Observation Matrix Design for Densifying MIMO Channel Estimation via 2D Ice Filling

Zijian Zhang, *Graduate Student Member, IEEE,* and Mingyao Cui, *Graduate Student Member, IEEE*

***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 fact 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 and TS-2DIF methods 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) has attracted considerable attention from the wireless communication community [2], [3], [4], [5], [6], [7]. 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 $\lambda/6$ [8], $\lambda/8$ [9], $\lambda/10$ [10], or even $\lambda/23$ [11]. 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) [12], holographic reconfigurable surfaces (RHSs) [3], continuous-aperture MIMO (CAP-MIMO) [4], superdirective antenna arrays [5], reconfigurable intelligent surfaces (RISs) [13], and fluid antenna systems (FASs) [7]. Particularly, in high-frequency (e.g., millimeter-wave or terahertz) communications, densifying MIMO systems are common due to their shorter wavelengths, reduced grating lobes, and enhanced beamforming precision [2]. Utilizing the extensive channel observations facilitated by a multitude of antennas, densifying MIMO is anticipated to achieve significant array gains and multiplexing-diversity gains [14], [15], [16], [17]. Furthermore, densifying MIMO can mitigate the effects of grating lobes and offer enhanced performance for large oblique angles of incidence [18]. Some studies have also highlighted their capabilities to realize super-directivity [5], [19] or super-bandwidth [20] 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 [21]. To implement effective precoding/combining, an indispensable technology for MIMO systems is the acquisition of channel state information (CSI) [22], [23]. 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 [24], such as the least square (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 [25], [26]. Relevant techniques include the orthogonal matching pursuit (OMP)-based estimator [27] and the approximate message passing (AMP)-based estimator [25], [28]. 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 [29], [30], [31].

Received 14 July 2025; revised 20 December 2025; accepted 19 January 2026. Date of publication 23 January 2026; date of current version 27 February 2026. This work was supported by the National Natural Science Foundation of China under Grant 624B2123. An earlier version of this paper was presented in part at the IEEE ICC'25, Montreal, Canada, 2025 [DOI:10.1109/ICC52391.2025.11161466]. Compared with the conference version [1], this journal version provides theoretical analysis, generalized channel modeling, and a new algorithm for hybrid MIMO architectures. The associate editor coordinating the review of this article and approving it for publication was Rodrigo C. de Lamare. *(Corresponding author: Mingyao Cui.)*

Zijian Zhang is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: zhangzij15@tsinghua.org.cn).

Mingyao Cui is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR 999077, China (e-mail: mycui@eee.hku.hk).

Digital Object Identifier 10.1109/TSP.2026.3657342

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 [24]. 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 [17]. Besides, the circuit mutual coupling induces signal interactions between adjacent antennas, which will influence the channel correlation in densifying MIMO systems [8], [9], [10], [11], [32]. These facts lead to the highly structured covariance matrices of densifying MIMO channels. Existing works have revealed that, such an informative covariance matrix can provide appreciable prior knowledge for the specific design of observation matrices (e.g., combiners and precoders) in channel estimation, thus significantly improving the accuracy of CSI acquisition [33], [34], [35].

To exploit the strong channel correlations for improved CSI acquisition, our prior work [36] proposes an ice filling (IF) based observation matrix design in dense array systems, which is inspired by the idea of Gaussian Process Regression (GPR). By maximizing the mutual information (MI) between the received pilot and wireless channel, which is characterized by the channel covariance 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 [36]. Nevertheless, 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 [36]. 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 covariance matrices 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 densifying MIMO channel estimation. Our key idea is to maximize the MI 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 [36] 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 analog 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. Different from the exiting works on hybrid MIMO beamforming that precoders and combiners can be designed independently, in our considered observation matrix design, the beamformer and the combiner are coupled due to a Kronecker-product structure. To overcome this challenge, the proposed TS-2DIF alternately optimizes the analog combiner, the digital combiner, and the precoder at the transceivers to approach the ideal observation matrix.

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.

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 a densifying MIMO system, consisting of an N_{R}-antenna base station (BS) equipped with N_{RF} RF chains and an N_{T}-antenna user. 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 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

$$\begin{aligned}\mathbf{y}_q &= \mathbf{W}_q^H \mathbf{H} \mathbf{v}_q s_q + \mathbf{W}_q^H \mathbf{z}_q \\ &= \left(\mathbf{v}_q^T \otimes \mathbf{W}_q^H\right) \mathbf{h} s_q + \mathbf{W}_q^H \mathbf{z}_q,\end{aligned} \tag{1}$$

where $\mathbf{h} \equiv \mathrm{vec}(\mathbf{H})$ is defined as the vectorized channel matrix, s_q 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 power constraint is satisfied: $\|\mathbf{v}_q\|^2 = P$, with P the per-pilot transmit power. 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, and 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 phases of $\mathbf{A}_q$ are adjustable, which imposes a structural constraint on the feasible set of $\mathbf{W}_q = \mathbf{A}_q \mathbf{D}_q$.

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

$$\mathbf{y} = \mathbf{X}^H \mathbf{h} + \mathbf{z}, \tag{2}$$

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 $\{\mathbf{v}_q\}_{q=1}^Q$.

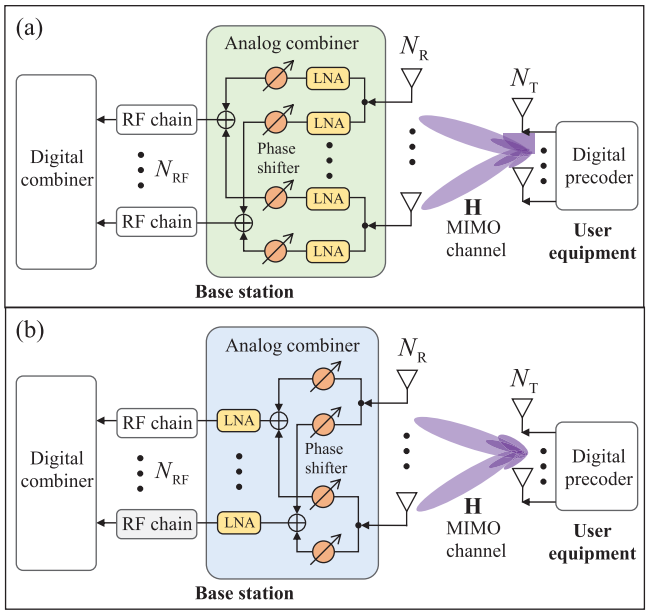

Fig. 1. An illustration of hybrid analog and digital MIMO architectures, where the analog combiner of BS can adopt either: (a) the amplitude-and-phase controllable structure; or (b) the phase-only controllable structure.

B. Channel Model

We consider the general correlated Rayleigh-fading channel model, which simultaneously takes the spatial correlation, spherical wavefront, and electromagnetic mutual coupling into account. Following this well-known model [37], [38], [39], the channel $\mathbf{H}$ between the BS and the user is expressed as:

$$\mathbf{H} = \mathbf{C}_{\mathrm{rx}}^{1/2} \mathbf{R}_{\mathrm{rx}}^{1/2} \mathbf{H}_{\mathrm{iid}} \mathbf{R}_{\mathrm{tx}}^{1/2} \mathbf{C}_{\mathrm{tx}}^{1/2}, \tag{3}$$

where $\mathbf{R}_{\mathrm{rx}} \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{R}}}$ and $\mathbf{R}_{\mathrm{tx}} \in \mathbb{C}^{N_{\mathrm{T}} \times N_{\mathrm{T}}}$ are the spatial correlation matrices at the BS and the user, respectively; $\mathbf{C}_{\mathrm{rx}} \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{R}}}$ and $\mathbf{C}_{\mathrm{tx}} \in \mathbb{C}^{N_{\mathrm{T}} \times N_{\mathrm{T}}}$ characterize the electromagnetic mutual coupling among the BS antennas and the user antennas, respectively; and $\mathbf{H}_{\mathrm{iid}}$ is a complex Gaussian matrix in which each elements are i.i.d with zero mean and unit variance. This channel model is widely recognized by the studies on correlated MIMO channels, also adopted by commercial channel simulators, such as Matlab Antenna toolbox [40].

Example 1 (Example of settings): Note that, the settings of $\mathbf{R}_{\mathrm{rx}}$, $\mathbf{R}_{\mathrm{tx}}$, $\mathbf{C}_{\mathrm{rx}}$, and $\mathbf{C}_{\mathrm{tx}}$ depend on the specific transmission scenarios in practice. As a typical example, the (m,n)-th entry of $\mathbf{R}_{\mathrm{rx}}$ is usually modeled as [12], [15], [41]:

$$\mathbf{R}_{\mathrm{rx}}(m,n) = \int \int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} f_{\mathrm{rx}}(\varphi, \theta) e^{j\mathbf{k}(\varphi,\theta)^T (\mathbf{r}_m - \mathbf{r}_n)} \mathrm{d}\theta \mathrm{d}\varphi, \tag{4}$$

where φ and θ denote the azimuth angle and elevation angle of incident signals, respectively; $f_{\mathrm{rx}}(\varphi, \theta)$ is the scenario-dependent spatial-scattering function,

which describes the angular multipath distribution and the directivity gain of receiver antennas in practice; $\mathbf{k}(\varphi,\theta) := \frac{2\pi}{\lambda}(\cos\theta\cos\varphi, \cos\theta\sin\varphi, \sin\theta)^T$ is the wave vector with λ being the wavelength; and $\mathbf{r}_m \in \mathbb{R}^3$ is the location of the m-th receiver antenna. To characterize the mutual coupling, $\mathbf{C}_{\rm rx}$ is modeled as [37], [42]:

$$\mathbf{C}_{\rm rx} = \left(\mathbf{Z}_{\rm rx} + R_{\rm rx}\mathbf{I}\right)^{-1}, \tag{5}$$

where $\mathbf{Z}_{\rm rx}$ is the mutual impedance matrix and $R_{\rm rx} > 0$ is the dissipation resistance of antennas. Given the specific array parameters, $\mathbf{Z}_{\rm rx}$ and $R_{\rm rx}$ can be calculated by the well-known induced electromagnetic fields (EMF) method [43]. Similar settings can be adopted to generate $\mathbf{R}_{\rm tx}$ and $\mathbf{C}_{\rm tx}$.

