\mathrm{R}(:,[1,\cdots,N_\mathrm{RF}])$, which coincides with the optimal solution in (15). Note that, according to **Lemma 3**, this process does not changes the eigenvectors of $\mathbf{\Sigma}_1$.[1] In this context, the desired $\mathbf{X}_2 = \sqrt{P}\mathbf{U}_1(:,[1,\cdots,N_\mathrm{RF}])$ at timeslot $t=2$ can also be achieved by carefully selecting $\mathbf{v}_2$ and $\mathbf{W}_2$ from $\left\{\sqrt{P}\mathbf{a}_{n_\mathrm{T}}^*\right\}_{n_\mathrm{T}=1}^{N_\mathrm{T}}$ and $\{\mathbf{b}_{n_\mathrm{R}}\}_{n_\mathrm{R}=1}^{N_\mathrm{R}}$ respectively, which does not influence the eigenvectors of $\mathbf{\Sigma}_2$. Analogously, all our desired $\{\mathbf{X}_q\}_{q=1}^{Q}$ can be obtained by this successive process. As a result, the problem is transformed into: *How to select appropriate eigenvectors such that the objective function in problem (18) is maximized?*

B. Eigenvector Selection

In this subsection, the problem of eigenvector selection is investigated. According to **Lemma 3** and **Corollary 1**, all posterior kernels $\{\mathbf{\Sigma}_t\}_{t=0}^{Q-1}$ can be rewritten as the form of

$$\begin{aligned}\mathbf{\Sigma}_t =& \mathbf{U}_0 \mathrm{diag}\left(\lambda_{1,1}^t, \lambda_{1,2}^t, \cdots, \lambda_{N_\mathrm{T},N_\mathrm{R}}^t\right)\mathbf{U}_0^H \\ =& \sum_{n_\mathrm{T}=1}^{N_\mathrm{T}}\sum_{n_\mathrm{R}=1}^{N_\mathrm{R}} \lambda_{n_\mathrm{T},n_\mathrm{R}}^t \left(\mathbf{a}_{n_\mathrm{T}} \otimes \mathbf{b}_{n_\mathrm{R}}\right)\left(\mathbf{a}_{n_\mathrm{T}}^H \otimes \mathbf{b}_{n_\mathrm{R}}^H\right), \end{aligned} \quad (21)$$

where $\lambda_{n_\mathrm{T},n_\mathrm{R}}^t$ is the $(n_\mathrm{T},n_\mathrm{R})$-th eigenvalue of $\mathbf{\Sigma}_t$ associated with the eigenvector $\mathbf{a}_{n_\mathrm{T}} \otimes \mathbf{b}_{n_\mathrm{R}}$. In particular, $\lambda_{n_\mathrm{T},n_\mathrm{R}}^0 = \alpha_{n_\mathrm{T}}\beta_{n_\mathrm{R}}$, and its update from t to $t+1$ is expressed by

$$\lambda_{n_\mathrm{T},n_\mathrm{R}}^{t+1} = \begin{cases} \dfrac{\lambda_{n_\mathrm{T},n_\mathrm{R}}^t \sigma^2}{P\lambda_{n_\mathrm{T},n_\mathrm{R}}^t + \sigma^2}, & \mathbf{a}_{n_\mathrm{T}}^* \otimes \mathbf{b}_{n_\mathrm{R}} \text{ is selected,} \\ \lambda_{n_\mathrm{T},n_\mathrm{R}}^t, & \text{else.} \end{cases} \quad (22)$$

Based on the above derivation, we prove the following lemma to further simplify the original problem (18):

Lemma 4: When $\mathbf{v}_{t+1} \in \left\{\sqrt{P}\mathbf{a}_{n_\mathrm{T}}^*\right\}_{n_\mathrm{T}=1}^{N_\mathrm{T}}$ and the columns of $\mathbf{W}_{t+1}$ are belonging to $\{\mathbf{b}_{n_\mathrm{R}}\}_{n_\mathrm{R}=1}^{N_\mathrm{R}}$, the original problem (18) can be transformed into an eigenvalue selection problem, written as

$$\begin{aligned} \max_{n_\mathrm{T},\{n_{\mathrm{R},k}\}_{k=1}^{N_\mathrm{RF}}} \quad & \sum_{k=1}^{N_\mathrm{RF}} \log_2\left(1 + \frac{P\lambda_{n_\mathrm{T},n_{\mathrm{R},k}}^t}{\sigma^2}\right) \\ \text{s.t.} \quad & n_\mathrm{T} \in \{1,\cdots,N_\mathrm{T}\}, \\ & n_{\mathrm{R},k} \in \{1,\cdots,N_\mathrm{R}\}, \forall k, \end{aligned}$$

[1]Thus, $\mathbf{U}_1$ and $\mathbf{U}_0 := \mathbf{U}_\mathrm{T} \otimes \mathbf{U}_\mathrm{R}$ share the same columns, but their column arrangement orders may be different due to the eigenvalue updates.

Algorithm 2 Linear Search for Eigenvalue Selection

Input: Eigenvalues $\{\lambda_{n_\mathrm{T},n_\mathrm{R}}^t\}_{n_\mathrm{T}=1,n_\mathrm{R}=1}^{N_\mathrm{T},N_\mathrm{R}}$ in timeslot t.
Output: Optimal eigenvalue indices $n_\mathrm{T}^{\mathrm{opt}}$ and $\{n_{\mathrm{R},k}^{\mathrm{opt}}\}_{k=1}^{N_\mathrm{RF}}$ that maximize $\sum_{k=1}^{N_\mathrm{RF}} \log_2(1+\lambda_{n_\mathrm{T},n_{\mathrm{R},k}}^t/\sigma^2)$.
Initialize indexes: $n_\mathrm{T}^{\mathrm{opt}} = 1$ and $[n_{\mathrm{R},1}^{\mathrm{opt}},\cdots,n_{\mathrm{R},N_\mathrm{RF}}^{\mathrm{opt}}] = [1,\cdots,N_\mathrm{RF}]$.
Initialize the maximum objective: $\zeta_{\max} = \sum_{k=1}^{N_\mathrm{RF}} \log_2(1+P\lambda_{n_\mathrm{T}^{\mathrm{opt}},n_{\mathrm{R},k}^{\mathrm{opt}}}^t/\sigma^2)$
for $n_\mathrm{T} = 1,\cdots,N_\mathrm{T}$ **do**
  Find the N_RF-largest values from $\{\lambda_{n_\mathrm{T},n_\mathrm{R}}^t\}_{n_\mathrm{R}=1}^{N_\mathrm{R}}$, and then denote their second indexes as $\{n_{\mathrm{R},k}\}_{k=1}^{N_\mathrm{RF}}$.
  if $\sum_{k=1}^{N_\mathrm{RF}} \log_2(1+P\lambda_{n_\mathrm{T},n_{\mathrm{R},k}}^t/\sigma^2) > \zeta_{\max}$ **then**
    Update the optimal indexes by $n_\mathrm{T}^{\mathrm{opt}} = n_\mathrm{T}$ and $[n_{\mathrm{R},1}^{\mathrm{opt}},\cdots,n_{\mathrm{R},N_\mathrm{RF}}^{\mathrm{opt}}] = [n_{\mathrm{R},1},\cdots,n_{\mathrm{R},N_\mathrm{RF}}]$
    Update the maximum objective: $\zeta_{\max} = \sum_{k=1}^{N_\mathrm{RF}} \log_2(1+P\lambda_{n_\mathrm{T}^{\mathrm{opt}},n_{\mathrm{R},k}^{\mathrm{opt}}}^t/\sigma^2)$
  end if
end for
return Optimal $n_\mathrm{T}^{\mathrm{opt}}$ and $\{n_{\mathrm{R},k}^{\mathrm{opt}}\}_{k=1}^{N_\mathrm{RF}}$.

$$n_{\mathrm{R},k} \neq n_{\mathrm{R},k'}, \forall k \neq k'. \quad (23)$$

Proof: See Appendix F. ∎

Problem (23) aims to find one eigenvector index n_T on the transmitter side and N_RF different eigenvector indices $\{n_{\mathrm{R},k}\}_{k=1}^{N_\mathrm{RF}}$ on the receiver side, such that the selected eigenvalues $\{\lambda_{n_\mathrm{T},n_{\mathrm{R},k}}^t\}_{k=1}^{N_\mathrm{RF}}$ can maximize the objective (23). A linear search algorithm is proposed to solve (23) optimally, as summarized in **Algorithm 2**. The key idea is to fix an n_T and then find N_RF-largest values from $\{\lambda_{n_\mathrm{T},n_\mathrm{R}}^t\}_{n_\mathrm{R}=1}^{N_\mathrm{R}}$ to calculate the objective $\sum_{k=1}^{N_\mathrm{RF}} \log_2(1+P\lambda_{n_\mathrm{T},n_{\mathrm{R},k}}^t/\sigma^2)$. After traversing all $n_\mathrm{T} \in \{1,\cdots,N_\mathrm{T}\}$, the optimal $n_\mathrm{T}^{\mathrm{opt}}$ and $\{n_{\mathrm{R},k}^{\mathrm{opt}}\}_{k=1}^{N_\mathrm{RF}}$ can be obtained from the indices of the maximum objective. Finally, the optimal precoder and the optimal combiner are thereby expressed as

$$\mathbf{v}_{t+1}^{\mathrm{opt}} = \sqrt{P}\mathbf{a}_{n_\mathrm{T}^{\mathrm{opt}}}^* \text{ and } \mathbf{W}_{t+1}^{\mathrm{opt}} = \left[\mathbf{b}_{n_{\mathrm{R},1}^{\mathrm{opt}}},\cdots,\mathbf{b}_{n_{\mathrm{R},N_\mathrm{RF}}^{\mathrm{opt}}}\right], \quad (24)$$

respectively, which generate a feasible observation matrix $\mathbf{X}_{t+1}^{\mathrm{opt}} = \mathbf{v}_{t+1}^{\mathrm{opt}\,*} \otimes \mathbf{W}_{t+1}^{\mathrm{opt}}$ at timeslot $t+1$. One can verify without difficulty that the initialized precoder and combiner obtained in (18) are a special case of (15) when $t=0$.

To summarize, the eigenvalue updating rule in (22), as well as the eigenvector selection method stated in **Algorithm 2** and (24), allow us to calculate all observation matrices.

C. Insightful Interpretation to 2DIF

In this subsection, we provide insightful explanations to the proposed 2DIF algorithm to clarify its physical significance. At first, the relationship between the well-known water-filling method and the proposed 2DIF method is discussed. Then, the superiority of the proposed 2DIF method over the existing IF method [32] is illustrated.

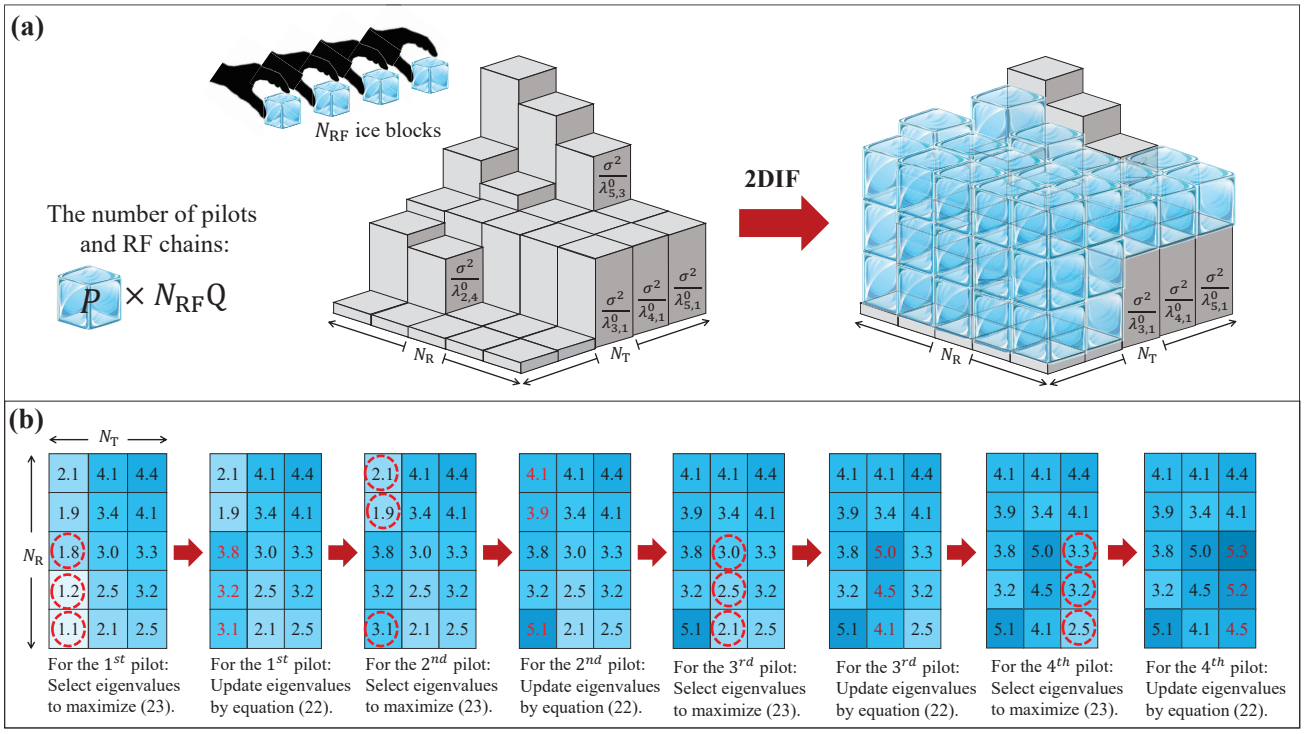

Fig. 2. (a) shows how the pilot allocation of the proposed 2DIF algorithm works to maximize the MI. (b) provides an example to show the implementation process of the proposed 2DIF, where $N_{\mathrm{R}} = 5$, $N_{\mathrm{T}} = 3$, $N_{\mathrm{RF}} = 3$, $P = 2$, and $Q = 4$. The number within the $(n_{\mathrm{T}}, n_{\mathrm{R}})$-th square is the ice-level $\frac{\sigma^2}{\lambda^t_{n_{\mathrm{T}},n_{\mathrm{R}}}}$.