C. Problem Formulation

As antennas of densifying MIMO are packed within a small area, the channels across close-by antennas are strongly correlated. Define the covariance of vectorized channel $\mathbf{h} = \mathrm{vec}(\mathbf{H})$ as $\mathsf{E}\left(\mathbf{h}\mathbf{h}^H\right) = \mathbf{\Sigma}_{\mathbf{h}} \in \mathbb{C}^{N_{\rm R}N_{\rm T}\times N_{\rm R}N_{\rm T}}$, which is also called the *kernel* of channel. The high-correlation property of wireless channel indicates that the kernel $\mathbf{\Sigma}_{\mathbf{h}}$ can provide prior knowledge to achieve accurate channel estimation [44], [45], [46], [47]. To this end, we follow the idea of GPR to design the estimator and the observation matrix. According to the model in (3), the channel is sampled from the Gaussian process $\mathcal{CN}\left(\mathbf{0}_{N_{\rm R}N_{\rm T}}, \mathbf{\Sigma}_{\mathbf{h}}\right)$. The joint probability distribution of $\mathbf{h}$ and $\mathbf{y}$ then satisfies

$$\begin{bmatrix}\mathbf{h}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{CN}\left(\begin{bmatrix}\mathbf{0}_{N_{\rm R}N_{\rm T}}\\ \mathbf{0}_{N_{\rm RF}Q}\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma}_{\mathbf{h}} & \mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}\\ \mathbf{X}^H\mathbf{\Sigma}_{\mathbf{h}} & \mathbf{X}^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}+\mathbf{\Xi}\end{bmatrix}\right), \tag{6}$$

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 $\mathbf{z}$. Then, the posterior mean and the posterior covariance of $\mathbf{h}$ are expressed as

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

$$\mathbf{\Sigma}_{\mathbf{h}|\mathbf{y}} = \mathbf{\Sigma}_{\mathbf{h}} - \mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}\left(\mathbf{X}^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}+\mathbf{\Xi}\right)^{-1}\mathbf{X}^H\mathbf{\Sigma}_{\mathbf{h}}. \tag{8}$$

The posterior mean $\boldsymbol{\mu}_{\mathbf{h}|\mathbf{y}}$ is exactly the adopted channel estimator, which is equivalent to the linear MMSE (LMMSE) estimator. Besides, the posterior covariance $\mathbf{\Sigma}_{\mathbf{h}|\mathbf{y}}$ characterizes the estimation error, which is largely dependent on the observation matrix $\mathbf{X}$. This dependence indicates that 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 $\mathbf{y}$. Henceforth, our objective is to find $\{\mathbf{W}_q\}_{q=1}^Q$ and $\{\mathbf{v}_q\}_{q=1}^Q$ that maximize the MI[1] between $\mathbf{y}$ and $\mathbf{h}$, which is formulated as:

$$\max_{\mathbf{W}\in\mathcal{W},\|\mathbf{v}_q\|^2=P} I(\mathbf{y};\mathbf{h}) = \log\det\left(\mathbf{I}_{N_{\rm RF}Q} + \mathbf{\Xi}^{-1}\mathbf{X}^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}\right). \tag{9}$$

[1]According to the information-theoretic properties [48], the maximization of MI is asymptotically the minimization of Cramer-Rao bound.

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 (9), 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 $\mathbf{\Xi}$ introduced by combiners $\{\mathbf{W}_q\}_{q=1}^Q$, the global optimal solution to (9) 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 $t+1$:

$$\max_{\mathbf{W}_{t+1},\mathbf{v}_{t+1}} \Delta I_{t+1} := I(\bar{\mathbf{y}}_{t+1};\mathbf{h}) - I(\bar{\mathbf{y}}_t;\mathbf{h}). \tag{10}$$

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.

1) When $t = 1$: For ease of understanding, we begin with handling the first timeslot, i.e., $t = 1$. In this context, problem (10) can be rewritten as

$$\max_{\|\mathbf{v}_1\|^2=P,\ \mathbf{W}_1} I(\mathbf{y}_1;\mathbf{h}) \tag{11}$$

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

$$I(\mathbf{y}_1;\mathbf{h}) = \log\det\left(\mathbf{I}_{N_{\rm RF}} + \frac{1}{\sigma^2}\left(\mathbf{W}_1^H\mathbf{W}_1\right)^{-1} \times \left(\mathbf{v}_1^T\otimes\mathbf{W}_1^H\right)\mathbf{\Sigma}_{\mathbf{h}}\left(\mathbf{v}_1^*\otimes\mathbf{W}_1\right)\right). \tag{12}$$

Since the reformulated problem (11) is still intricate, we seek to simplify it using the following lemmas.

Lemma 1: For the correlated Rayleigh-fading model in (3), the covariance of the vectored channel $\mathbf{\Sigma}_{\mathbf{h}}$ can be rewritten as the form of a Kronecker product of two kernels, i.e.,

$$\mathbf{\Sigma}_{\mathbf{h}} = \mathbf{\Sigma}_{\rm T}\otimes\mathbf{\Sigma}_{\rm R}, \tag{13}$$

Algorithm 1 2DIF Based Combiner and Precoder Design

Input: Number of pilots Q, kernel $\mathbf{\Sigma_h}$.

Output: Designed precoders $\{\mathbf{v}_q^{\rm opt}\}_{q=1}^Q$ and combiners $\{\mathbf{W}_q^{\rm opt}\}_{q=1}^Q$.

1: Rewrite kernel as $\mathbf{\Sigma_h} = \mathbf{\Sigma}_{\rm T} \otimes \mathbf{\Sigma}_{\rm R}$

2: Find the eigenvectors $[\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{N_{\rm T}}]$ and the corresponding eigenvalues $[\alpha_1, \alpha_2, \cdots, \alpha_{N_{\rm T}}]$ of $\mathbf{\Sigma}_{\rm T}$

3: Find the eigenvectors $[\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_{N_{\rm R}}]$ and the corresponding eigenvalues $[\beta_1, \beta_2, \cdots, \beta_{N_{\rm R}}]$ of $\mathbf{\Sigma}_{\rm R}$

4: Initialize: $[\lambda_{1,1}^0, \lambda_{1,2}^0, \cdots, \lambda_{N_{\rm T},N_{\rm R}}^0] = [\alpha_1\beta_1, \alpha_1\beta_2, \cdots, \alpha_{N_{\rm T}}\beta_{N_{\rm R}}]$

5: **for** $t = 0, \cdots, Q-1$ **do**

6: Find the optimal $n_{\rm T}^{\rm opt}$ and $\{n_{{\rm R},k}^{\rm opt}\}_{k=1}^{N_{\rm RF}}$ via **Algorithm 2**

7: Eigenvector-assignment: $\mathbf{v}_{t+1}^{\rm opt} = \sqrt{P}\mathbf{a}_{n_{\rm T}^{\rm opt}}^*$ and $\mathbf{W}_{t+1}^{\rm opt} = [\mathbf{b}_{n_{{\rm R},1}^{\rm opt}}, \cdots, \mathbf{b}_{n_{{\rm R},N_{\rm RF}}^{\rm opt}}]$

8: Eigenvalue-update for all $n_{\rm T} \in \{1, \cdots, N_{\rm T}\}$ and $n_{\rm R} \in \{1, \cdots, N_{\rm R}\}$ via

9: $\lambda_{n_{\rm T},n_{\rm R}}^{t+1} = \begin{cases} \frac{\lambda_{n_{\rm T},n_{\rm R}}^t \sigma^2}{P\lambda_{n_{\rm T},n_{\rm R}}^t + \sigma^2}, & n_{\rm T} = n_{\rm T}^{\rm opt} \ \&\ n_{\rm R} \in \{n_{{\rm R},k}^{\rm opt}\}_{k=1}^{N_{\rm RF}}, \\ \lambda_{n_{\rm T},n_{\rm R}}^t, & \text{else.} \end{cases}$

10: **end for**

11: **return** Designed precoders $\{\mathbf{v}_q^{\rm opt}\}_{q=1}^Q$ and combiners $\{\mathbf{W}_q^{\rm opt}\}_{q=1}^Q$ for channel estimation.

where $\mathbf{\Sigma}_{\rm T} := (\mathbf{C}_{\rm tx}^{1/2})^T \mathbf{R}_{\rm tx}^* (\mathbf{C}_{\rm tx}^{1/2})^* \in \mathbb{C}^{N_{\rm T}\times N_{\rm T}}$ and $\mathbf{\Sigma}_{\rm R} := \mathbf{C}_{\rm rx}^{1/2}\mathbf{R}_{\rm rx}(\mathbf{C}_{\rm rx}^{1/2})^H \in \mathbb{C}^{N_{\rm R}\times N_{\rm R}}$ characterize the channel correlation among transmit antennas and that among receive antennas, respectively.

Proof: (See Appendix A). ■

Lemma 2: Introducing the orthogonality constraints $\mathbf{W}_q^H\mathbf{W}_q = \mathbf{I}_{N_{\rm RF}}$ for all $q \in \{1, \cdots, Q\}$ does not influence the optimal value of MI $I(\mathbf{y}; \mathbf{h})$ in (9).

Proof: (See Appendix B). ■

Utilizing **Lemma 1** and **Lemma 2**, problem (11) can be equivalently rewritten as

$$\begin{aligned} \max_{\mathbf{v}_1, \mathbf{W}_1} \ & I(\mathbf{y}_1; \mathbf{h}) = \log\det\left(\mathbf{I}_{N_{\rm RF}} + \frac{\mathbf{v}_1^T\mathbf{\Sigma}_{\rm T}\mathbf{v}_1^*}{\sigma^2}\mathbf{W}_1^H\mathbf{\Sigma}_{\rm R}\mathbf{W}_1\right) \\ \text{s.t.} \ & \|\mathbf{v}_1\|^2 = P, \\ & \mathbf{W}_1^H\mathbf{W}_1 = \mathbf{I}_{N_{\rm RF}}, \end{aligned} \tag{14}$$

where the reorganized objective function $I(\mathbf{y}_1; \mathbf{h})$ is obtained by substituting $\mathbf{\Sigma_h} = \mathbf{\Sigma}_{\rm T} \otimes \mathbf{\Sigma}_{\rm R}$ and $\mathbf{W}_1^H\mathbf{W}_1 = \mathbf{I}_{N_{\rm RF}}$ into (11). Observing (14), one can prove with ease that the optimal $\mathbf{v}_1^*$ is the eigenvector of $\mathbf{\Sigma}_{\rm T}$ associated with its largest eigenvalue, and the optimal $\mathbf{W}_1$ is composed of the eigenvectors of $\mathbf{\Sigma}_{\rm R}$ associated with its top $N_{\rm RF}$ eigenvalues. Define the eigenvalue decompositions (EVDs) $\mathbf{\Sigma}_{\rm T} = \mathbf{U}_{\rm T}\mathbf{\Lambda}_{\rm T}\mathbf{U}_{\rm T}^H$ and $\mathbf{\Sigma}_{\rm R} = \mathbf{U}_{\rm R}\mathbf{\Lambda}_{\rm R}\mathbf{U}_{\rm R}^H$, wherein the eigenvalues in $\mathbf{\Lambda}_{\rm T} = {\rm diag}\,(\alpha_1, \cdots, \alpha_{N_{\rm T}})$ and $\mathbf{\Lambda}_{\rm R} = {\rm diag}\,(\beta_1, \cdots, \beta_{N_{\rm R}})$ are arranged in descending order. In this way, the maximum MI in (14) can be derived as

$$\max_{\mathbf{v}_1, \mathbf{W}_1} I(\mathbf{y}_1; \mathbf{h}) = \sum_{n=1}^{N_{\rm RF}} \log_2\left(1 + \frac{P\alpha_1\beta_n}{\sigma^2}\right), \tag{15}$$

and its achievable solution are expressed as

$$\mathbf{v}_1^{\rm opt} = \sqrt{P}\mathbf{U}_{\rm T}^*(:, 1) \ \text{and} \ \mathbf{W}_1^{\rm opt} = \mathbf{U}_{\rm R}(:, [1, \cdots, N_{\rm RF}]). \tag{16}$$

Equation (16) can be viewed as the optimal initialization settings for our proposed observation matrix design.

2) From t to $t+1$: Given the optimized 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. By invoking the principle of recursion, we only need to address the design of precoder $\mathbf{v}_{t+1}$ and combiner $\mathbf{W}_{t+1}$ with 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

$$\Delta I_{t+1} = \log_2\det\left(\mathbf{I}_{N_{\rm RF}} + \frac{1}{\sigma^2}\mathbf{X}_{t+1}^H\mathbf{\Sigma}_t\mathbf{X}_{t+1}\right), \tag{17}$$

wherein $\mathbf{X}_{t+1} := \mathbf{v}_{t+1}^* \otimes \mathbf{W}_{t+1}$ refers to the Kronecker-constrained observation matrix, and

$$\mathbf{\Sigma}_t = \mathbf{\Sigma_h} - \mathbf{\Sigma_h}\bar{\mathbf{X}}_t\left(\bar{\mathbf{X}}_t^H\mathbf{\Sigma_h}\bar{\mathbf{X}}_t + \sigma^2\mathbf{I}_{N_{\rm RF}t}\right)^{-1}\bar{\mathbf{X}}_t^H\mathbf{\Sigma_h} \tag{18}$$

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

$$\begin{aligned} \max_{\mathbf{v}_{t+1}, \mathbf{W}_{t+1}} \ & \log\det\left(\mathbf{I}_{N_{\rm RF}} + \frac{1}{\sigma^2}(\mathbf{v}_{t+1}^T \otimes \mathbf{W}_{t+1}^H)\mathbf{\Sigma}_t(\mathbf{v}_{t+1}^* \otimes \mathbf{W}_{t+1})\right) \\ \text{s.t.} \ & \|\mathbf{v}_{t+1}\|^2 = P, \\ & \mathbf{W}_{t+1}^H\mathbf{W}_{t+1} = \mathbf{I}_{N_{\rm RF}}. \end{aligned} \tag{19}$$

Note that, in problem (14), the kernel $\mathbf{\Sigma_h}$ is decoupled into $\mathbf{\Sigma}_{\rm T} \otimes \mathbf{\Sigma}_{\rm R}$ such that $\mathbf{v}_1$ and $\mathbf{W}_1$ can be obtained by selecting the appropriate eigenvectors of $\mathbf{\Sigma}_{\rm T}$ and $\mathbf{\Sigma}_{\rm R}$ as in (16). We attempt the similar idea to solve for $\mathbf{v}_{t+1}$ and $\mathbf{W}_{t+1}$. To this end, we first define the EVD: $\mathbf{\Sigma}_t = \mathbf{U}_t\mathbf{\Lambda}_t\mathbf{U}_t^H$ where $\mathbf{U}_t \in \mathbb{C}^{N_{\rm T}N_{\rm R}\times N_{\rm T}N_{\rm R}}$. Notice that the constraints $\|\mathbf{v}_{t+1}\|^2 = P$ and $\mathbf{W}_{t+1}^H\mathbf{W}_{t+1} = \mathbf{I}_{N_{\rm RF}}$ make the overall matrix observation $\mathbf{X}_{t+1}$ orthogonal, i.e., $\mathbf{X}_{t+1}^H\mathbf{X}_{t+1} = P\mathbf{I}_{\rm RF}$. If we temporarily omit the Kronecker constraint $\mathbf{X}_{t+1} = \mathbf{v}_{t+1}^* \otimes \mathbf{W}_{t+1}$ and try to solve (19) by considering the orthogonal constraint $\mathbf{X}_{t+1}^H\mathbf{X}_{t+1} = P\mathbf{I}_{\rm RF}$ only, it becomes evident that the global optimal solution to (19) is $\mathbf{X}_{t+1} = \sqrt{P}\mathbf{U}_t(:, [1, \cdots, N_{\rm RF}])$. 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) Influence of setting $\mathbf{X}_{t+1}$ as the principal eigenvectors of $\mathbf{\Sigma}_t$. Before evaluating the feasibility of setting $\mathbf{X}_{t+1} = \sqrt{P}\mathbf{U}_t(:, [1, \cdots, N_{\rm RF}])$, we first need to exploit its influence on the evolution rule of the posterior kernel $\mathbf{\Sigma}_{t+1}$. The following lemma characterizes the relationship between $\mathbf{\Sigma}_{t+1}$ and $\mathbf{\Sigma}_t$.

Lemma 3: Let $\lambda_n(\cdot)$ denote the n-th largest eigenvalue of the matrix in its argument, e.g., $\lambda_n(\mathbf{\Sigma}_t) = \mathbf{\Lambda}_t(n, n)$ 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

$$\begin{aligned}\mathbf{\Sigma}_{t+1} = \mathbf{U}_t \mathrm{diag}\Bigg(& \frac{\lambda_1\left(\mathbf{\Sigma}_t\right)\sigma^2}{P\lambda_1\left(\mathbf{\Sigma}_t\right)+\sigma^2}, \frac{\lambda_2\left(\mathbf{\Sigma}_t\right)\sigma^2}{P\lambda_2\left(\mathbf{\Sigma}_t\right)+\sigma^2}, \cdots, \\ & \frac{\lambda_{N_{\mathrm{RF}}}\left(\mathbf{\Sigma}_t\right)\sigma^2}{P\lambda_{N_{\mathrm{RF}}}\left(\mathbf{\Sigma}_t\right)+\sigma^2}, \lambda_{N_{\mathrm{RF}}+1}\left(\mathbf{\Sigma}_t\right), \lambda_{N_{\mathrm{RF}}+2}\left(\mathbf{\Sigma}_t\right), \\ & \cdots, \lambda_{N_{\mathrm{R}}N_{\mathrm{T}}}\left(\mathbf{\Sigma}_t\right)\Bigg)\mathbf{U}_t^H. \end{aligned} \tag{20}$$

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 $\left\{\frac{\lambda_n(\mathbf{\Sigma}_t)\sigma^2}{P\lambda_n(\mathbf{\Sigma}_t)+\sigma^2}\right\}_{n=1}^{N_{\mathrm{RF}}}$ in $\mathbf{\Sigma}_{t+1}$. Considering the generality of t, we can conclude that, the eigenvectors of channel covariance $\mathbf{\Sigma}_0 := \mathbf{\Sigma}_{\mathbf{h}}$ 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 $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_{\mathbf{h}}$. To assess this feasibility, we examine the structure of the eigenspace of $\mathbf{\Sigma}_{\mathbf{h}}$ below.

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

$$\begin{aligned}\mathbf{\Sigma}_{\mathbf{h}} &= \mathbf{U}_0\mathbf{\Lambda}_0\mathbf{U}_0^H \\ &= \underbrace{\left(\mathbf{U}_{\mathrm{T}} \otimes \mathbf{U}_{\mathrm{R}}\right)}_{\text{Orthogonal matrix}} \underbrace{\left(\mathbf{\Lambda}_{\mathrm{T}} \otimes \mathbf{\Lambda}_{\mathrm{R}}\right)}_{\text{Eigenvalue matrix}} \left(\mathbf{U}_{\mathrm{T}}^H \otimes \mathbf{U}_{\mathrm{R}}^H\right) \\ &= \sum_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}}\sum_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}} \alpha_{n_{\mathrm{T}}}\beta_{n_{\mathrm{R}}}\left(\mathbf{a}_{n_{\mathrm{T}}} \otimes \mathbf{b}_{n_{\mathrm{R}}}\right)\left(\mathbf{a}_{n_{\mathrm{T}}}^H \otimes \mathbf{b}_{n_{\mathrm{R}}}^H\right), \end{aligned} \tag{21}$$