1) Ideal water-filling: To better understand the proposed observation matrix design, we first interpret problem (8) from the view of *water-filling*. Specifically, the orthogonal property of $\mathbf{W}_q$ proved in **Lemma 2** allows us to replace the noise covariance matrix $\mathbf{\Xi}$ in (8) by $\sigma^2\mathbf{I}_{N_{\mathrm{RF}}Q}$ without affecting the optimal $I(\mathbf{y};\mathbf{h})$. Then, by further relaxing the constraints $\mathbf{W}_q^T\mathbf{W}_q = \mathbf{I}_{N_{\mathrm{RF}}}$ and $\|\mathbf{v}_q\|^2 = P$, we focus only on the total power constraint imposed on the overall observation matrix $\mathbf{X}$, i.e., $\mathrm{Tr}\left(\mathbf{X}\mathbf{X}^H\right) = PN_{\mathrm{RF}}Q$. In this case, the optimal value of problem (8) is shown to have an upper bound:

$$\max_{\mathbf{X}} \ \log_2 \det\left(\mathbf{I}_{N_{\mathrm{RF}}Q} + \frac{1}{\sigma^2}\mathbf{X}^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}\right) \quad \text{s.t. } \mathrm{Tr}\left(\mathbf{X}\mathbf{X}^H\right) = PN_{\mathrm{RF}}Q. \tag{25}$$

Notably, this upper bound is equivalent to the channel capacity of a point-to-point MIMO system equipped with $N_{\mathrm{T}}N_{\mathrm{R}}$ transmit antennas and $N_{\mathrm{RF}}Q$ receive antennas. Thereafter, the overall observation matrix can be optimally solved as[2]

$$\mathbf{X}^{\mathrm{ideal}} = \mathbf{U}_0\left(:, [1, \cdots, N_{\mathrm{RF}}Q]\right)\mathbf{P}, \tag{26}$$

where $\mathbf{P} = \mathrm{diag}\left(\sqrt{p_1}, \cdots, \sqrt{p_{N_{\mathrm{RF}}Q}}\right)$ is the power allocation matrix. The power allocated to the n-th eigenvector is determined by the water-filling principle, i.e., $p_n = \left(\beta - \frac{\sigma^2}{\lambda_n(\mathbf{\Sigma}_{\mathbf{h}})}\right)^+$, where the water-level β is adjusted to satisfy the total power constraint $\mathrm{Tr}\left(\mathbf{X}^{\mathrm{ideal}}(\mathbf{X}^{\mathrm{ideal}})^H\right) = \sum_{n=1}^{N_{\mathrm{RF}}Q} p_n = PN_{\mathrm{RF}}Q$.

[2]For ease of discussion, we assume that $N_{\mathrm{RF}}Q$ is smaller than the rank of channel covariance, $\mathbf{\Sigma}_{\mathbf{h}}$.

Although the ideal observation matrix $\mathbf{X}^{\mathrm{ideal}}$, that maximizes the upper bound (25), might not be implementable in practice (as $\mathbf{X}^{\mathrm{ideal}}$ may violate the constraints $\mathbf{W}_q^T\mathbf{W}_q = \mathbf{I}_{N_{\mathrm{RF}}}$ and $\|\mathbf{v}_q\|^2 = P$), it can give us two pivotal intuitions. First, the observation matrix should align with the eigenspace $\{\mathbf{a}_{n_{\mathrm{T}}} \otimes \mathbf{b}_{n_{\mathrm{R}}}\}_{n_{\mathrm{T}}=1,n_{\mathrm{R}}=1}^{N_{\mathrm{T}},N_{\mathrm{R}}}$ of the full MIMO channel covariance, $\mathbf{\Sigma}_{\mathbf{h}} = \mathbf{\Sigma}_{\mathrm{T}} \otimes \mathbf{\Sigma}_{\mathrm{R}}$. Second, we need to fill in more power (water) to the eigenvectors having larger eigenvalues $\{\lambda_n(\mathbf{\Sigma}_{\mathbf{h}})\}_{n=1}^{N_{\mathrm{RF}}Q}$ (or equivalently lower base levels $\{\frac{\sigma^2}{\lambda_n(\mathbf{\Sigma}_{\mathbf{h}})}\}_{n=1}^{N_{\mathrm{RF}}Q}$).

2) 2DIF versus water-filling: The proposed 2DIF algorithm materializes the above two intuitions under the practical constraints $\mathbf{W}_q^T\mathbf{W}_q = \mathbf{I}_{N_{\mathrm{RF}}}$ and $\|\mathbf{v}_q\|^2 = P$ via the eigenvector selection process in (23). The first intuition is automatically achieved by assigning $\mathbf{v}_{t+1}$ with a eigenvector from $\{\sqrt{P}\mathbf{a}^*_{n_{\mathrm{T}}}\}_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}}$ and assigning the columns of $\mathbf{W}_{t+1}$ with different eigenvectors from $\{\mathbf{b}_{n_{\mathrm{R}}}\}_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}}$, as proved in **Lemma 4**. The second intuition is approximately accomplished via selecting eigenvectors that have lower base levels, $\{\frac{\sigma^2}{\lambda_n(\mathbf{\Sigma}_{\mathbf{h}})}\}$, by more times. This is attributed to the fact that the maximization of (23) tends to select a eigenvalue combination $\{\lambda^t_{n_{\mathrm{T}},n_{\mathrm{R},k}}\}_{k=1}^{N_{\mathrm{RF}}}$ that has the lowest $\{\frac{\sigma^2}{\lambda^t_{n_{\mathrm{T}},n_{\mathrm{R},k}}}\}_{k=1}^{N_{\mathrm{RF}}}$ on average. To see this approximation more clearly, we rewrite the updating rule of the selected eigenvalue in (22) as

$$\underbrace{\frac{\sigma^2}{\lambda^{t+1}_{n_{\mathrm{T}},n_{\mathrm{R}}}}}_{\text{Updated ice-level}} = \underbrace{\frac{\sigma^2}{\lambda^{t}_{n_{\mathrm{T}},n_{\mathrm{R}}}}}_{\text{Current ice-level}} + \underbrace{P}_{\text{Height of an ice block}}. \tag{27}$$

Equation (27) reveals that every time the eigenvector $\mathbf{a}^*_{n_{\mathrm{T}}} \otimes \mathbf{b}_{n_{\mathrm{R}}}$ is selected, the value of $\frac{\sigma^2}{\lambda^t_{n_{\mathrm{T}},n_{\mathrm{R}}}}$ increases by P. Similar

to the water-filling process, the process described by (27) can be vividly interpreted as allocating an ice block having P-unit power to the $(n_{\mathrm{T}}, n_{\mathrm{R}})$-th orthogonal channel, where $\frac{\sigma^2}{\lambda^t_{n_{\mathrm{T}},n_{\mathrm{R}}}}$ is viewed as the ice level in the t-th timeslot. To summarize, due to the consideration of N_{RF} RF chains and $N_{\mathrm{T}} \times N_{\mathrm{R}}$ MIMO systems, in each timeslot, the 2DIF first selects N_{RF} orthogonal channels, which have the deepest ice-levels $\{\frac{\sigma^2}{\lambda^t_{n_{\mathrm{T}},n_{\mathrm{R},k}}}\}_{k=1}^{N_{\mathrm{RF}}}$ on average, from the total $N_{\mathrm{T}} \times N_{\mathrm{R}}$ channels. Then, the 2DIF will fill N_{RF} ice blocks (i.e., N_{RF} pilots) of height P onto them. As illustrated in Fig. 2 (a), after Q timeslots, the final ice-levels of all channels can have a similar height with the water-level, β, determined by the water-filling principle. In this case, the second intuition is approximately achieved.

Example 1 (Example for Executing 2DIF). We present an example to show the growth of ice-level in Fig. 2 (b). Here, we set $N_{\mathrm{R}} = 5$, $N_{\mathrm{T}} = 3$, $N_{\mathrm{RF}} = 3$, $Q = 2$, and $P = 2$. The number table in Fig. 2 (b) records all ice-levels in each timeslot, i.e., $\frac{\sigma^2}{\lambda^t_{n_{\mathrm{T}},n_{\mathrm{R}}}}$ for $n_{\mathrm{T}} \in \{1,2,3\}$ and $n_{\mathrm{R}} \in \{1,2,3,4,5\}$. For each pilot, we select 3 deepest ice-levels, that maximize (23), within one column of the number table, i.e., $\{1.1, 1.2, 1.8\}$ for the 1^{st} pilot, $\{3.1, 1.9, 2.1\}$ for the 2^{nd} pilot, $\{3.0, 2.5, 2.1\}$ for the 3^{rd} pilot, and $\{2.5, 3.2, 3.3\}$ for the 4^{th} pilot. Then, every selected ice-level is increased by $P = 2$ for eigenvalue update, i.e., ice-filling, and the corresponding eigenvectors are assigned to the precoders, $\mathbf{v}_t$, and combiners, $\mathbf{W}_t$. As a result, the ice-levels gradually grow and can finally approximate the ideal water-level, which is calculated as $\beta = 4.31$.

3) 2DIF versus IF: We now comprehensively compare the proposed 2DIF algorithm with the IF algorithm devised in our prior work [32]. In a nutshell, the IF algorithm is a special case of 2DIF algorithm, when $N_T = 1$, $N_{\mathrm{RF}} = 1$, and $P = 1$. Specifically, the IF algorithm is tailored to the channel estimation scenario of a *single-antenna* transmitter with *unit* transmit power and a multi-antenna receiver with *single RF chain*. It only exploits the eigenvalue selection and updating rules of the receive kernel $\mathbf{\Sigma}_{\mathrm{R}}$ to design combiners. Attributed to the general design in this paper, our proposed 2DIF has the following two advantages over IF.