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 $\mathbf{\Sigma}_{\mathbf{h}}$, 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 $\mathbf{\Sigma}_{\mathbf{h}}$, 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 (19) 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 $\mathbf{X}_1 = \sqrt{P}\mathbf{U}_0\left(:, [1, \cdots, N_{\mathrm{RF}}]\right)$ at timeslot $t = 1$, we can set $\mathbf{v}_1 = \sqrt{P}\mathbf{a}_1^* = \sqrt{P}\mathbf{U}_{\mathrm{T}}^*\left(:, 1\right)$ and $\mathbf{W}_1 = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_{N_{\mathrm{RF}}}] = \mathbf{U}_{\mathrm{R}}\left(:, [1, \cdots, N_{\mathrm{RF}}]\right)$, which coincides with the optimal solution in (16). Note that, according to **Lemma 3**, this process does not changes the eigenvectors of $\mathbf{\Sigma}_1$.[2] In this context, the desired $\mathbf{X}_2 = \sqrt{P}\mathbf{U}_1\left(:, [1, \cdots, N_{\mathrm{RF}}]\right)$ 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 (19) 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

$$\begin{aligned}\mathbf{\Sigma}_t &= \mathbf{U}_0\mathrm{diag}\left(\lambda_{1,1}^t, \lambda_{1,2}^t, \cdots, \lambda_{N_{\mathrm{T}},N_{\mathrm{R}}}^t\right)\mathbf{U}_0^H \\ &= \sum_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}}\sum_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}} \lambda_{n_{\mathrm{T}},n_{\mathrm{R}}}^t\left(\mathbf{a}_{n_{\mathrm{T}}} \otimes \mathbf{b}_{n_{\mathrm{R}}}\right)\left(\mathbf{a}_{n_{\mathrm{T}}}^H \otimes \mathbf{b}_{n_{\mathrm{R}}}^H\right), \end{aligned} \tag{22}$$

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

$$\lambda_{n_{\mathrm{T}},n_{\mathrm{R}}}^{t+1} = \begin{cases} \dfrac{\lambda_{n_{\mathrm{T}},n_{\mathrm{R}}}^t\sigma^2}{P\lambda_{n_{\mathrm{T}},n_{\mathrm{R}}}^t+\sigma^2}, & \mathbf{a}_{n_{\mathrm{T}}}^* \otimes \mathbf{b}_{n_{\mathrm{R}}} \text{ is selected}, \\ \lambda_{n_{\mathrm{T}},n_{\mathrm{R}}}^t, & \text{else}. \end{cases} \tag{23}$$

Based on the above derivation, we prove the following lemma to further simplify the original problem (19):

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 (19) can be transformed into an eigenvector selection problem, written as

$$\begin{aligned} \max_{n_{\mathrm{T}},\{n_{\mathrm{R},k}\}_{k=1}^{N_{\mathrm{RF}}}} \quad & \sum_{k=1}^{N_{\mathrm{RF}}} \log_2\left(1 + \frac{P\lambda_{n_{\mathrm{T}},n_{\mathrm{R},k}}^t}{\sigma^2}\right) \\ \text{s.t.} \quad & n_{\mathrm{T}} \in \{1, \cdots, N_{\mathrm{T}}\}, \\ & n_{\mathrm{R},k} \in \{1, \cdots, N_{\mathrm{R}}\}, \forall k, \\ & n_{\mathrm{R},k} \neq n_{\mathrm{R},k'}, \forall k \neq k'. \end{aligned} \tag{24}$$

Proof: (See Appendix F). ∎

Problem (24) 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 (24). A linear search algorithm is proposed to solve (24) 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 desired precoder and combiner are thereby expressed as

$$\mathbf{v}_{t+1}^{\mathrm{opt}} = \sqrt{P}\mathbf{a}_{n_{\mathrm{T}}^{\mathrm{opt}}}^* \text{ and } \mathbf{W}_{t+1}^{\mathrm{opt}} = \left[\mathbf{b}_{n_{\mathrm{R},1}^{\mathrm{opt}}}, \cdots, \mathbf{b}_{n_{\mathrm{R},N_{\mathrm{RF}}}^{\mathrm{opt}}}\right], \tag{25}$$

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

respectively, which generate a feasible observation matrix $\mathbf{X}^{\rm opt}_{t+1}=\mathbf{v}^{\rm opt\,*}_{t+1}\otimes\mathbf{W}^{\rm opt}_{t+1}$ at timeslot $t+1$. One can verify without difficulty that the initialized precoder and combiner obtained in (16) are a special case of (25) when $t=0$.

To summarize, the eigenvalue updating rule in (23), as well as the eigenvector selection method stated in **Algorithm 2** and (25), 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 [36] is illustrated.

1) Ideal Water-Filling: To better understand the proposed observation matrix design, we first interpret problem (9) 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 (9) by $\sigma^2\mathbf{I}_{N_{\rm RF}Q}$ without affecting the optimal $I(\mathbf{y};\mathbf{h})$. Then, by further relaxing the constraints $\mathbf{W}^T_q\mathbf{W}_q=\mathbf{I}_{N_{\rm RF}}$ and $\|\mathbf{v}_q\|^2=P$, we focus only on the total power constraint imposed on the overall observation matrix $\mathbf{X}$, i.e., ${\rm Tr}\left(\mathbf{X}\mathbf{X}^H\right)=PN_{\rm RF}Q$. In this case, the optimal value of problem (9) is shown to have an upper bound:

$$\max_{\mathbf{X}}\ \log_2\det\left(\mathbf{I}_{N_{\rm RF}Q}+\frac{1}{\sigma^2}\mathbf{X}^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}\right)\quad {\rm s.t.}\ \ {\rm Tr}\left(\mathbf{X}\mathbf{X}^H\right)=PN_{\rm RF}Q. \tag{26}$$

Notably, this upper bound is equivalent to the channel capacity of a point-to-point MIMO system equipped with $N_{\rm T}N_{\rm R}$ transmit antennas and $N_{\rm RF}Q$ receive antennas. Thereafter, the overall observation matrix can be optimally solved as[3]

$$\mathbf{X}^{\rm ideal}=\mathbf{U}_0\left(:,[1,\cdots,N_{\rm RF}Q]\right)\mathbf{P}, \tag{27}$$

where $\mathbf{P}={\rm diag}\left(\sqrt{p_1},\cdots,\sqrt{p_{N_{\rm RF}Q}}\right)$ 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 ${\rm Tr}\left(\mathbf{X}^{\rm ideal}(\mathbf{X}^{\rm ideal})^H\right)=\sum^{N_{\rm RF}Q}_{n=1}p_n=PN_{\rm RF}Q$.

Although the ideal observation matrix $\mathbf{X}^{\rm ideal}$, that maximizes the upper bound (26), might not be implementable in practice (as $\mathbf{X}^{\rm ideal}$ may violate the constraints $\mathbf{W}^T_q\mathbf{W}_q=\mathbf{I}_{N_{\rm RF}}$ and $\|\mathbf{v}_q\|^2=P$), it can give us two pivotal intuitions. First, the observation matrix should align with the eigenspace $\{\mathbf{a}_{n_{\rm T}}\otimes\mathbf{b}_{n_{\rm R}}\}^{N_{\rm T},N_{\rm R}}_{n_{\rm T}=1,n_{\rm R}=1}$ of the full MIMO channel covariance, $\mathbf{\Sigma}_{\mathbf{h}}=\mathbf{\Sigma}_{\rm T}\otimes\mathbf{\Sigma}_{\rm R}$. Second, we need to fill in more power (water) to the eigenvectors having larger eigenvalues $\{\lambda_n(\mathbf{\Sigma}_{\mathbf{h}})\}^{N_{\rm RF}Q}_{n=1}$ (or equivalently lower base levels $\left\{\frac{\sigma^2}{\lambda_n(\mathbf{\Sigma}_{\mathbf{h}})}\right\}^{N_{\rm RF}Q}_{n=1}$).

2) 2DIF Versus Water-Filling: The proposed 2DIF algorithm materializes the above two intuitions under the practical constraints $\mathbf{W}^T_q\mathbf{W}_q=\mathbf{I}_{N_{\rm RF}}$ and $\|\mathbf{v}_q\|^2=P$ via the eigenvector selection process in (24). The first intuition is automatically achieved by assigning $\mathbf{v}_{t+1}$ with a eigenvector from $\{\sqrt{P}\mathbf{a}^*_{n_{\rm T}}\}^{N_{\rm T}}_{n_{\rm T}=1}$ and assigning the columns of $\mathbf{W}_{t+1}$ with different eigenvectors from $\{\mathbf{b}_{n_{\rm R}}\}^{N_{\rm R}}_{n_{\rm R}=1}$, as proved in **Lemma 4**. The second intuition is approximately accomplished via selecting eigenvectors that have lower base levels, $\left\{\frac{\sigma^2}{\lambda_n(\mathbf{\Sigma}_{\mathbf{h}})}\right\}$, by more times. This is attributed to the fact that the maximization of (24) tends to select an eigenvalue combination $\{\lambda^t_{n_{\rm T},n_{{\rm R},k}}\}^{N_{\rm RF}}_{k=1}$ that has the lowest $\left\{\frac{\sigma^2}{\lambda^t_{n_{\rm T},n_{{\rm R},k}}}\right\}^{N_{\rm RF}}_{k=1}$ on average. To see this approximation more clearly, we rewrite the updating rule of the selected eigenvalue in (23) as

$$\underbrace{\frac{\sigma^2}{\lambda^{t+1}_{n_{\rm T},n_{\rm R}}}}_{\text{Updated ice-level}}=\underbrace{\frac{\sigma^2}{\lambda^t_{n_{\rm T},n_{\rm R}}}}_{\text{Current ice-level}}+\underbrace{P}_{\text{Height of an ice block}}. \tag{28}$$

Equation (28) reveals that every time the eigenvector $\mathbf{a}_{n_{\rm T}}\otimes\mathbf{b}_{n_{\rm R}}$ is selected, the value of $\frac{\sigma^2}{\lambda^t_{n_{\rm T},n_{\rm R}}}$ increases by P. Similar to the water-filling process, the process described by (28) can be vividly interpreted as allocating an ice block having P-unit power to the $(n_{\rm T},n_{\rm R})$-th orthogonal channel, where $\frac{\sigma^2}{\lambda^t_{n_{\rm T},n_{\rm R}}}$ is viewed as the ice level in the t-th timeslot. To summarize, due to the consideration of $N_{\rm RF}$ RF chains and $N_{\rm T}\times N_{\rm R}$ MIMO systems, in each timeslot, the 2DIF first selects $N_{\rm RF}$ orthogonal channels, which have the deepest ice-levels $\left\{\frac{\sigma^2}{\lambda^t_{n_{\rm T},n_{{\rm R},k}}}\right\}^{N_{\rm RF}}_{k=1}$ on average, from the total $N_{\rm T}\times N_{\rm R}$ channels. Then, the 2DIF will fill $N_{\rm RF}$ ice blocks (i.e., $N_{\rm RF}$ 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.

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

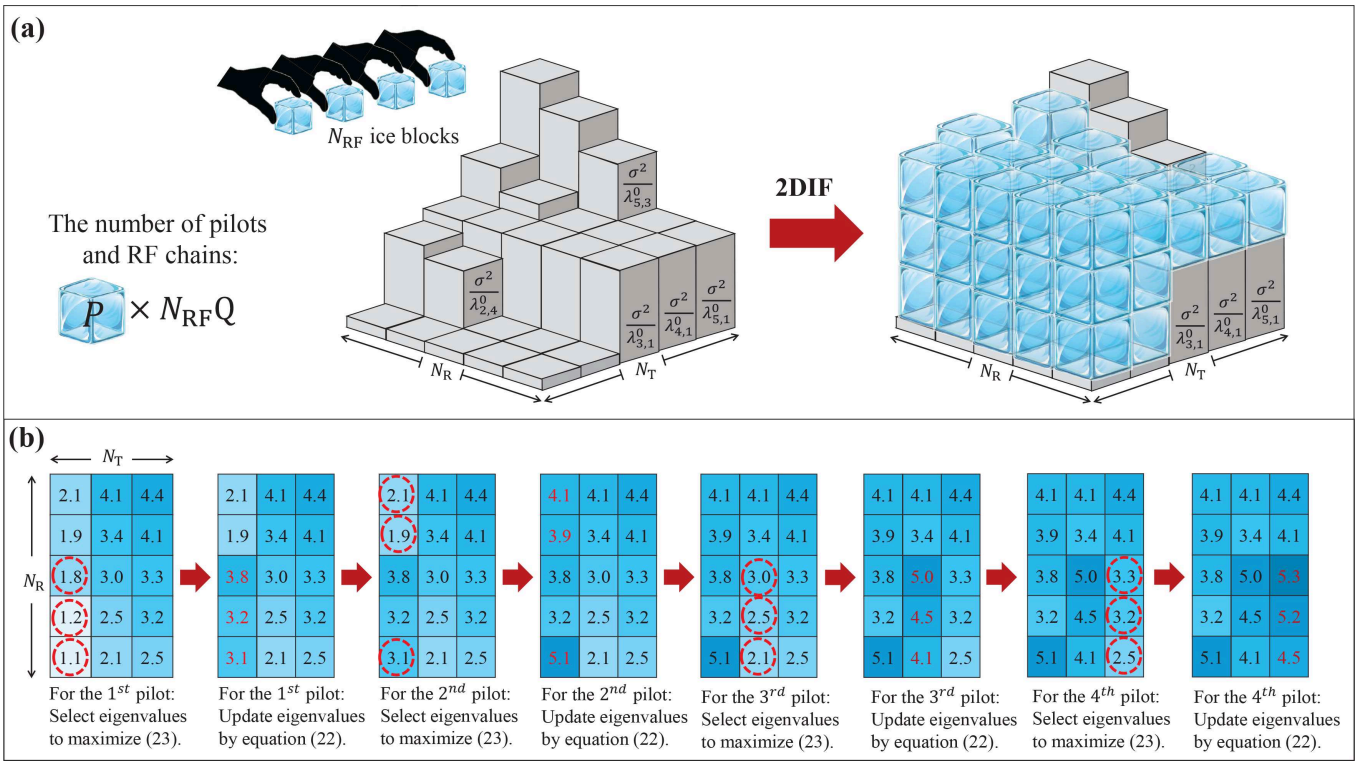

Fig. 2. (a) Shows how the pilot allocation of the proposed 2DIF algorithm works to maximize the MI. (b) Provides an example to show the implementation process of the proposed 2DIF, where $N_{\mathrm{R}} = 5$, $N_{\mathrm{T}} = 3$, $N_{\mathrm{RF}} = 3$, $P = 2$, and $Q = 4$. The number within the $(n_{\mathrm{T}},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

Example 2 (Example for Executing 2DIF): We present an example to show the growth of ice-level in Fig. 2(b). Here, we set $N_{\mathrm{R}} = 5$, $N_{\mathrm{T}} = 3$, $N_{\mathrm{RF}} = 3$, $P = 2$, and $Q = 4$. 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 (24), 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 [36]. 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 duplicate observations: $\mathbf{w}_t^H(\mathbf{H}\mathbf{v}_q + \mathbf{z}_q) = \mathbf{w}_{t+1}^H(\mathbf{H}\mathbf{v}_q + \mathbf{z}_q)$. These make no contribution to the channel estimation accuracy. In contrast, the proposed 2DIF algorithm efficiently circumvents this issue 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. 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 $\{\mathbf{b}_{n_{\mathrm{R},k}}\}_{k=1}^{N_{\mathrm{R}}}$. This fact guarantees the discrepancy of the observed channel information by each RF chain, contributing to improved channel estimation accuracy.

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:

$$\min_{\mathbf{v}_q, \mathbf{A}_q, \mathbf{D}_q} \quad \|\mathbf{X}_q^{\mathrm{IF}} - \mathbf{v}_q^* \otimes (\mathbf{A}_q\mathbf{D}_q)\|_F^2, \tag{29}$$

$$\text{s.t.} \quad \|\mathbf{v}_q\|^2 = P, \tag{29a}$$

$$|\mathbf{A}_q| = \frac{1}{\sqrt{N_{\mathrm{R}}}}\mathbf{1}_{N_{\mathrm{R}} \times N_{\mathrm{RF}}}, \tag{29b}$$

where $\mathbf{1}_{N_{\mathrm{R}} \times N_{\mathrm{RF}}}$ is an N_{R}-by-N_{RF} all-one matrix. By solving problem (29), 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 (29) owing to the non-convex modulus constraint in (29b) as well as the coupled relationship of $\mathbf{v}_q$, $\mathbf{A}_q$, and $\mathbf{D}_q$ in the objective (29). 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

$$\min_{\mathbf{D}_q} \sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{X}_{q,n}^{\mathrm{IF}} - v_q^*(n)\mathbf{A}_q\mathbf{D}_q\|_F^2. \tag{30}$$

Problem (30) is a standard quadradic programming (QP), which can be optimally solved by making the gradient of objective function to zero, i.e.,

$$\begin{aligned}\mathbf{D}_q^{\mathrm{opt}} &= \left(\sum_{n=1}^{N_{\mathrm{T}}} |v_q(n)|^2 \mathbf{A}_q^H\mathbf{A}_q\right)^{-1} \left(\sum_{n=1}^{N_{\mathrm{T}}} v_q(n)\mathbf{A}_q^H\mathbf{X}_{q,n}^{\mathrm{IF}}\right) \\ &\overset{(a)}{=} \sum_{n=1}^{N_{\mathrm{T}}} \frac{v_q(n)}{P}\left(\mathbf{A}_q^H\mathbf{A}_q\right)^{-1}\mathbf{A}_q^H\mathbf{X}_{q,n}^{\mathrm{IF}},\end{aligned} \tag{31}$$

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:

$$\min_{\mathbf{A}_q} \quad \sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{X}_{q,n}^{\mathrm{IF}} - \mathbf{A}_q\mathbf{D}_q v_q^*(n)\|_F^2, \tag{32}$$

$$\text{s.t.} \quad |\mathbf{A}_q| = \frac{1}{\sqrt{N_{\mathrm{R}}}}\mathbf{1}_{N_{\mathrm{R}} \times N_{\mathrm{RF}}}. \tag{32a}$$

Directly optimizing problem (32) is challenging owing to the constant modulus constraint (32a) 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,

$$\begin{aligned}&\sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{X}_{q,n}^{\mathrm{IF}} - \mathbf{A}_q\mathbf{D}_q v_q^*(n)\|_F^2 \\ &\le \sum_{n=1}^{N_{\mathrm{T}}} \|\mathbf{X}_{q,n}^{\mathrm{IF}}\mathbf{D}_q^{-1} - \mathbf{A}_q v_q^*(n)\|_F^2 \|\mathbf{D}_q\|_F^2.\end{aligned} \tag{33}$$

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 (31)

    Update the analog combiner $\mathbf{A}_q$ by (36)

    Update the precoder $\mathbf{v}_q$ by (38)

  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.

In (33), 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

$$\begin{aligned}&\sum_{n=1}^{N_\mathrm{T}} \|\mathbf{J}_{q,n} - \mathbf{A}_q v_q^*(n)\|_F^2 \\ &= C_1 + \sum_{n=1}^{N_\mathrm{T}} \left(\|\mathbf{A}_q\|_F^2 \||v_q(n)|^2 - 2\mathrm{Tr}\left\{\Re\left(v_q^*(n)\mathbf{J}_{q,n}^H\mathbf{A}_q\right)\right\}\right) \\ &= C_1 + N_{\mathrm{RF}}P - 2\mathrm{Tr}\left(\Re\left\{\mathbf{J}^H\mathbf{A}_q\right\}\right), \end{aligned} \tag{34}$$

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 (33) and (34), the new optimization problem is formulated as

$$\max_{|\mathbf{A}_q| = \frac{1}{\sqrt{N_\mathrm{R}}}\mathbf{1}_{N_\mathrm{R}\times N_{\mathrm{RF}}}} \mathrm{Tr}\left(\Re\left\{\mathbf{J}^H\mathbf{A}_q\right\}\right), \tag{35}$$

Evidently, the optimal solution of (35) is given by

$$\mathbf{A}_q^{\mathrm{opt}} = \frac{1}{\sqrt{N_\mathrm{R}}}\exp\left(j\angle\mathbf{J}\right), \tag{36}$$

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 (29) for optimizing $\mathbf{v}_q$ can be rewritten as

$$\begin{aligned}&\sum_{n=1}^{N_\mathrm{T}} \|\mathbf{X}_{q,n}^{\mathrm{IF}} - \mathbf{A}_q\mathbf{D}_q v_q^*(n)\|_F^2 \\ &= \|\mathbf{X}_q^{\mathrm{IF}}\|_F^2 + \|\mathbf{A}_q\mathbf{D}_q\|_F^2 - 2\Re\{\mathbf{c}_q^H\mathbf{v}_q\}, \end{aligned} \tag{37}$$

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 (29a), the optimal $\mathbf{v}_q^{\mathrm{opt}}$ is given as[4]

$$\mathbf{v}_q^{\mathrm{opt}} = \sqrt{P}\mathbf{c}_q/\|\mathbf{c}_q\|. \tag{38}$$

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 (31), (36), and (38) until convergence. After obtaining the hybrid combiners and precoders for pilot transmission, the channel estimator (7) 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.