- **Utilizing the correlation at transmitter**: The proposed 2DIF algorithm reveals the eigenvalue/eigenspace evolution mode of the full MIMO channel kernel, $\mathbf{\Sigma}_{\mathrm{T}} \otimes \mathbf{\Sigma}_{\mathrm{R}}$. It enables the joint precoder and combiner design to simultaneously exploit both the transmit and receive channel covariance. The IF algorithm, however, lacks the exploitation of the channel covariance at the transmitter. Thereby, the proposed 2DIF algorithm enjoys a better channel estimation performance.
- **Utilizing multi-RF-chain observations**: The IF algorithm is not suitable to multi-RF-chain receivers. Specifically, IF algorithm can only produce one combiner vector $\mathbf{w}_t \in \mathbb{C}^{N_{\mathrm{R}} \times 1}$ in each timeslot t via the eigenvector-selection process. Thus, it is possible that the combiners $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$ designed in two adjacent timeslots are selected as the same eigenvector. At this moment, if $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$ are used in the same timeslot q in a multi-RF-chain receiver, they will lead to identical observations, i.e., $\mathbf{w}_t^H(\mathbf{H}\mathbf{v}_q + \mathbf{z}_q) = \mathbf{w}_{t+1}^H(\mathbf{H}\mathbf{v}_q + \mathbf{z}_q)$. The additional, but identical, channel observation will thus make no contribution to the channel estimation accuracy. In contrast, the proposed 2DIF algorithm efficiently addresses this problem by simultaneously producing all N_{RF} combiner vectors, i.e., $\mathbf{W}_t \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{RF}}}$, in each timeslot. In particular, the orthogonal constraint $\mathbf{W}_t^H \mathbf{W}_t = \mathbf{I}$ ensures that the columns of $\mathbf{W}_t$ are selected as different eigenvectors $\{\mathbf{b}_{n_{\mathrm{R},k}}\}_{k=1}^{N_{\mathrm{R}}}$, where $n_{\mathrm{R},k} \neq n_{\mathrm{R},k'}$ for all $k \neq k'$. This fact guarantees that each RF chain could observe the channel from a unique eigenvector, which can capture the characteristics from different channel patterns to improve the estimation accuracy.

Algorithm 3 TS-2DIF Based Combiner and Precoder Design

Input: Number of pilots Q, kernel $\mathbf{\Sigma}_{\mathbf{h}}$.
Output: Designed precoders $\{\mathbf{v}_q^{\mathrm{opt}}\}_{q=1}^Q$ and hybrid combiners $\{\mathbf{A}_q^{\mathrm{opt}}\}_{q=1}^Q$ and $\{\mathbf{D}_q^{\mathrm{opt}}\}_{q=1}^Q$.
Stage 1 (Optimal observation matrix design)
Obtain the ideal precoders $\{\mathbf{v}_q^{\mathrm{IF}}\}_{q=1}^Q$ and overall combiners $\{\mathbf{W}_q^{\mathrm{IF}}\}_{q=1}^Q$ from **Algorithm 1**
Get the ideal observation matrix $\mathbf{X}_q^{\mathrm{IF}} = (\mathbf{v}_q^{\mathrm{IF}})^* \otimes \mathbf{W}_q^{\mathrm{IF}}$ for all $q \in \{1, \cdots, Q\}$
Stage 2 (Joint hybrid combiner and precoder design)
for $q = 1, \cdots, Q$ **do**
  while no convergence of $\|\mathbf{X}_q^{\mathrm{IF}} - \mathbf{v}_q^* \otimes (\mathbf{A}_q\mathbf{D}_q)\|_F^2$ **do**
    Update the digital combiner $\mathbf{D}_q$ by (30)
    Update the analog combiner $\mathbf{A}_q$ by (35)
    Update the precoder $\mathbf{v}_q$ by (37)
  end while
end for
return Designed precoders $\{\mathbf{v}_q\}_{q=1}^Q$ and hybrid combiners $\{\mathbf{A}_q\}_{q=1}^Q$ and $\{\mathbf{D}_q\}_{q=1}^Q$ for channel estimation.

IV. Proposed Two-Stage 2DIF (TS-2DIF) Based Observation Matrix Design

Turn now to the receiver architecture with phase-only controllable analog combiner presented in Fig. 1 (b). As the coefficients of matrices $\{\mathbf{W}_q = \mathbf{A}_q\mathbf{D}_q\}_{q=1}^Q$ can no longer be freely manipulated, the proposed 2DIF algorithm might not be implementable in these scenarios. To address this problem, a TS-2DIF algorithm is proposed in this section.

A. Overview of TS-2DIF Observation Matrix Design

Recall that the phase-only controllable structure in Fig. 1 (b) requires to express the hybrid combiner as $\mathbf{W}_q = \mathbf{A}_q\mathbf{D}_q$. Each element of the analog combiner $\mathbf{A}_q$ is restricted by the modulus constraint $|\mathbf{A}_q(n_{\mathrm{R}}, n_{\mathrm{rf}})| = \frac{1}{\sqrt{N_{\mathrm{R}}}}$, for $n_{\mathrm{R}} \in \{1, \cdots, N_{\mathrm{R}}\}$ and $n_{\mathrm{rf}} \in \{1, \cdots, N_{\mathrm{RF}}\}$. This poses a structural constraint on the feasible set of $\mathbf{W}_q$, destroying its eigenvector structure, thereby making the 2DIF algorithm inapplicable. Therefore, it becomes necessary to redesign a new set of observation matrices $\{\mathbf{v}_q\}_{q=1}^Q$ and $\{\mathbf{W}_q = \mathbf{A}_q\mathbf{D}_q\}_{q=1}^Q$ tailored for the phase-only controllable architecture.

For this purpose, we propose a TS-2DIF algorithm as summarized in **Algorithm 3**, which consists of two stages. In the first stage, **Algorithm 1** is carried out to obtain the ideal observation matrix $\mathbf{X}_q^{\mathrm{IF}} = (\mathbf{v}_q^{\mathrm{IF}})^* \otimes \mathbf{W}_q^{\mathrm{IF}}$ for $\forall q$, where the superscript "IF" is used to indicate the observation matrices generated by the 2DIF algorithm. Subsequently, the second stage aims to make the newly designed observation matrix $\mathbf{X}_q = \mathbf{v}_q^* \otimes (\mathbf{A}_q\mathbf{D}_q)$ sufficiently close to the ideal observation matrix $\mathbf{X}_q^{\mathrm{IF}}$. To this end, the joint optimization of $\{\mathbf{v}_q\}_{q=1}^Q$ and $\{\mathbf{A}_q\mathbf{D}_q\}_{q=1}^Q$ are formulated as:

$$\min_{\mathbf{v}_q,\mathbf{A}_q,\mathbf{D}_q} \ \|\mathbf{X}_q^{\mathrm{IF}} - \mathbf{v}_q^* \otimes (\mathbf{A}_q\mathbf{D}_q)\|_F^2, \tag{28}$$

$$\text{s.t.} \ \ \|\mathbf{v}_q\|^2 = P, \tag{28a}$$

$$|\mathbf{A}_q| = \frac{1}{\sqrt{N_{\mathrm{R}}}}\mathbf{1}_{N_{\mathrm{R}}\times N_{\mathrm{RF}}}, \tag{28b}$$

where $\mathbf{1}_{N_{\mathrm{R}}\times N_{\mathrm{RF}}}$ is an N_{R}-by-N_{RF} all-one matrix. By solving problem (28), the newly designed observation matrices $\{\mathbf{X}_q\}_{q=1}^Q$ are expected to achieve a comparable channel estimation performance with $\{\mathbf{X}_q^{\mathrm{IF}}\}_{q=1}^Q$.

B. Joint Optimization of Precoders and Hybrid Combiners

It is intricate to directly solve problem (28) owing to the non-convex modulus constraint in (28b) as well as the coupled relationship of $\mathbf{v}_q$, $\mathbf{A}_q$, and $\mathbf{D}_q$ in the objective (28). To overcome these challenges, we exploit the alternating minimization method to iteratively update $\mathbf{A}_q$ and $\mathbf{D}_q$, and $\mathbf{v}_q$ until an convergence condition triggers. The detailed optimization procedures are elaborated one by one as follows.

1) Fix $\mathbf{A}_q$ and $\mathbf{v}_q$, and optimize $\mathbf{D}_q$: For ease of discussion, we denote $\mathbf{v}_q = [v_q(1), v_q(2), \cdots, v_q(N_{\mathrm{T}})]^T$ and define $\mathbf{X}_{q,n}^{\mathrm{IF}} \in \mathbb{C}^{N_{\mathrm{R}}\times N_{\mathrm{RF}}}$ as the n-th block component of $\mathbf{X}_q^{\mathrm{IF}}$ such that $\mathbf{X}_q^{\mathrm{IF}} = [(\mathbf{X}_{q,1}^{\mathrm{IF}})^T, (\mathbf{X}_{q,2}^{\mathrm{IF}})^T, \cdots, (\mathbf{X}_{q,N_{\mathrm{T}}}^{\mathrm{IF}})^T]^T$. Then, when keeping the combiner $\mathbf{A}_q$ and the precoder $\mathbf{v}_q$ fixed, the sub-problem for optimizing $\mathbf{D}_q$ is expressed as

$$\min_{\mathbf{D}_q} \sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{X}_{q,n}^{\mathrm{IF}} - v_q^*(n)\mathbf{A}_q\mathbf{D}_q\|_F^2. \tag{29}$$

Problem (29) is a standard quadradic programming (QP), which can be optimally solved by making the gradient of objective function to zero, i.e.,

$$\begin{aligned}\mathbf{D}_q^{\mathrm{opt}} &= \left(\sum_{n=1}^{N_{\mathrm{T}}} |v_q(n)|^2 \mathbf{A}_q^H\mathbf{A}_q\right)^{-1}\left(\sum_{n=1}^{N_{\mathrm{T}}} v_q(n)\mathbf{A}_q^H\mathbf{X}_{q,n}^{\mathrm{IF}}\right)\\ &\overset{(a)}{=} \sum_{n=1}^{N_{\mathrm{T}}} \frac{v_q(n)}{P}\left(\mathbf{A}_q^H\mathbf{A}_q\right)^{-1}\mathbf{A}_q^H\mathbf{X}_{q,n}^{\mathrm{IF}},\end{aligned} \tag{30}$$

where (a) holds because $\|\mathbf{v}_q\|^2 = \sum_{n=1}^{N_T} |v_q(n)|^2 = P$.

2) Fix $\mathbf{D}_q$ and $\mathbf{v}_q$, and optimize $\mathbf{A}_q$: We then fix the digital combiner $\mathbf{D}_q$ and precoder $\mathbf{v}_q$, and seeks an analog combiner that optimizes the following sub-problem:

$$\min_{\mathbf{A}_q} \ \sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{X}_{q,n}^{\mathrm{IF}} - \mathbf{A}_q\mathbf{D}_q v_q^*(n)\|_F^2, \tag{31}$$

$$\text{s.t.} \ |\mathbf{A}_q| = \frac{1}{\sqrt{N_{\mathrm{R}}}}\mathbf{1}_{N_{\mathrm{R}}\times N_{\mathrm{RF}}}. \tag{31a}$$

Directly optimizing problem (31) is challenging owing to the constant modulus constraint (31a) and the product of $\mathbf{A}_q$ and $\mathbf{D}_q$. To address this issue, we notice that the objective function has the an upper bound due to the Cauchy-Schwarz inequality,

$$\begin{aligned}&\sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{X}_{q,n}^{\mathrm{IF}} - \mathbf{A}_q\mathbf{D}_q v_q^*(n)\|_F^2\\ &\qquad \le \sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{X}_{q,n}^{\mathrm{IF}}\mathbf{D}_q^{-1} - \mathbf{A}_q v_q^*(n)\|_F^2 \|\mathbf{D}_q\|_F^2.\end{aligned} \tag{32}$$

In (32), the analog combiner $\mathbf{A}_q$ has escaped from the product form with $\mathbf{D}_q$, which can significantly simplify the optimization problem. Taking this into account, we replace the original objective function with $\sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{X}_{q,n}^{\mathrm{IF}}\mathbf{D}_q^{-1} - \mathbf{A}_q v_q^*(n)\|_F^2$. Then, by defining $\mathbf{J}_{q,n} = \mathbf{X}_{q,n}^{\mathrm{IF}}\mathbf{D}_q^{-1}$, the new objective function can be further simplified as

$$\begin{aligned}&\sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{J}_{q,n} - \mathbf{A}_q v_q^*(n)\|_F^2\\ &= C_1 + \sum_{n=1}^{N_{\mathrm{T}}} \left(\|\mathbf{A}_q\|_F^2 |v_q(n)|^2 - 2\mathrm{Tr}\left\{\Re\left(v_q^*(n)\mathbf{J}_{q,n}^H\mathbf{A}_q\right)\right\}\right)\\ &= C_1 + N_{\mathrm{RF}}P - 2\mathrm{Tr}\left(\Re\left\{\mathbf{J}^H\mathbf{A}_q\right\}\right),\end{aligned} \tag{33}$$

where $C_1 = \sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{J}_{q,n}\|_F^2$ and $\mathbf{J} := \sum_{n=1}^{N_{\mathrm{T}}} v_q(n)\mathbf{J}_{q,n}$. By combining (32) and (33), the new optimization problem is formulated as

$$\max_{|\mathbf{A}_q| = \frac{1}{\sqrt{N_{\mathrm{R}}}}\mathbf{1}_{N_{\mathrm{R}}\times N_{\mathrm{RF}}}} \ \mathrm{Tr}\left(\Re\left\{\mathbf{J}^H\mathbf{A}_q\right\}\right), \tag{34}$$

Evidently, the optimal solution of (34) is given by

$$\mathbf{A}_q^{\mathrm{opt}} = \frac{1}{\sqrt{N_{\mathrm{R}}}}\exp\left(j\angle\mathbf{J}\right), \tag{35}$$

which completes the update of $\mathbf{A}_q$ in step 6 of **Algorithm 3**.

3) Fix $\mathbf{A}_q$ and $\mathbf{D}_q$, and optimize $\mathbf{v}_q$: With given combiner matrices $\mathbf{A}_q$ and $\mathbf{D}_q$, the objective function in (28) for optimizing $\mathbf{v}_q$ can be rewritten as

$$\begin{aligned}&\sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{X}_{q,n}^{\mathrm{IF}} - \mathbf{A}_q\mathbf{D}_q v_q^*(n)\|_F^2\\ &\qquad = \|\mathbf{X}_q^{\mathrm{IF}}\|_F^2 + \|\mathbf{A}_q\mathbf{D}_q\|_F^2 - 2\Re\{\mathbf{c}_q^H\mathbf{v}_q\},\end{aligned} \tag{36}$$

where $\mathbf{c}_q = [\mathrm{Tr}\{(\mathbf{X}_{q,1}^{\mathrm{IF}})^H\mathbf{A}_q\mathbf{D}_q\}, \cdots, \mathrm{Tr}\{(\mathbf{X}_{q,N_T}^{\mathrm{IF}})^H\mathbf{A}_q\mathbf{D}_q\}]^T$. By further considering the power constraint in (28a), the optimal $\mathbf{v}_q^{\mathrm{opt}}$ is given as

$$\mathbf{v}_q^{\mathrm{opt}} = \sqrt{P}\mathbf{c}_q/\|\mathbf{c}_q\|. \tag{37}$$

This completes the update of $\mathbf{v}_q$ in step 7 of **Algorithm 3**.

To summarize, *Stage 2* of **Algorithm 3** alternatively updates $\mathbf{D}_q$, $\mathbf{A}_q$, and $\mathbf{v}_q$ using (30), (35), and (37) until convergence. After obtaining the hybrid combiners and precoders for pilot transmission, the channel estimator (6) can be utilized to recover wireless channel matrices.

V. Computational Complexity Analysis and Kernel Selection

In this section, the computational complexities of the proposed algorithms are analyzed. Then, the kernel selection for the proposed channel estimator is discussed.

A. Computational Complexity Analysis

Consider **Algorithm 1** at first. The EVDs for $\mathbf{\Sigma}_{\mathrm{T}}$ and $\mathbf{\Sigma}_{\mathrm{R}}$ require $\mathcal{O}\left(N_{\mathrm{T}}^3+N_{\mathrm{R}}^3\right)$ FLOPS. The update of eigenvalues requires $\mathcal{O}\left(QN_{\mathrm{RF}}\right)$ FLOPS in total. In **Algorithm 2**, the linear search requires the sort operations with the complexity of $\mathcal{O}\left(N_{\mathrm{T}}N_{\mathrm{R}}\log_2(N_{\mathrm{R}})\right)$, and calculating the objective requires $\mathcal{O}\left(N_{\mathrm{T}}N_{\mathrm{RF}}\right)$ FLOPS. Thus, the overall complexity of **Algorithm 1** is $\mathcal{O}\left(N_{\mathrm{T}}^3+N_{\mathrm{R}}^3+N_{\mathrm{T}}N_{\mathrm{R}}\log_2(N_{\mathrm{R}})+\left(Q+N_{\mathrm{T}}\right)N_{\mathrm{RF}}\right)$. The computational complexity of **Algorithm 3** is dominated by the alternating optimizations of $\mathbf{D}_q$, $\mathbf{A}_q$, and $\mathbf{v}_q$. In particular, their computations require $\mathcal{O}(Q(N_{\mathrm{RF}}^2N_{\mathrm{R}}+N_{\mathrm{RF}}^3))$, $\mathcal{O}(Q(N_{\mathrm{RF}}^2N_{\mathrm{R}}+N_{\mathrm{RF}}^3+N_{\mathrm{T}}N_{\mathrm{R}}N_{\mathrm{RF}}^2))$, and $\mathcal{O}(QN_{\mathrm{T}}N_{\mathrm{R}}N_{\mathrm{RF}}^2)$ FLOPS, respectively. Assuming that the number of iterations is I_o, the overall computational complexity of **Algorithm 3** is $O\left(I_oQ\left(N_{\mathrm{RF}}^3+N_{\mathrm{T}}N_{\mathrm{R}}N_{\mathrm{RF}}^2\right)\right)$.

It is worth noting that, **Algorithms 1∼3** only rely on the given kernel $\mathbf{\Sigma}_{\mathbf{h}}$ instead of the instantaneous channels or received pilots, thus they do not need to be implemented in real time. Since the channel covariance does not change so frequently, the designed observation matrices can be deployed online for channel estimation for a long time. Thereby, from the long-term perspective, the computational complexity of online deploying the proposed channel estimator is not dominated by the observation matrix design.

B. Kernel Selection

Selecting an appropriate covariance matrix, i.e., kernel $\mathbf{\Sigma}$, is crucial for constructing a robust estimator. The kernel $\mathbf{\Sigma}$ dictates the shape and adaptability of the estimator, thereby influencing its performances to detect functional trends and provide precise predictions. Given the localized-correlation characteristic of MIMO channels, the ideal kernel should enhance the similarity between adjacent antennas while diminishing its impact as the distance increases. In this section, two kinds of kernels are recommended.

1) Statistical Kernel: Given the kernel's role as the prior covariance of channel $\mathbf{h}=\mathrm{vec}(\mathbf{H})$, the optimal strategy is to utilize the actual covariance for channel estimation, i.e., $\mathbf{\Sigma}_{\mathbf{h}}=\mathsf{E}\left(\mathrm{vec}(\mathbf{H})\mathrm{vec}(\mathbf{H})^{\mathrm{H}}\right)$. Prior to deploying the proposed estimator online, it is feasible to train an approximation of $\mathbf{\Sigma}_{\mathbf{h}}$ in advance by leveraging some existing channel models or channel datasets [35]–[38]. Concretely, according to the law of large numbers, $\mathbf{\Sigma}_{\mathbf{h}}$ can be trained by

$$\mathbf{\Sigma}_{\mathbf{h}}\approx\frac{1}{R}\sum_{r=1}^{R}\mathrm{vec}(\mathbf{H}_r)\mathrm{vec}(\mathbf{H}_r)^{\mathrm{H}}, \tag{38}$$

where R is the number of channel realizations and $\mathbf{H}_r$ is the r-th channel realization used for kernel training. As for the covariance matrices $\mathbf{\Sigma}_{\mathrm{T}}$ and $\mathbf{\Sigma}_{\mathrm{R}}$ that characterize the transmitter-side and receive-side channel covariance, respectively, they can be obtained by $\mathbf{\Sigma}_{\mathrm{T}}=\frac{1}{RN_{\mathrm{R}}}\sum_{r=1}^{R}\sum_{n=1}^{N_{\mathrm{R}}}\mathbf{H}_r^T(n,:)\mathbf{H}_r^*(n,:)$ and $\mathbf{\Sigma}_{\mathrm{R}}=\frac{1}{RN_{\mathrm{T}}}\sum_{r=1}^{R}\sum_{n=1}^{N_{\mathrm{T}}}\mathbf{H}_r(:,n)\mathbf{H}_r^H(:,n)$.

2) Artificial Kernels: In practical scenarios where obtaining an explicit channel model or channel dataset is challenging, it is preferred to train an artificial kernel to replace $\mathbf{\Sigma}_{\mathbf{h}}$ [39]. For simplicity, we assume that the ULAs are deployed at both the BS and the user, while it can be easily extended to the UPA case. To characterize the localized-correlation characteristic of channels, two artificial kernels are popular [40]:

- **Laplace Kernel**: The Laplace kernel $\mathbf{\Sigma}_{\mathrm{La}}$ is the most popular choice in Bayesian estimation. Let $\mathbf{n}_{\mathrm{T}}=\left[-\frac{N_{\mathrm{T}}-1}{2},-\frac{N_{\mathrm{T}}-3}{2},\cdots,\frac{N_{\mathrm{T}}-1}{2}\right]^T$ and $\mathbf{n}_{\mathrm{R}}=\left[-\frac{N_{\mathrm{R}}-1}{2},-\frac{N_{\mathrm{R}}-3}{2},\cdots,\frac{N_{\mathrm{R}}-1}{2}\right]^T$. Then, the Laplace kernels, which respectively characterize the channel correlations at the user and the BS, can be modeled as

$$\mathbf{\Sigma}_{\mathrm{La,T}}=\exp\left(-\eta^2\frac{d^2}{\lambda^2}\left|\mathbf{1}_{N_{\mathrm{T}}}^T\otimes\mathbf{n}_{\mathrm{T}}-\mathbf{n}_{\mathrm{T}}^T\otimes\mathbf{1}_{N_{\mathrm{T}}}\right|^{\odot2}\right), \tag{39}$$

$$\mathbf{\Sigma}_{\mathrm{La,R}}=\exp\left(-\eta^2\frac{d^2}{\lambda^2}\left|\mathbf{1}_{N_{\mathrm{R}}}^T\otimes\mathbf{n}_{\mathrm{R}}-\mathbf{n}_{\mathrm{R}}^T\otimes\mathbf{1}_{N_{\mathrm{R}}}\right|^{\odot2}\right), \tag{40}$$

  where λ is the carrier wavelength; d is the antenna spacing; $\eta>0$ is an adjustable hyperparameter; and $\mathbf{Z}^{\odot2}$ denotes the element-wise product of two matrices $\mathbf{Z}$. Thus, the overall kernel can be written as $\mathbf{\Sigma}_{\mathrm{La}}=\mathbf{\Sigma}_{\mathrm{La,T}}\otimes\mathbf{\Sigma}_{\mathrm{La,R}}$.
- **Bessel Kernel**: By exploiting the inherent periodic property of Bessel functions, the Bessel kernel, denoted as $\mathbf{\Sigma}_{\mathrm{Be}}$, is well-suited for recovering data with oscillatory patterns. The Bessel kernels, which respectively characterizes the channel correlations at the user and the BS, can be modeled as

$$\mathbf{\Sigma}_{\mathrm{Be,T}}=J_0\left(\eta\frac{d}{\lambda}\left|\mathbf{1}_{N_{\mathrm{T}}}^T\otimes\mathbf{n}_{\mathrm{T}}-\mathbf{n}_{\mathrm{T}}^T\otimes\mathbf{1}_{N_{\mathrm{T}}}\right|\right), \tag{41}$$

$$\mathbf{\Sigma}_{\mathrm{Be,R}}=J_0\left(\eta\frac{d}{\lambda}\left|\mathbf{1}_{N_{\mathrm{R}}}^T\otimes\mathbf{n}_{\mathrm{R}}-\mathbf{n}_{\mathrm{R}}^T\otimes\mathbf{1}_{N_{\mathrm{R}}}\right|\right), \tag{42}$$

  where J_0 is the zero-order Bessel function of the first kind and $\eta>0$ is a hyperparameter. Thus, the overall kernel to can be written as $\mathbf{\Sigma}_{\mathrm{Be}}=\mathbf{\Sigma}_{\mathrm{Be,T}}\otimes\mathbf{\Sigma}_{\mathrm{Be,R}}$.