[4]This work assumes the digital precoder $\mathbf{v}_q$ because it is technically feasible and commercially relevant for user equipment. In the analog precoding case, (38) should be modified as $\mathbf{v}_q^{\mathrm{opt}} = \sqrt{P/N_\mathrm{T}}\exp\left(j\angle\mathbf{c}_q\right)$. In the hybrid precoding case, as proved in [49], vector $\mathbf{v}_q^{\mathrm{opt}}$ in (38) can be perfectly realized by the hybrid precoder with at least 2 RF chains.

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 $\mathcal{O}\left(N_\mathrm{T}^3 + N_\mathrm{R}^3 + N_\mathrm{T}N_\mathrm{R}\log_2(N_\mathrm{R}) + (Q+N_\mathrm{T})N_{\mathrm{RF}}\right)$. 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. This advantage allows the proposed 2DIF to achieve lower computational complexity compared with the existing channel estimators [50], [51]. Besides, since the proposed algorithms only aim to design precoders and combiners, they can naturally share the same running process as the current 5G new radio (NR) systems.

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, three 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 [44], [45], [46], [47]. Concretely, according to the law of large numbers, $\mathbf{\Sigma}_\mathbf{h}$ can be trained by

$$\mathbf{\Sigma}_\mathbf{h} \approx \frac{1}{R}\sum_{r=1}^{R}\mathrm{vec}(\mathbf{H}_r)\mathrm{vec}(\mathbf{H}_r)^{\mathrm{H}}, \tag{39}$$

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 $\mathbf{\Sigma_h}$ [52]. For simplicity, we assume that the uniform linear arrays (ULAs) are deployed at both the BS and the user, while it can be easily extended to the UPA case. The consistency between the artificial kernels and the statistical kernel is that both of them assign higher similarity to nearby antennas and decrease influence with distance. Given array parameters, the mutual coupling matrices $\mathbf{C}_{\mathrm{rx}}$ and $\mathbf{C}_{\mathrm{tx}}$ can be calculated or measured. Thus here we focus on characterizing the spatial correlations $\mathbf{R}_{\mathrm{rx}}$ and $\mathbf{R}_{\mathrm{tx}}$, two artificial kernels are recommended [53]:

i) 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 spatial correlations at the user and the BS, can be modeled as

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

$$\mathbf{R}_{\mathrm{La,R}} = \exp\left(-\eta^2 \frac{d^2}{\lambda^2} \left|\mathbf{1}_{N_{\mathrm{R}}}^T \otimes \mathbf{n}_{\mathrm{R}} - \mathbf{n}_{\mathrm{R}}^T \otimes \mathbf{1}_{N_{\mathrm{R}}}\right|^{\odot 2}\right), \quad (41)$$

where $\eta > 0$ is an adjustable hyperparameter; and $\mathbf{Z}^{\odot 2}$ denotes the element-wise product of two matrices $\mathbf{Z}$. Thus, the overall kernel can be written as $\mathbf{\Sigma}_{\mathrm{La}} = ((\mathbf{C}_{\mathrm{tx}}^{1/2})^T \mathbf{R}_{\mathrm{La,T}} (\mathbf{C}_{\mathrm{tx}}^{1/2})^*) \otimes (\mathbf{C}_{\mathrm{rx}}^{1/2} \mathbf{\Sigma}_{\mathrm{La,R}} (\mathbf{C}_{\mathrm{rx}}^{1/2})^H)$.

ii) 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 spatial correlations at the user and the BS, can be modeled as

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

$$\mathbf{R}_{\mathrm{Be,R}} = J_0\left(\eta \frac{d}{\lambda} \left|\mathbf{1}_{N_{\mathrm{R}}}^T \otimes \mathbf{n}_{\mathrm{R}} - \mathbf{n}_{\mathrm{R}}^T \otimes \mathbf{1}_{N_{\mathrm{R}}}\right|\right), \quad (43)$$

where J_0 is the zero-order Bessel function of the first kind and $\eta > 0$ is a hyperparameter. The overall kernel is written as $\mathbf{\Sigma}_{\mathrm{Be}} = ((\mathbf{C}_{\mathrm{tx}}^{1/2})^T \mathbf{R}_{\mathrm{Be,T}} (\mathbf{C}_{\mathrm{tx}}^{1/2})^*) \otimes (\mathbf{C}_{\mathrm{rx}}^{1/2} \mathbf{\Sigma}_{\mathrm{Be,R}} (\mathbf{C}_{\mathrm{rx}}^{1/2})^H)$.

The hyperparameter η plays a pivotal role in closely mirroring the real channel covariance, whose value can be determined by a maximum likelihood (ML) estimator. We assume that R channel realizations are utilized to train a kernel $\mathbf{\Sigma}$, where $\mathbf{\Sigma}$ can be either $\mathbf{\Sigma}_{\mathrm{La}}$ or $\mathbf{\Sigma}_{\mathrm{Be}}$. Then, the estimation of η is written as

$$\eta^{\mathrm{opt}} = \arg\max_{\eta > 0} \sum_{r=1}^{R} \ln\left(\mathsf{P}\left(\mathbf{y}_r|\eta\right)\right), \quad (44)$$

wherein the likelihood function is given by

$$\mathsf{P}\left(\mathbf{y}_r|\eta\right) = \frac{\exp\left(-\mathbf{y}_r^H \left(\mathbf{X}_r^H \mathbf{\Sigma} \mathbf{X}_r + \mathbf{\Xi}_r\right)^{-1} \mathbf{y}_r\right)}{\pi^{QN_{\mathrm{RF}}} \det\left(\mathbf{X}_r^H \mathbf{\Sigma} \mathbf{X}_r + \mathbf{\Xi}_r\right)}, \quad (45)$$

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 $\mathbf{z}_r$. One-dimensional search method is adopted to obtain η^{opt}.

3) Adaptive Kernel: In cases where the statistical kernel is unavailable and the artificial kernels are inaccurate, we introduce a novel adaptive kernel training strategy. It uses channels estimated by the 2DIF algorithm to adaptively update the channel kernel for implementing 2DIF. The method smoothly integrates the kernel training and channel estimation stages, eliminating the additional time period required for estimating the covariance matrix.

Consider T_f consecutive frames, where the inter-frame channels $\mathbf{H}_{t_f} \in \mathbb{C}^{N_R \times N_T}, \forall t_f \in \{1, 2, \cdots, T_f\}$ are i.i.d distributed, following $\mathrm{vec}(\mathbf{H}_{t_f}) \sim \mathcal{CN}(0, \mathbf{\Sigma}_{\mathrm{T}} \otimes \mathbf{\Sigma}_{\mathrm{R}})$. Our strategy iteratively estimates the channels $\hat{\mathbf{H}}_{t_f}$ and updates the kernels $\hat{\mathbf{\Sigma}}_T^{(t_f)}$ and $\hat{\mathbf{\Sigma}}_R^{(t_f)}$, targeting at gradually improving the accuracy of $\hat{\mathbf{H}}_{t_f}$ and converging $\hat{\mathbf{\Sigma}}_T^{(t_f)} \otimes \hat{\mathbf{\Sigma}}_R^{(t_f)}$ to the real kernel $\mathbf{\Sigma}_{\mathrm{T}} \otimes \mathbf{\Sigma}_{\mathrm{R}}$. To be specific, at frame t_f, the channel $\mathbf{H}_{t_f}$ is estimated by 2DIF using the kernel $\hat{\mathbf{\Sigma}}_T^{(t_f-1)} \otimes \hat{\mathbf{\Sigma}}_R^{(t_f-1)}$ trained during frames $1 \sim t_f - 1$:

$$\hat{\mathbf{\Sigma}}_T^{(t_f-1)} \otimes \hat{\mathbf{\Sigma}}_R^{(t_f-1)} \overset{\text{2DIF}}{\rightarrow} \hat{\mathbf{H}}_{t_f}. \quad (47)$$

Then, we accumulate the information of the newly estimated channel $\hat{\mathbf{H}}_{t_f}$ to update the kernels $\hat{\mathbf{\Sigma}}_T^{(t_f)}$ and $\hat{\mathbf{\Sigma}}_R^{(t_f)}$ as

$$\begin{cases} \hat{\mathbf{\Sigma}}_T^{(t_f)} = \frac{t_f-1}{t_f} \hat{\mathbf{\Sigma}}_T^{(t_f-1)} + \frac{1}{t_f N_R} \sum_{n=1}^{N_R} \hat{\mathbf{H}}_{t_f}^T(n,:) \hat{\mathbf{H}}_{t_f}^*(n,:) \\ \hat{\mathbf{\Sigma}}_R^{(t_f)} = \frac{t_f-1}{t_f} \hat{\mathbf{\Sigma}}_R^{(t_f-1)} + \frac{1}{t_f N_T} \sum_{n=1}^{N_T} \hat{\mathbf{H}}_{t_f}(:,n) \hat{\mathbf{H}}_{t_f}^H(:,n). \end{cases} \quad (48)$$

As a result, the proposed adaptive kernel training strategy proceeds as in (46), Shown at the bottom of the page. To trigger the working flow in (46), the initial kernels $\hat{\mathbf{\Sigma}}_T^{(0)}$ and $\hat{\mathbf{\Sigma}}_R^{(0)}$ are set as identity matrices: $\hat{\mathbf{\Sigma}}_T^{(0)} = \mathbf{I}_{N_T}$ and $\hat{\mathbf{\Sigma}}_R^{(0)} = \mathbf{I}_{N_R}$.