The hyperparameter η plays a pivotal role in shaping the performance of artificial kernels. The value of η can be determined by a maximum likelihood (ML) estimator, which ensures that the trained artificial kernel closely mirrors the real channel covariance. We assume that R channel realizations are utilized to train a kernel $\mathbf{\Sigma}$, where $\mathbf{\Sigma}$ can be $\mathbf{\Sigma}_{\mathrm{La}}$ or $\mathbf{\Sigma}_{\mathrm{Be}}$. By viewing hyperparameter η as a variable to be optimized, the ML method to estimate η can be written as

$$\eta^{\mathrm{opt}}=\underset{\eta>0}{\mathrm{argmax}}\ \sum_{r=1}^{R}\ln\left(\mathsf{P}\left(\mathbf{y}_r|\eta\right)\right), \tag{43}$$

wherein the likelihood function is given by

$$\mathsf{P}\left(\mathbf{y}_r|\eta\right)=\frac{\exp\left(-\mathbf{y}_r^H\left(\mathbf{X}_r^H\mathbf{\Sigma}\mathbf{X}_r+\mathbf{\Xi}_r\right)^{-1}\mathbf{y}_r\right)}{\pi^{QN_{\mathrm{RF}}}\det\left(\mathbf{X}_r^H\mathbf{\Sigma}\mathbf{X}_r+\mathbf{\Xi}_r\right)}, \tag{44}$$

in which $\mathbf{y}_r = \mathbf{X}_r^H \mathbf{h}_r + \mathbf{z}_r \in \mathbb{C}^{QN_{\mathrm{RF}}}$ denotes the received pilot associated with the r-th channel realization $\mathbf{h}_r$ for kernel training; $\mathbf{X}_r$ is obtained based on the randomly generated precoders and combiners; and $\mathbf{\Xi}_r$ is the covariance of AWGN $\mathbf{z}_r$. Since η is a positive scalar in (43), the one-dimensional search method can be adopted to obtain the optimal η^{opt}.

VI. Simulation Results

In this section, simulation results are carried out to verify the effectiveness of the proposed 2DIF based and TS-2DIF based channel estimators.

A. Simulation Setup and Baselines

We consider a single-user MIMO system. The ULAs are equipped on both the BS and the user. To account for a practical wireless environment, the 3GPP TR 38.901 channel model is utilized for simulations [34]. In specific, the carrier frequency is set to 3.5 GHz. The number of clusters is 23, and each cluster contributes 20 rays. The angles of arrival (AoAs) and the angles of departure (AoDs) associated with each cluster are randomly selected from $\mathcal{U}(-90°, +90°)$. For each ray in a cluster, the angle spread and the delay spread are randomly selected from $\mathcal{U}(-5°, +5°)$ and $\mathcal{U}(-30\ \mathrm{ns}, +30\ \mathrm{ns})$, respectively. The path gains are randomly selected from $\mathcal{CN}(0, 1)$. Otherwise specifically specified, the numbers of transceiver antennas and RF chains are set as: $N_{\mathrm{T}} = 4$, $N_{\mathrm{R}} = 64$, and $N_{\mathrm{RF}} = 4$, respectively. The antenna spacing is set to be $\frac{\lambda}{8}$. The signal-to-noise ratio (SNR) is defined as $\mathrm{SNR} = \frac{P}{\sigma^2}\mathsf{E}(\|\mathbf{h}\|^2)$, whose default value is set to 10 dB. The evaluation criterion of estimation accuracy is the normalized mean square error (NMSE), which is expressed as $\mathrm{NMSE} = \mathsf{E}(\frac{\|\mathbf{h}-\hat{\mathbf{h}}\|^2}{\|\mathbf{h}\|^2})$. The number of channel realizations for kernel training is set to $R = 100$. The default value of pilot length is set to $Q = 48$.

To verify the effectiveness of the proposed 2DIF based channel estimator and TS-2DIF based channel estimator, the following seven schemes are simulated for comparison:

- **LS**: The least-square (LS) channel estimator is feasible only when observation dimension is no less than the channel dimension, i.e., $QN_{\mathrm{RF}} \geq N_{\mathrm{T}}N_{\mathrm{R}}$. To realize this, we assume the pilot length is $Q = \lceil N_{\mathrm{T}}N_{\mathrm{R}}/N_{\mathrm{RF}} \rceil = 64$, and all combiners/precoders are generated from the columns of discrete Fourier transform (DFT) matrices.
- **MMSE**: Under the same setting of LS estimator, the minimum mean squared error (MMSE) estimator is implemented to recover channel $\mathbf{H}$ via (6).
- **AMP**: Utilizing the channel sparsity in angular domain, the approximate message passing (AMP) method proposed in [25] is implemented to estimate channel $\mathbf{H}$.
- **IF scheme**: By viewing the considered MIMO system as N_{T} independent SIMO systems, the IF-based channel estimator [32] can be utilized to recover $\mathbf{H}$ in a column-by-column way. Note that, since IF scheme is only applicable to the single-RF-chain case, the pilot length used should be modified as $\lceil QN_{\mathrm{RF}} \rceil = 192$ to ensure that it has the same number of observations with the 2DIF.

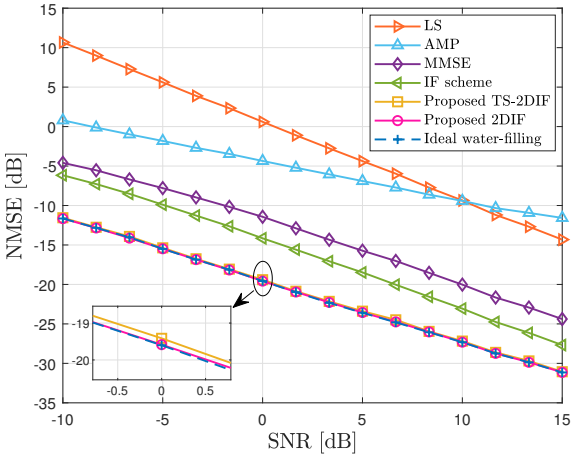

Fig. 3. The NMSE as a function of SNR for different schemes.

- **Proposed 2DIF**: The proposed 2DIF method in **Algorithm 1** is employed to design the precoders and combiners of MIMO system in Fig. 1 (a). Based on these designed observation matrices/vectors, the Bayesian estimator in (6) is employed to estimate channel $\mathbf{H}$.
- **Proposed TS-2DIF**: The proposed TS-2DIF method in **Algorithm 3** is employed to design the precoders and combiners of MIMO system in Fig. 1 (b). The Bayesian estimator in (6) is employed to recover channel $\mathbf{H}$.
- **Ideal water-filling**: To provide a fundamental performance limit, the ideal (but may not be practically achievable) observation matrices $\{\mathbf{X}_q\}_{q=1}^{Q}$ are directly obtained by solving (25) via water-filling method. Then, (6) is employed to recover channel $\mathbf{H}$.

B. Estimation Accuracy Under Statistical Kernel

In this subsection, we consider the ideal case when the statistical kernel $\mathbf{\Sigma}_{\mathbf{h}} := \mathsf{E}(\mathbf{h}\mathbf{h}^H)$ can be trained thanks to the known channel models or datasets. Then, $\mathbf{\Sigma}_{\mathbf{h}}$ is employed for all required estimators for the channel recovery.

Firstly, we plot the NMSE as a function of SNR in Fig. 3. One can observe that, thanks to the carefully designed observation matrices/vectors, the proposed 2DIF and TS-2DIF schemes remarkably outperform the benchmark schemes in estimation accuracy. The reason is that the proposed methods fully exploit the spatial correlations among the transceiver antennas for channel estimation. In particular, the NMSEs for the proposed 2DIF and TS-2DIF schemes are about 5 dB lower than that for the IF scheme. It is because the IF schemes realizes the MIMO channel estimation by viewing it as N_{T} independent SIMO channel estimations, which ignores the spatial correlation of transmitter antennas. Besides, we note that the proposed 2DIF scheme achieves very similar performance to the ideal water-filling scheme. This phenomenon implies that the practical pilot allocation can behave almost the same as the theoretically-optimal "continuous" pilot allocation.

Then, the NMSE versus the number of pilots Q is provided in Fig. 4. One can find that the superiority of the proposed schemes still holds. Although the dimension of the estimated parameters is high as $N_{\mathrm{T}}N_{\mathrm{R}} = 256$, using a small number

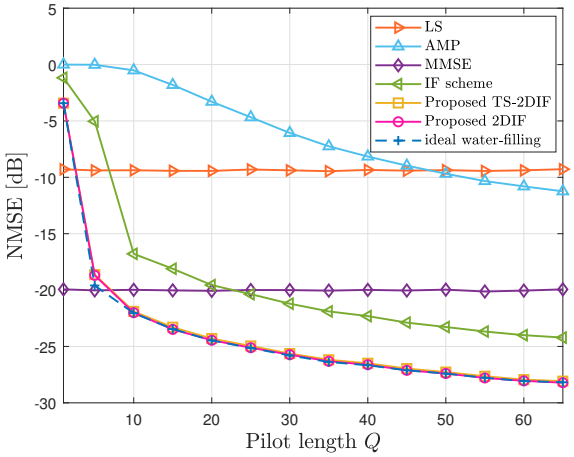

Fig. 4. The NMSE as a function of pilot length Q for different schemes.

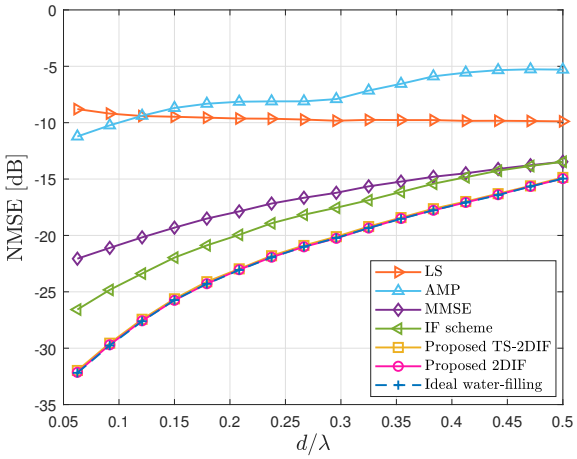

Fig. 5. The NMSE as a function of the normalized antenna spacing d/λ for different schemes.

of pilots $Q = 20$, the NMSEs for the proposed schemes can be lower than -20 dB. In contrast, even if the pilot length is longer than $Q = 60$, the conventional AMP estimator is still unable to achieve such high accuracy. It indicates that utilizing the correlation of compact antennas is of great significance for high-accuracy channel estimation. In addition, observing Fig. 3 and Fig. 4, one can conclude that the TS-2DIF scheme can achieve almost the same estimation accuracy as the 2DIF scheme. This indicates that, from the perspective of CSI acquisition, both hybrid MIMO structures in Fig. 1 have no obvious performance gap.

To observe the impact of spatial correlations on the estimation accuracy, we plot the NMSE as a function of the normalized antenna spacing d/λ in Fig. 5. One can observe that, as the antenna spacing decreases, the estimation accuracy of the proposed schemes becomes higher. It is because a smaller antenna spacing leads to stronger spatial correlations, which can provide more prior knowledge for Bayesian estimators. In this case, the more informative kernel allows the proposed schemes to realize more accurate channel estimation. Besides, we also note that, even if the antenna spacing is $\lambda/2$, the proposed 2DIF and TS-2DIF methods can still hold the superiority. This fact indicates that, for a conventional massive MIMO system, as long as its channels are not i.i.d. Rayleigh-

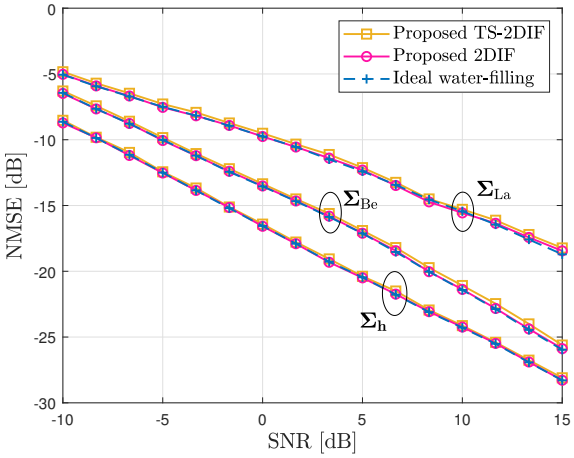

Fig. 6. The NMSE as a function of SNR for different kernels.

fading (otherwise $\mathbf{\Sigma}_{\mathbf{h}} = \mathbf{I}_{N_{\mathrm{T}}N_{\mathrm{R}}}$), the non-diagonal kernel $\mathbf{\Sigma}_{\mathbf{h}}$ with some structural properties can still contribute to the improvement of the estimation accuracy. This result verifies the generality of our proposed schemes in MIMO channel estimations.

C. Estimation Accuracy Under Artificial Kernels

In practical scenarios where obtaining an explicit channel model or channel dataset is challenging, it is preferred to use an artificial kernel for channel estimation, as discussed in Subsection V-B. In this subsection, two popular artificial kernels, i.e., Laplace kernel $\mathbf{\Sigma}_{\mathrm{La}}$ and Bessel kernel $\mathbf{\Sigma}_{\mathrm{Be}}$, are compared with the ideal statistical kernel $\mathbf{\Sigma}_{\mathbf{h}}$. We plot the NMSE as a function of the SNR in Fig. 6 and the NMSE as a function of the pilot length Q in Fig. 7, respectively.