The proposed adaptive kernel learning method exhibits two key advantages. Firstly, as the initial kernels are identity matrices, the proposed method requires no prior information about the real kernels, making it applicable to general and practical communication systems. Secondly, the proposed method allows

$$\hat{\mathbf{\Sigma}}_T^{(0)} \otimes \hat{\mathbf{\Sigma}}_R^{(0)} \underbrace{\overset{\text{2DIF}}{\rightarrow} \hat{\mathbf{H}}_1 \overset{(48)}{\rightarrow} \hat{\mathbf{\Sigma}}_T^{(1)} \otimes \hat{\mathbf{\Sigma}}_R^{(1)}}_{\text{frame 1}} \underbrace{\overset{\text{2DIF}}{\rightarrow} \hat{\mathbf{H}}_2 \overset{(48)}{\rightarrow} \hat{\mathbf{\Sigma}}_T^{(2)} \otimes \hat{\mathbf{\Sigma}}_R^{(2)}}_{\text{frame 2}} \overset{\text{2DIF}}{\rightarrow} \cdots \underbrace{\overset{\text{2DIF}}{\rightarrow} \hat{\mathbf{H}}_{T_f} \overset{(48)}{\rightarrow} \hat{\mathbf{\Sigma}}_T^{(T_f)} \otimes \hat{\mathbf{\Sigma}}_R^{(T_f)}}_{\text{frame } T_f}. \quad (46)$$

the coexistence of channel estimation and kernel training within one frame, as it leverages the channels estimated by 2DIF itself to train the kernel. This integration eliminates the additional time period required for kernel learning, greatly simplifying the frame structure and protocol of 2DIF in practical systems. Besides, it is worth noting that, the Kronecker structure in **Lemma 1** should be viewed as a design and approximation tool rather than a strict requirement. For some non-separable channel models such as [54], the proposed schemes can hold effectiveness by adopting the kernel selection methods provided in this subsection.

VI. Simulation Results

In this section, simulations are carried out to verify the effectiveness of the proposed 2DIF based channel estimation schemes.

A. Simulation Setup and Baselines

We consider a single-user MIMO system, where ULAs are equipped on both the BS and the user. The general correlated Rayleigh-fading channel model in (3) is considered to generate $\mathbf{H}$. In specific, the carrier frequency is set to 3.5 GHz. Following the setup in [41], the elements of the spatial correlation matrices $\mathbf{R}_{\rm rx}$ are calculated via (4), and the spatial correlation matrices $\mathbf{R}_{\rm tx}$ is obtained from the similar process. As a typical example to analyze antenna coupling effect [40], here we consider the dipole antennas with $\lambda/2$ length and $\lambda/100$ width for both transceivers. The antenna spacing is set to be $\lambda/8$. In this context, the spatial-scattering function $f_{\rm rx}(\varphi,\theta)$ for the BS and that for the user $f_{\rm tx}(\varphi,\theta)$ are given by $f_{\rm rx}(\varphi,\theta) = f_{\rm tx}(\varphi,\theta) = \frac{1.67}{2\pi}\cos^4(\theta)$ [43]. Given these array parameters, the mutual coupling matrices $\mathbf{C}_{\rm rx}$ and $\mathbf{C}_{\rm tx}$ are calculated using Matlab Antenna Toolbox [40]. Otherwise specifically specified, the numbers of transceiver antennas and RF chains are set as: $N_{\rm T}=4$, $N_{\rm R}=64$, and $N_{\rm RF}=4$, respectively. The signal-to-noise ratio (SNR) is defined as ${\rm SNR}=\frac{P}{\sigma^2}\mathsf{E}(\|\mathbf{h}\|^2)$, 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 ${\rm NMSE}=\mathsf{E}\left(\frac{\|\mathbf{h}-\hat{\mathbf{h}}\|^2}{\|\mathbf{h}\|^2}\right)$. 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 LS channel estimator is feasible only when observation dimension is no less than the channel dimension, i.e., $QN_{\rm RF}\geq N_{\rm T}N_{\rm R}$. To realize this, we assume the pilot length is $Q=\lceil N_{\rm T}N_{\rm R}/N_{\rm RF}\rceil=64$, and all combiners/precoders are generated by discrete Fourier transform (DFT) matrices.
- **MMSE**: Under the same setting of LS estimator, the classic MMSE estimator with DFT observation matrices is implemented to recover channel $\mathbf{H}$ via (7).
- **AMP**: Utilizing the channel sparsity in angular domain, the approximate message passing (AMP) method proposed in [28] is implemented to estimate channel $\mathbf{H}$. The combiners and precoders are randomly generated from Gaussian random measurement matrices.
- **IF scheme**: By viewing the considered MIMO system as $N_{\rm T}$ independent SIMO systems, the IF-based channel estimator [36] 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 $\lceil QN_{\rm RF}\rceil=192$ to ensure that it has the same number of observations with the 2DIF.
- **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 MMSE estimator in (7) 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 MMSE estimator in (7) 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 (26) via water-filling method. Then, (7) is employed to recover channel $\mathbf{H}$.

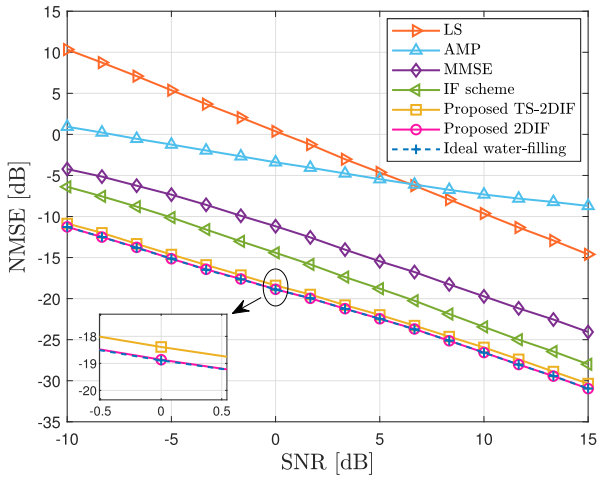

Fig. 3. The NMSE as a function of SNR for different schemes.

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

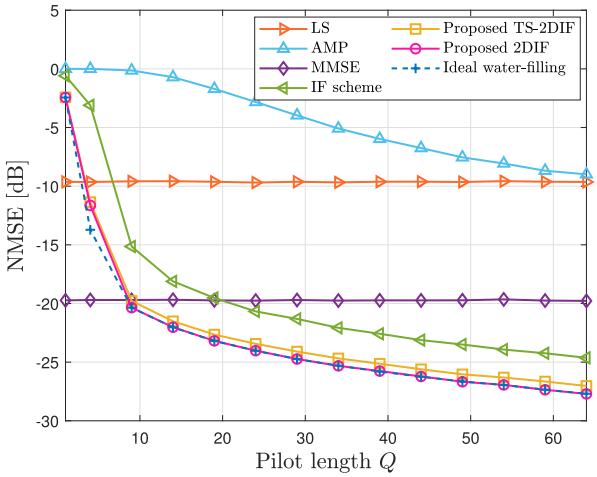

Fig. 4. The NMSE as a function of pilot length Q for different schemes.

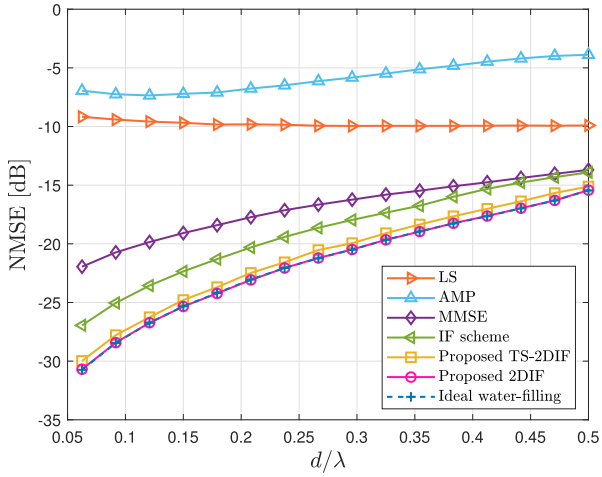

Fig. 5. The NMSE as a function of the normalized antenna spacing d/λ for different schemes.

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 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 Figs. 3 and 4, one can conclude that the TS-2DIF scheme can achieve almost the same estimation accuracy as the 2DIF. 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. As the antenna spacing increases, due to the reduced channel correlation, the proposed schemes gradually converge to the classical MMSE scheme. However, 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-fading (otherwise

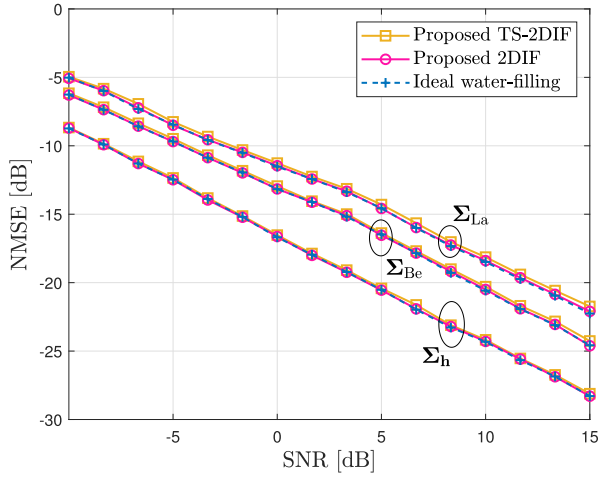

Fig. 6. The NMSE as a function of SNR for different kernels.

$\mathbf{\Sigma}_{\mathbf{h}} = \mathbf{I}_{N_{\mathrm{T}}N_{\mathrm{R}}}$), the non-diagonal kernel $\mathbf{\Sigma}_{\mathbf{h}}$ with some structural properties can still contribute to the improvement of the estimation accuracy.

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 $\mathrm{SNR} = 10$

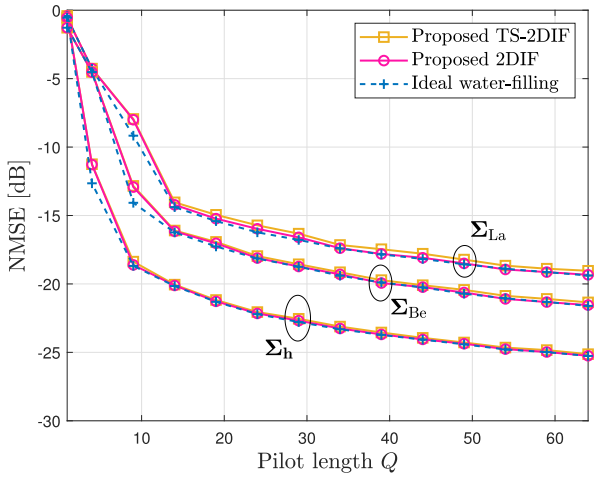

Fig. 7. The NMSE as a function of pilot length Q for different kernels.