One can observe that, for each type of kernel, the three proposed schemes exhibit very similar trends in estimation accuracy. This implies that our proposed estimators have the similar robustness for different kernels in channel estimation. Compared to the ideal kernel, the performance losses for both artificial kernels are acceptable. For examples, when $\mathrm{SNR} = 10$ dB, the NMSEs for Laplace kernel, Bessel kernel, and statistical kernel are about -24 dB, -21 dB, and -16 dB, respectively. When the pilot length is $Q = 21$, the NMSEs for these three kernels are about -14 dB, -18 dB, and -21 dB, respectively. We can conclude that, even if the real covariance (statistical kernel) is unknown, the proposed channel estimator can still hold their performance advantages by training artificial kernels. This fact encourages the potential applications of the proposed schemes in practice.

Besides, it is interesting to find that the NMSE for the Bessel kernel is about 5 dB $\sim$ 7 dB lower than that for the Laplace kernel. The reason is that, unlike the Laplace kernel whose correlation decreases monotonically with the distance, the Bessel kernel can characterize the data with oscillatory or periodic patterns thanks to its embedded Bessel function. Fortunately, the clustered channel covariance often exhibits spatial periodicity within a limited aperture space, such as the well-known Clarke model and Fourier model [12]. As a result, compared to the Laplace kernel, the Bessel kernel

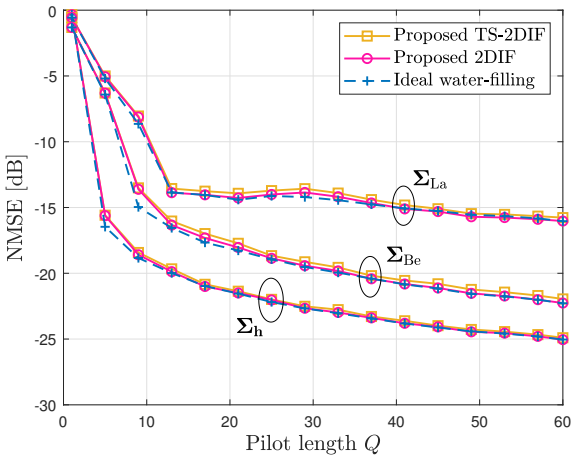

Fig. 7. The NMSE as a function of pilot length Q for different kernels.

with periodic property is more suitable to estimate densifying MIMO channels.

VII. Conclusions

By fully exploiting the channel correlations among antennas, this work has proposed a generalized channel estimation framework for densifying MIMO systems, focusing on the design of observation matrices. By maximizing the MI between channels and received pilots, the 2DIF method has been proposed to design observation matrices through jointly optimizing the precoders and combiners. Subsequently, the TS-2DIF method has been proposed to extend the applicability of our framework to the typical hybrid MIMO whose analog combiner is phase-only controllable. Simulation results have validated the superiority of our proposed methods compared to the conventional channel estimators.

In future works, due to the frequency-domain correlation of multi-carrier channels, investigating the wideband channel estimation can be interesting. Considering the channel-correlated scenario where different users share same scatterers, the proposed 2DIF can be extended to the multi-user MIMO case. For RIS-aided communication systems, analyzing and exploiting the spatial correlations of RIS-aided cascaded channels can be a potential direction.

Appendix A
Proof of Lemma 1

Given the channel model in (4) and $\mathbf{h} \equiv \mathrm{vec}\left(\mathbf{H}\right)$, the vectorized channel can be rewritten as

$$\mathbf{h} = \sqrt{\frac{N_{\mathrm{T}}N_{\mathrm{R}}}{CR}} \sum_{c=1}^{C}\sum_{r=1}^{R} g_{c,r}\mathbf{b}^{*}\left(\varphi_{c,r}\right) \otimes \mathbf{a}\left(\theta_{c,r}\right). \tag{45}$$

Utilizing the commutative law of Kronecker product $\left(\mathbf{A}\otimes\mathbf{B}\right)\left(\mathbf{C}\otimes\mathbf{D}\right) = \left(\mathbf{AC}\right)\otimes\left(\mathbf{BD}\right)$, we have

$$\begin{aligned}&\left(\mathbf{b}^{*}\left(\varphi_{c,r}\right)\otimes\mathbf{a}\left(\theta_{c,r}\right)\right)\left(\mathbf{b}^{T}\left(\varphi_{c,r}\right)\otimes\mathbf{a}^{H}\left(\theta_{c,r}\right)\right) = \\ &\left(\mathbf{b}^{*}\left(\varphi_{c,r}\right)\mathbf{b}^{T}\left(\varphi_{c,r}\right)\right)\otimes\left(\mathbf{a}\left(\theta_{c,r}\right)\mathbf{a}^{H}\left(\theta_{c,r}\right)\right).\end{aligned} \tag{46}$$

Then, the covariance of channel $\mathbf{h}$ can be derived as (47), where (a) holds since the gains of different rays $\{g_{c,r}\}_{c=1,r=1}^{C,R}$ are i.i.d. with zero mean and normalized power. (b) holds according to (46). (c) holds since $\mathsf{E}\left(\mathbf{a}\left(\theta_{c,r}\right)\mathbf{a}^{H}\left(\theta_{c,r}\right)\right) = \mathsf{E}\left(\mathbf{a}\left(\theta_{c',r'}\right)\mathbf{a}^{H}\left(\theta_{c',r'}\right)\right)$ and $\mathsf{E}\left(\mathbf{b}^{*}\left(\varphi_{c,r}\right)\mathbf{b}^{T}\left(\varphi_{c,r}\right)\right) = \mathsf{E}\left(\mathbf{b}^{*}\left(\varphi_{c',r'}\right)\mathbf{b}^{T}\left(\varphi_{c',r'}\right)\right)$ hold for any $c, c' \in \{1,\cdots,C\}$ and $r, r' \in \{1,\cdots,R\}$. (d) holds by defining

$$\mathbf{\Sigma}_{\mathrm{T}} = N_{\mathrm{T}}\mathsf{E}\left(\mathbf{b}^{*}\left(\varphi_{c,r}\right)\mathbf{b}^{T}\left(\varphi_{c,r}\right)\right), \tag{48}$$

$$\mathbf{\Sigma}_{\mathrm{R}} = N_{\mathrm{R}}\mathsf{E}\left(\mathbf{a}\left(\theta_{c,r}\right)\mathbf{a}^{H}\left(\theta_{c,r}\right)\right), \tag{49}$$

wherein c and r can be arbitrarily selected from $\{1,\cdots,C\}$ and $\{1,\cdots,R\}$, respectively. One can find that, the matrix $\mathbf{\Sigma}_{\mathrm{T}}$ only depends on the steering vector $\mathbf{b}\left(\varphi\right)$ at the transmitter, while the matrix $\mathbf{\Sigma}_{\mathrm{R}}$ is only associated with the steering vector $\mathbf{a}\left(\theta\right)$ at the receiver. Thus, $\mathbf{\Sigma}_{\mathrm{T}}$ and $\mathbf{\Sigma}_{\mathrm{R}}$ can be viewed as the kernels that characterize the correlation among the transmitter antennas and that among the receiver antennas, respectively. This completes the proof.

Appendix B
Proof of Lemma 2

Using some matrix techniques, the MI $I(\mathbf{y};\mathbf{h})$ can be rewritten as equation (50), where (a) holds since $\det\left(\mathbf{I}+\mathbf{AB}\right) = \det\left(\mathbf{I}+\mathbf{BA}\right)$ and $\mathbf{\Xi} = \sigma^2\mathrm{blkdiag}\left(\mathbf{W}_1^H\mathbf{W}_1,\cdots,\mathbf{W}_Q^H\mathbf{W}_Q\right)$; (b) holds according to the property that $\left(\mathbf{a}\otimes\mathbf{B}\right)\mathbf{C}\left(\mathbf{a}^H\otimes\mathbf{D}\right) = \left(\mathbf{a}\mathbf{a}^H\right)\otimes\left(\mathbf{BCD}\right)$ if all dimensions meet the requirements of matrix multiplications. To find more insights, we perform singular value decomposition (SVD) on all $\{\mathbf{W}_q\}_{q=1}^{Q}$ and then substitute all decomposition formulas $\mathbf{W}_q = \mathbf{\Pi}_q\mathbf{\Omega}_q\mathbf{\Upsilon}_q^H$ into (50). It is evident that $\mathbf{W}_q\left(\mathbf{W}_q^H\mathbf{W}_q\right)^{-1}\mathbf{W}_q^H = \mathbf{\Pi}_q\mathbf{\Pi}_q^H$, thus the MI $I(\mathbf{y};\mathbf{h})$ can be rewritten as

$$I(\mathbf{y};\mathbf{h}) = \log_2\det\left(\mathbf{I}_{N_{\mathrm{R}}N_{\mathrm{T}}} + \frac{1}{\sigma^2}\sum_{q=1}^{Q}\left(\left(\mathbf{v}_q^{*}\mathbf{v}_q^{T}\right)\otimes\left(\mathbf{\Pi}_q\mathbf{\Pi}_q^H\right)\right)\mathbf{\Sigma}_{\mathbf{h}}\right). \tag{51}$$

Observing (51), one can find that the MI $I(\mathbf{y};\mathbf{h})$ in (8) only relies on the orthogonal matrix $\mathbf{\Pi}_q \in \mathbb{C}^{N\times N_{\mathrm{RF}}}$ decomposed from $\mathbf{W}_q$ for all $q \in \{1,\cdots,Q\}$, while it does not depend on any $\mathbf{\Omega}_q$ or $\mathbf{\Upsilon}_q$. It indicates that imposing $\mathbf{W}_q = \mathbf{\Pi}_q$ does not change the value of $I(\mathbf{y};\mathbf{h})$. As a result, the orthogonality constraint $\mathbf{W}_q^H\mathbf{W}_q = \mathbf{\Pi}_q^H\mathbf{\Pi}_q = \mathbf{I}_{\mathrm{RF}}$ can be safely introduced into the problem formulation regarding $I(\mathbf{y};\mathbf{h})$, which completes the proof.

Appendix C
Proof of MI increment $I(\bar{\mathbf{y}}_{t+1};\mathbf{h}) - I(\bar{\mathbf{y}}_t;\mathbf{h})$

Using some matrix partition operations, the MI $I(\bar{\mathbf{y}}_{t+1};\mathbf{h})$ can be rewritten as

$$\begin{aligned}I(\bar{\mathbf{y}}_{t+1};\mathbf{h}) &\overset{(a)}{=} \log_2\det\left(\mathbf{I}_{N_{\mathrm{RF}}Q} + \frac{1}{\sigma^2}\bar{\mathbf{X}}_{t+1}^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_{t+1}\right) \\ &= \log_2\det\begin{bmatrix}\mathbf{I}_{N_{\mathrm{RF}}t}+\frac{1}{\sigma^2}\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t & \frac{1}{\sigma^2}\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}_{t+1} \\ \frac{1}{\sigma^2}\mathbf{X}_{t+1}^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t & \mathbf{I}_{N_{\mathrm{RF}}}+\frac{1}{\sigma^2}\mathbf{X}_{t+1}^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}_{t+1}\end{bmatrix} \\ &\overset{(b)}{=} \log_2\det\begin{bmatrix}\mathbf{I}_{N_{\mathrm{RF}}t}+\frac{1}{\sigma^2}\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t & \frac{1}{\sigma^2}\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}_{t+1} \\ \mathbf{0}_{N_{\mathrm{RF}}\times N_{\mathrm{RF}}t} & \mathbf{I}_{N_{\mathrm{RF}}}+\frac{1}{\sigma^2}\mathbf{X}_{t+1}^H\mathbf{\Sigma}_t\mathbf{X}_{t+1}\end{bmatrix}\end{aligned}$$

$$= I(\bar{\mathbf{y}}_t;\mathbf{h}) + \log_2\det\left(\mathbf{I}_{N_{\mathrm{RF}}} + \frac{1}{\sigma^2}\mathbf{X}_{t+1}^H\boldsymbol{\Sigma}_{\mathbf{h}}\mathbf{X}_{t+1}\right), \quad (52)$$

where (a) holds since according to **Lemma 2** and (b) holds by performing matrix triangularization. In particular, $\boldsymbol{\Sigma}_t$ is given by $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{\mathbf{h}} - \boldsymbol{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t\left(\bar{\mathbf{X}}_t^H\boldsymbol{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t + \sigma^2\mathbf{I}_{N_{\mathrm{RF}}t}\right)^{-1}\bar{\mathbf{X}}_t^H\boldsymbol{\Sigma}_{\mathbf{h}}$, which completes the proof.

Appendix D
Proof of Lemma 3

The key idea of the proof is to rewrite the $\bar{\mathbf{X}}_t$-related terms in (17) as $\boldsymbol{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t = \boldsymbol{\Sigma}_{\mathbf{h}}\left[\bar{\mathbf{X}}_{t-1}, \mathbf{X}_t\right]$ and

$$\bar{\mathbf{X}}_t^H\boldsymbol{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t = \begin{bmatrix}\bar{\mathbf{X}}_{t-1}^H\boldsymbol{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_{t-1} & \bar{\mathbf{X}}_{t-1}^H\boldsymbol{\Sigma}_{\mathbf{h}}\mathbf{X}_t \\ \mathbf{X}_t^H\boldsymbol{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_{t-1} & \mathbf{X}_t^H\boldsymbol{\Sigma}_{\mathbf{h}}\mathbf{X}_t\end{bmatrix}. \quad (53)$$

Then, using the Schur's matrix inversion formula to expand the term $\left(\bar{\mathbf{X}}_t^H\boldsymbol{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t + \sigma^2\mathbf{I}_{N_{\mathrm{RF}}t}\right)^{-1}$ in (17), the following recursion formula of can be obtained:

$$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t\mathbf{X}_{t+1}\left(\mathbf{X}_{t+1}^H\boldsymbol{\Sigma}_t\mathbf{X}_{t+1} + \sigma^2\mathbf{I}_{N_{\mathrm{RF}}}\right)^{-1}\mathbf{X}_{t+1}^H\boldsymbol{\Sigma}_t, \quad (54)$$

When $\mathbf{X}_{t+1} = \sqrt{P}\mathbf{U}_t(:,[1,\cdots,N_{\mathrm{RF}}])$, we have $\boldsymbol{\Sigma}_t\mathbf{X}_{t+1} = \mathbf{X}_{t+1}\mathrm{diag}\left(\lambda_1(\boldsymbol{\Sigma}_t),\cdots,\lambda_{N_{\mathrm{RF}}}(\boldsymbol{\Sigma}_t)\right)$ and $\mathbf{X}_{t+1}^H\boldsymbol{\Sigma}_t\mathbf{X}_{t+1} = P\mathrm{diag}\left(\lambda_1(\boldsymbol{\Sigma}_t),\cdots,\lambda_{N_{\mathrm{RF}}}(\boldsymbol{\Sigma}_t)\right)$. Thus, the following equality holds:

$$\boldsymbol{\Sigma}_{t+1} = \mathbf{U}_t\boldsymbol{\Lambda}_t\mathbf{U}_t^H - \mathbf{X}_{t+1}\mathrm{diag}\Bigg(\frac{\lambda_1^2(\boldsymbol{\Sigma}_t)}{P\lambda_1(\boldsymbol{\Sigma}_t)+\sigma^2},\cdots, \frac{\lambda_{N_{\mathrm{RF}}}^2(\boldsymbol{\Sigma}_t)}{P\lambda_{N_{\mathrm{RF}}}(\boldsymbol{\Sigma}_t)+\sigma^2}\Bigg)\mathbf{X}_{t+1}^H. \quad (55)$$

Given that $\mathbf{X}_{t+1}\mathrm{diag}(\frac{\lambda_1^2(\boldsymbol{\Sigma}_t)}{P\lambda_1(\boldsymbol{\Sigma}_t)+\sigma^2},\cdots,\frac{\lambda_{N_{\mathrm{RF}}}^2(\boldsymbol{\Sigma}_t)}{P\lambda_{N_{\mathrm{RF}}}(\boldsymbol{\Sigma}_t)+\sigma^2})\mathbf{X}_{t+1}^H = \mathbf{U}_t\mathrm{diag}(\frac{P\lambda_1^2(\boldsymbol{\Sigma}_t)}{P\lambda_1(\boldsymbol{\Sigma}_t)+\sigma^2},\cdots,\frac{P\lambda_{N_{\mathrm{RF}}}^2(\boldsymbol{\Sigma}_t)}{P\lambda_{N_{\mathrm{RF}}}(\boldsymbol{\Sigma}_t)+\sigma^2},\underbrace{0,\cdots,0}_{N_{\mathrm{R}}N_{\mathrm{T}}-N_{\mathrm{RF}}})\mathbf{U}_t^H$ and $\boldsymbol{\Sigma}_t = \mathbf{U}_t\boldsymbol{\Lambda}_t\mathbf{U}_t^H$, the equality in (19) can be derived from (55), which completes the proof.

Appendix E
Proof of Corollary 1

According to **Lemma 1** and equality $\left(\mathbf{ABA}^H\right)\otimes\left(\mathbf{CDC}^H\right) = \left(\mathbf{A}\otimes\mathbf{C}\right)\left(\mathbf{B}\otimes\mathbf{D}\right)\left(\mathbf{A}^H\otimes\mathbf{C}^H\right)$, the kernel $\boldsymbol{\Sigma}_{\mathbf{h}}$ can be decomposed as

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{h}} &= \left(\mathbf{U}_{\mathrm{T}}\boldsymbol{\Lambda}_{\mathrm{T}}\mathbf{U}_{\mathrm{T}}{}^H\right)\otimes\left(\mathbf{U}_{\mathrm{T}}\boldsymbol{\Lambda}_{\mathrm{T}}\mathbf{U}_{\mathrm{T}}{}^H\right)\\ &= \underbrace{\left(\mathbf{U}_{\mathrm{T}}\otimes\mathbf{U}_{\mathrm{R}}\right)}_{\mathbf{U}_0}\underbrace{\left(\boldsymbol{\Lambda}_{\mathrm{T}}\otimes\boldsymbol{\Lambda}_{\mathrm{R}}\right)}_{\text{Eigenvalue matrix}}\left(\mathbf{U}_{\mathrm{T}}^H\otimes\mathbf{U}_{\mathrm{R}}^H\right)\\ &= \sum_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}}\sum_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}}\alpha_{n_{\mathrm{T}}}\beta_{n_{\mathrm{R}}}\left(\mathbf{a}_{n_{\mathrm{T}}}\otimes\mathbf{b}_{n_{\mathrm{R}}}\right)\left(\mathbf{a}_{n_{\mathrm{T}}}^H\otimes\mathbf{b}_{n_{\mathrm{R}}}^H\right), \quad (56)\end{aligned}$$

Based on (56), one can verify without difficulty that (20) is exactly the eigenvalue decomposition of $\boldsymbol{\Sigma}_{\mathbf{h}}$, which completes the proof.

Appendix F
Proof of Lemma 4

Given the new constraints $\mathbf{v}_{t+1} \in \left\{\sqrt{P}\mathbf{a}_{n_{\mathrm{T}}}^*\right\}_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}}$ and $\mathbf{w}_{t+1,k} \in \{\mathbf{b}_{n_{\mathrm{R}}}\}_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}}$ for all $k \in \{1,\cdots,N_{\mathrm{RF}}\}$, problem (18) can be reorganized as

$$\begin{aligned}\max_{\mathbf{v}_{t+1},\mathbf{W}_{t+1}}\ & f\left(\mathbf{v}_{t+1},\mathbf{W}_{t+1}\right)\\ \text{s.t. }& \mathbf{v}_{t+1} \in \left\{\sqrt{P}\mathbf{a}_{n_{\mathrm{T}}}^*\right\}_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}},\\ & \mathbf{w}_{t+1,k} \in \{\mathbf{b}_{n_{\mathrm{R}}}\}_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}}, \forall k \in \{1,\cdots,N_{\mathrm{RF}}\},\\ & \mathbf{w}_{t+1,k} \neq \mathbf{w}_{t+1,k'}, \forall k \neq k', \quad (57)\end{aligned}$$

where the objective function is given in (58), in which (a) holds according to the definition in (21) and (b) holds by utilizing the property to the property that $\left(\mathbf{A}\otimes\mathbf{B}\right)\left(\mathbf{C}\otimes\mathbf{D}\right) = \left(\mathbf{AC}\right)\otimes\left(\mathbf{BD}\right)$. Note that, the constraint $\mathbf{w}_{t+1,k} \neq \mathbf{w}_{t+1,k'}$ for all $k \neq k'$ in (57) ensures the orthogonality of $\mathbf{W}_{t+1}$. Observing (57), one can find that our goal becomes finding optimal indexes n_{T} and $\{n_{\mathrm{R},k}\}_{k=1}^{N_{\mathrm{RF}}}$ that maximize the MI increment $f\left(\mathbf{v}_{t+1},\mathbf{W}_{t+1}\right)$. Assuming that the optimal indexes

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{h}} = \mathsf{E}\left(\mathbf{h}\mathbf{h}^H\right) &\overset{(a)}{=} \frac{N_{\mathrm{T}}N_{\mathrm{R}}}{CR}\sum_{c=1}^{C}\sum_{r=1}^{R}\mathsf{E}\left(\left(\mathbf{b}^*\left(\varphi_{c,r}\right)\otimes\mathbf{a}\left(\theta_{c,r}\right)\right)\left(\mathbf{b}^T\left(\varphi_{c,r}\right)\otimes\mathbf{a}^H\left(\theta_{c,r}\right)\right)\right)\\ &\overset{(b)}{=} \frac{N_{\mathrm{T}}N_{\mathrm{R}}}{CR}\sum_{c=1}^{C}\sum_{r=1}^{R}\mathsf{E}\left(\mathbf{b}^*\left(\varphi_{c,r}\right)\mathbf{b}^T\left(\varphi_{c,r}\right)\right)\otimes\mathsf{E}\left(\mathbf{a}\left(\theta_{c,r}\right)\mathbf{a}^H\left(\theta_{c,r}\right)\right)\\ &\overset{(c)}{=} N_{\mathrm{T}}N_{\mathrm{R}}\mathsf{E}\left(\mathbf{b}^*\left(\varphi_{c,r}\right)\mathbf{b}^T\left(\varphi_{c,r}\right)\right)\otimes\mathsf{E}\left(\mathbf{a}\left(\theta_{c,r}\right)\mathbf{a}^H\left(\theta_{c,r}\right)\right)\overset{(d)}{=}\boldsymbol{\Sigma}_{\mathrm{T}}\otimes\boldsymbol{\Sigma}_{\mathrm{R}}. \quad (47)\end{aligned}$$

$$\begin{aligned}I(\mathbf{y};\mathbf{h}) &\overset{(a)}{=} \log_2\det\Bigg(\mathbf{I}_{N_{\mathrm{R}}N_{\mathrm{T}}} +\\ &\frac{1}{\sigma^2}\left[\mathbf{v}_1^*\otimes\mathbf{W}_1,\cdots,\mathbf{v}_Q^*\otimes\mathbf{W}_Q\right]\mathrm{blkdiag}\left(\left(\mathbf{W}_1^H\mathbf{W}_1\right)^{-1},\cdots,\left(\mathbf{W}_Q^H\mathbf{W}_Q\right)^{-1}\right)\left[\mathbf{v}_1^*\otimes\mathbf{W}_1,\cdots,\mathbf{v}_Q^*\otimes\mathbf{W}_Q\right]^H\boldsymbol{\Sigma}_{\mathbf{h}}\Bigg)\\ &\overset{(b)}{=} \log_2\det\left(\mathbf{I}_{N_{\mathrm{R}}N_{\mathrm{T}}} + \frac{1}{\sigma^2}\sum_{q=1}^{Q}\left(\left(\mathbf{v}_q^*\mathbf{v}_q^T\right)\otimes\left(\mathbf{W}_q\left(\mathbf{W}_q^H\mathbf{W}_q\right)^{-1}\mathbf{W}_q^H\right)\right)\boldsymbol{\Sigma}_{\mathbf{h}}\right). \quad (50)\end{aligned}$$

are expressed by $n_{\mathrm{T}}^{\mathrm{opt}}$ and $\{n_{\mathrm{R},k}^{\mathrm{opt}}\}_{k=1}^{N_{\mathrm{RF}}}$, the optimal precoder and the optimal combiner are

$$\mathbf{v}_{t+1}^{\mathrm{opt}} = \sqrt{P}\mathbf{a}_{n_{\mathrm{T}}^{\mathrm{opt}}}^{*} \text{ and } \mathbf{W}_{t+1}^{\mathrm{opt}} = \left[\mathbf{b}_{n_{\mathrm{R},1}^{\mathrm{opt}}}, \cdots, \mathbf{b}_{n_{\mathrm{R},N_{\mathrm{RF}}}^{\mathrm{opt}}}\right], \tag{59}$$

respectively. Then, we have

$$\mathbf{a}_{n_{\mathrm{T}}}^{H}(\mathbf{v}_{t+1}^{\mathrm{opt}})^{*} = \begin{cases} \sqrt{P}, & n_{\mathrm{T}} = n_{\mathrm{T}}^{\mathrm{opt}} \\ 0, & \text{else} \end{cases}, \tag{60a}$$

$$\mathbf{b}_{n_{\mathrm{R}}}^{H}\mathbf{W}_{t+1}^{\mathrm{opt}} = \begin{cases} \mathbf{e}_{n_R}^{T}, & n_{\mathrm{R}} \in \{n_{\mathrm{R},k}^{\mathrm{opt}}\}_{k=1}^{N_{\mathrm{RF}}} \\ \mathbf{0}_{N_{\mathrm{RF}}}^{T}, & \text{else} \end{cases}, \tag{60b}$$

where $\mathbf{e}_{n_R}$ denotes an N_{RF}-dimensional vector whose n_R-th entry is one and the other entries are zero. By substituting (60) into (57), the optimal MI increment $f\left(\mathbf{v}_{t+1}^{\mathrm{opt}}, \mathbf{W}_{t+1}^{\mathrm{opt}}\right)$ can be expressed by

$$\begin{aligned} & f\left(\mathbf{v}_{t+1}^{\mathrm{opt}}, \mathbf{W}_{t+1}^{\mathrm{opt}}\right) \\ &= \log_2 \det\left(\mathbf{I}_{N_{\mathrm{RF}}} + \frac{P}{\sigma^2}\sum_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}} \lambda_{t,n_{\mathrm{T}}^{\mathrm{opt}},n_{\mathrm{R}}} (\mathbf{W}_{t+1}^{\mathrm{opt}})^{H}\mathbf{b}_{n_{\mathrm{R}}}\mathbf{b}_{n_{\mathrm{R}}}^{H}\mathbf{W}_{t+1}^{\mathrm{opt}}\right) \\ &= \log_2 \det\left(\mathbf{I}_{N_{\mathrm{RF}}} + \frac{P}{\sigma^2}\mathrm{diag}\left(\lambda_{t,n_{\mathrm{T}}^{\mathrm{opt}},n_{\mathrm{R},1}^{\mathrm{opt}}}, \cdots, \lambda_{t,n_{\mathrm{T}}^{\mathrm{opt}},n_{\mathrm{R},N_{\mathrm{RF}}}^{\mathrm{opt}}}\right)\right) \\ &= \sum_{k=1}^{N_{\mathrm{RF}}} \log_2\left(1 + \frac{P\lambda_{t,n_{\mathrm{T}}^{\mathrm{opt}},n_{\mathrm{R},k}^{\mathrm{opt}}}}{\sigma^2}\right), \end{aligned} \tag{61}$$

which only relies on the eigenvalues of $\mathbf{\Sigma}_t$. In this context, the problem becomes finding n_{T} and $\{n_{\mathrm{R},k}\}_{k=1}^{N_{\mathrm{RF}}}$ that maximize $f\left(\mathbf{v}_{t+1}, \mathbf{W}_{t+1}\right)$, as formulated in (23). This completes the proof.

REFERENCES

[1] C. Huang, S. Hu, G. C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. D. Renzo, and M. Debbah, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[2] R. Deng, Y. Zhang, H. Zhang, B. Di, H. Zhang, H. V. Poor, and L. Song, "Reconfigurable holographic surfaces for ultra-massive MIMO in 6G: Practical design, optimization and implementation," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2367–2379, Aug. 2023.

[3] Z. Zhang and L. Dai, "Pattern-division multiplexing for multi-user continuous-aperture MIMO," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2350–2366, Aug. 2023.

[4] I. Kanbaz, O. Yurduseven, and M. Matthaiou, "Super-directive antenna arrays: How many elements do we need?" *arXiv preprint arXiv:2401.09179*, Jan. 2024.

[5] Z. Zhang and L. Dai, "Reconfigurable intelligent surfaces for 6G: Nine fundamental issues and one critical problem," *Tsinghua Sci. Technol.*, vol. 28, no. 5, pp. 929–939, Oct. 2023.

[6] K.-K. Wong and K.-F. Tong, "Fluid antenna multiple access," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 4801–4815, Jul. 2021.

[7] D. González-Ovejero, G. Minatti, G. Chattopadhyay, and S. Maci, "Multibeam by metasurface antennas," *IEEE Trans. Antennas Propag.*, vol. 65, no. 6, pp. 2923–2930, Jun. 2017.

[8] R.-B. Hwang, "Binary meta-hologram for a reconfigurable holographic metamaterial antenna," *Sci. Rep.*, vol. 10, no. 1, p. 8586, May 2020.

[9] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 162–169, Sep. 2018.

[10] M. Liu *et al.*, "Deeply subwavelength metasurface resonators for terahertz wavefront manipulation," *Adv. Opt. Mater.*, vol. 7, no. 21, p. 1900736, 2019.

[11] L. Wei, C. Huang, G. C. Alexandropoulos, W. E. I. Sha, Z. Zhang, M. Debbah, and C. Yuen, "Multi-user holographic MIMO surfaces: Channel modeling and spectral efficiency analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 1112–1124, Aug. 2022.

[12] J. An, C. Yuen, C. Huang, M. Debbah, H. Vincent Poor, and L. Hanzo, "A tutorial on holographic MIMO communications—part I: Channel modeling and channel estimation," *IEEE Commun. Lett.*, vol. 27, no. 7, pp. 1664–1668, Jul. 2023.

[13] J. An, C. Yuen, C. Huang, M. Debbah, H. V. Poor, and L. Hanzo, "A tutorial on holographic MIMO communications—part II: Performance analysis and holographic beamforming," *IEEE Commun. Lett.*, vol. 27, no. 7, pp. 1669–1673, Jul. 2023.

[14] Y. Liu, M. Zhang, T. Wang, A. Zhang, and M. Debbah, "Densifying MIMO: Channel modeling, physical constraints, and performance evaluation for holographic communications," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 6, pp. 1504–1518, Jun. 2024.

[15] M. Di Renzo, D. Dardari, and N. Decarli, "LoS MIMO-arrays vs. LoS MIMO-surfaces," in *Proc. 17th European Conference on Antennas and Propagation (EuCAP'23)*, 2023, pp. 1–5.

[16] J. Xie, H. Yin, and L. Han, "A genetic algorithm based superdirective beamforming method under excitation power range constraints," *arXiv preprint arXiv:2307.02063*, Jul. 2023.

[17] M. Akrout, V. Shyianov, F. Bellili, A. Mezghani, and R. W. Heath, "Super-wideband massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2414–2430, Aug. 2023.

[18] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. U.K.: Cambridge University Press, 2005.

[19] M. Cui, Z. Wu, Y. Lu, X. Wei, and L. Dai, "Near-field MIMO communications for 6G: Fundamentals, challenges, potentials, and future directions," *IEEE Commun. Mag.*, vol. 61, no. 1, pp. 40–46, Jan. 2022.

[20] T. Zheng, J. Zhu, Q. Yu, Y. Yan, and L. Dai, "Coded beam training," *arXiv preprint arXiv:2401.01673*, Mar. 2024.

[21] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. USA: Prentice-Hall, Inc., 1993.

[22] C. Huang, L. Liu, C. Yuen, and S. Sun, "Iterative channel estimation using LSE and sparse message passing for mmWave MIMO systems," *IEEE Trans. Signal Process.*, vol. 67, no. 1, pp. 245–259, Jan. 2019.

[23] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764–779, Jan. 2020.

[24] M. Cui and L. Dai, "Channel estimation for extremely large-scale MIMO: Far-field or near-field?" *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2663–2677, Apr. 2022.

[25] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, Oct. 2019.

[26] Y. Jin, J. Zhang, S. Jin, and B. Ai, "Channel estimation for cell-free mmwave massive MIMO through deep learning," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 10 325–10 329, Oct. 2019.

[27] Z. Wan, Z. Gao, B. Shim, K. Yang, G. Mao, and M.-S. Alouini, "Compressive sensing based channel estimation for millimeter-wave full-dimensional MIMO with lens-array," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 2337–2342, Feb. 2020.

[28] X. Ma, Z. Gao, F. Gao, and M. Di Renzo, "Model-driven deep learning based channel estimation and feedback for millimeter-wave massive

$$\begin{aligned} f\left(\mathbf{v}_{t+1}, \mathbf{W}_{t+1}\right) &\overset{(a)}{=} \log_2 \det\left(\mathbf{I}_{N_{\mathrm{RF}}} + \frac{1}{\sigma^2}\sum_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}}\sum_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}} \lambda_{t,n_{\mathrm{T}},n_{\mathrm{R}}}\left(\mathbf{v}_{t+1}^{T}\otimes\mathbf{W}_{t+1}^{H}\right)\left(\mathbf{a}_{n_{\mathrm{T}}}\otimes\mathbf{b}_{n_{\mathrm{R}}}\right)\left(\mathbf{a}_{n_{\mathrm{T}}}^{H}\otimes\mathbf{b}_{n_{\mathrm{R}}}^{H}\right)\left(\mathbf{v}_{t+1}^{*}\otimes\mathbf{W}_{t+1}\right)\right) \\ &\overset{(b)}{=} \log_2 \det\left(\mathbf{I}_{N_{\mathrm{RF}}} + \frac{1}{\sigma^2}\sum_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}}\sum_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}} \lambda_{t,n_{\mathrm{T}},n_{\mathrm{R}}}\left|\mathbf{a}_{n_{\mathrm{T}}}^{H}\mathbf{v}_{t+1}^{*}\right|^2\mathbf{W}_{t+1}^{H}\mathbf{b}_{n_{\mathrm{R}}}\mathbf{b}_{n_{\mathrm{R}}}^{H}\mathbf{W}_{t+1}\right) \end{aligned} \tag{58}$$

hybrid MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2388–2406, Aug. 2021.
[29] C. Williams and C. Rasmussen, "Gaussian processes for regression," in *Adv. Neural Inf. Process. Syst.*, vol. 8, 1995.
[30] W. K. New, K.-K. Wong, H. Xu, F. R. Ghadi, R. Murch, and C.-B. Chae, "Channel estimation and reconstruction in fluid antenna system: Oversampling is essential," *IEEE J. Sel. Areas Commun.*, Nov. 2024.
[31] J. An, C. Yuen, C. Huang, M. Debbah, H. V. Poor, and L. Hanzo, "A tutorial on holographic MIMO communications—part III: Open opportunities and challenges," *IEEE Commun. Lett.*, vol. 27, no. 7, pp. 1674–1678, Jul. 2023.
[32] M. Cui, Z. Zhang, L. Dai, and K. Huang, "Ice-filling: Near-optimal channel estimation for dense array systems," *arXiv preprint arXiv:2404.06806*, Apr. 2024.
[33] A. F. Molisch, K. Balakrishnan, C.-C. Chong, S. Emami, A. Fort, J. Karedal, J. Kunisch, H. Schantz, U. Schuster, and K. Siwiak, "IEEE 802.15. 4a channel model-final report," *IEEE P802*, vol. 15, no. 4, p. 0662, 2004.
[34] 3GPP TR, "Study on channel model for frequencies from 0.5 to 100 GHz," *3GPP TR 38.901 version 14.0.0 Release*, Dec. 2019.
[35] K. Werner and M. Jansson, "Estimating MIMO channel covariances from training data under the kronecker model," *Signal Process.*, vol. 89, no. 1, pp. 1–13, Jan. 2009.
[36] S. Park and R. W. Heath, "Spatial channel covariance estimation for the hybrid MIMO architecture: A compressive sensing-based approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8047–8062, Dec. 2018.
[37] K. Upadhya and S. A. Vorobyov, "Covariance matrix estimation for massive MIMO," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 546–550, Apr. 2018.
[38] M. B. Khalilsarai, T. Yang, S. Haghighatshoar, and G. Caire, "Structured channel covariance estimation from limited samples in massive mimo," in *Proc. IEEE Int. Conf. Commun. (ICC'20)*, Jun. 2020, pp. 1–7.
[39] J. Zhu, X. Su, Z. Wan, L. Dai, and T. J. Cui, "The benefits of electromagnetic information theory for channel estimation," in *Proc. IEEE Int. Conf. Commun. (ICC'24)*, Jun. 2024, pp. 4869–4874.
[40] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3250–3265, May 2012.