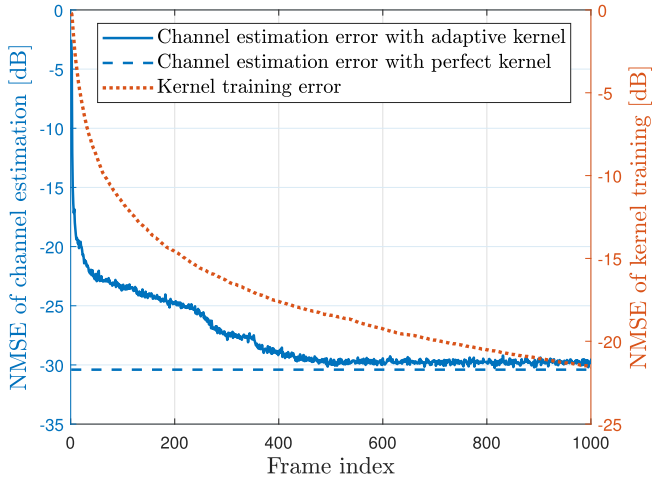

Fig. 8. The NMSE performance of the adaptive kernel training strategy.

dB, the NMSEs for Laplace kernel, Bessel kernel, and statistical kernel are about -18 dB, -20 dB, and -24 dB, respectively. When the pilot length is $Q = 19$, the NMSEs for these three kernels are about -15 dB, -17 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.

D. Estimation Accuracy Under Adaptive Kernel Training

We now evaluate the performance of the proposed adaptive kernel training strategy. Fig. 8 illustrates the NMSE of the estimated channel $\hat{\mathbf{H}}_{t_f}$ by 2DIF, accompanied by the NMSE of learned kernel $\hat{\mathbf{\Sigma}}_T^{(t_f)} \otimes \hat{\mathbf{\Sigma}}_R^{(t_f)}$ by adaptive training. The frame index t_f increases from 0 to 1000. The other settings are as follows: $N_T = 4$, $N_R = 64$, $d = \frac{\lambda}{8}$, $Q = 48$, $\mathrm{SNR} = 15$ dB. It is observed that, as the frame index increases, the proposed 2DIF algorithm and adaptive kernel training method achieve a mutual beneficial relationship. The newly estimated channels continuously improve the fidelity of learned kernels, while a more accurate kernel further decreases channel estimation error in future frames. Particularly, the NMSE of channel estimation rapidly declines below -20 dB after 16 frames of kernel training and closely approaches that achieved by the perfect kernel after 400 frames. Moreover, as the channel covariance matrix almost remains unchanged over time, we can keep using the learned kernel after adaptive training to achieve a near-optimal channel estimation performance (see frames 500$\sim$ 1000). This fact validates the feasibility and superiority of the proposed designs in practical situations where no prior information about the kernel is available.

VII. Conclusion

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 channel estimation schemes.

Appendix A
Proof of Lemma 1

Given the channel model in (3) and $\mathbf{h} \equiv \mathrm{vec}\,(\mathbf{H})$, the vectorized channel can be rewritten as

$$\begin{aligned}\mathbf{h} &= \mathrm{vec}\left(\mathbf{C}_{\mathrm{rx}}^{1/2}\mathbf{R}_{\mathrm{rx}}^{1/2}\mathbf{H}_{\mathrm{iid}}\mathbf{R}_{\mathrm{tx}}^{1/2}\mathbf{C}_{\mathrm{tx}}^{1/2}\right)\\ &= \left(\left(\mathbf{R}_{\mathrm{tx}}^{1/2}\mathbf{C}_{\mathrm{tx}}^{1/2}\right)^{T} \otimes \left(\mathbf{C}_{\mathrm{rx}}^{1/2}\mathbf{R}_{\mathrm{rx}}^{1/2}\right)\right)\mathrm{vec}\,(\mathbf{H}_{\mathrm{iid}}),\end{aligned} \tag{49}$$

where the second equation holds since $\mathrm{vec}\,(\mathbf{ABC}) = \left(\mathbf{C}^T \otimes \mathbf{A}\right)\mathrm{vec}\,(\mathbf{B})$. Then, the covariance of channel $\mathbf{h}$ can be derived as (52) at the bottom of the next page, where (a) holds since $\mathsf{E}\left(\mathrm{vec}\,(\mathbf{H}_{\mathrm{iid}})\left(\mathrm{vec}\,(\mathbf{H}_{\mathrm{iid}})\right)^H\right) = \mathbf{I}_{N_{\mathrm{R}}N_{\mathrm{T}}}$, (b) holds due to the commutative law of Kronecker product $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD})$; and (c) holds by defining

$$\mathbf{\Sigma}_{\mathrm{T}} = \left(\mathbf{C}_{\mathrm{tx}}^{1/2}\right)^{T}\mathbf{R}_{\mathrm{tx}}^{*}\left(\mathbf{C}_{\mathrm{tx}}^{1/2}\right)^{*}, \tag{50}$$

$$\mathbf{\Sigma}_{\mathrm{R}} = \mathbf{C}_{\mathrm{rx}}^{1/2}\mathbf{R}_{\mathrm{rx}}\left(\mathbf{C}_{\mathrm{rx}}^{1/2}\right)^{H}. \tag{51}$$

One can find that, the matrix $\mathbf{\Sigma}_{\mathrm{T}}$ only depends on the spatial correlation matrix $\mathbf{R}_{\mathrm{tx}}$ and the mutual coupling matrix $\mathbf{C}_{\mathrm{tx}}$ at the user, while the matrix $\mathbf{\Sigma}_{\mathrm{R}}$ is only associated with the spatial correlation matrix $\mathbf{R}_{\mathrm{rx}}$ and the mutual coupling matrix $\mathbf{C}_{\mathrm{rx}}$ at the BS. 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 (53), shown at the bottom of the page, where (a) holds since $\det(\mathbf{I}+\mathbf{AB})=\det(\mathbf{I}+\mathbf{BA})$ and $\mathbf{\Xi}=\sigma^2\text{blkdiag}\left(\mathbf{W}_1^H\mathbf{W}_1,\cdots,\mathbf{W}_Q^H\mathbf{W}_Q\right)$; (b) holds according to the property that $(\mathbf{a}\otimes\mathbf{B})\,\mathbf{C}\,(\mathbf{a}^H\otimes\mathbf{D})=(\mathbf{a}\mathbf{a}^H)\otimes(\mathbf{BCD})$ 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 (53). 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

$$I(\mathbf{y};\mathbf{h})=\log_2\det\left(\mathbf{I}_{N_{\mathrm{R}}N_{\mathrm{T}}}+\frac{1}{\sigma^2}\sum_{q=1}^{Q}\left(\left(\mathbf{v}_q^*\mathbf{v}_q^T\right)\otimes\left(\mathbf{\Pi}_q\mathbf{\Pi}_q^H\right)\right)\mathbf{\Sigma}_{\mathbf{h}}\right). \tag{54}$$

Observing (54), one can find that the MI $I(\mathbf{y};\mathbf{h})$ in (9) 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

$$\begin{aligned}I(\bar{\mathbf{y}}_{t+1};\mathbf{h})&\overset{(a)}{=}\log_2\det\left(\mathbf{I}_{N_{\mathrm{RF}}Q}+\frac{1}{\sigma^2}\bar{\mathbf{X}}_{t+1}^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_{t+1}\right)\\&=\log_2\det\begin{bmatrix}\mathbf{I}_{N_{\mathrm{RF}}t}+\frac{1}{\sigma^2}\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t & \frac{1}{\sigma^2}\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}_{t+1}\\ \frac{1}{\sigma^2}\mathbf{X}_{t+1}^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t & \mathbf{I}_{N_{\mathrm{RF}}}+\frac{1}{\sigma^2}\mathbf{X}_{t+1}^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}_{t+1}\end{bmatrix}\\&\overset{(b)}{=}\log_2\det\begin{bmatrix}\mathbf{I}_{N_{\mathrm{RF}}t}+\frac{1}{\sigma^2}\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t & \frac{1}{\sigma^2}\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}_{t+1}\\ \mathbf{0}_{N_{\mathrm{RF}}\times N_{\mathrm{RF}}t} & \mathbf{I}_{N_{\mathrm{RF}}}+\frac{1}{\sigma^2}\mathbf{X}_{t+1}^H\mathbf{\Sigma}_t\mathbf{X}_{t+1}\end{bmatrix}\\&=I(\bar{\mathbf{y}}_t;\mathbf{h})+\log_2\det\left(\mathbf{I}_{N_{\mathrm{RF}}}+\frac{1}{\sigma^2}\mathbf{X}_{t+1}^H\mathbf{\Sigma}_t\mathbf{X}_{t+1}\right),\end{aligned} \tag{55}$$

where (a) holds since according to **Lemma 2** and (b) holds by performing matrix triangularization. In particular, $\mathbf{\Sigma}_t$ is given by $\mathbf{\Sigma}_t=\mathbf{\Sigma}_{\mathbf{h}}-\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t\left(\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t+\sigma^2\mathbf{I}_{N_{\mathrm{RF}}t}\right)^{-1}\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}$, 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 (18) as $\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t=\mathbf{\Sigma}_{\mathbf{h}}\left[\bar{\mathbf{X}}_{t-1},\mathbf{X}_t\right]$ and

$$\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t=\begin{bmatrix}\bar{\mathbf{X}}_{t-1}^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_{t-1} & \bar{\mathbf{X}}_{t-1}^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}_t\\ \mathbf{X}_t^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_{t-1} & \mathbf{X}_t^H\mathbf{\Sigma}_{\mathbf{h}}\mathbf{X}_t\end{bmatrix}. \tag{57}$$

Then, using the Schur's matrix inversion formula to expand the term $\left(\bar{\mathbf{X}}_t^H\mathbf{\Sigma}_{\mathbf{h}}\bar{\mathbf{X}}_t+\sigma^2\mathbf{I}_{N_{\mathrm{RF}}t}\right)^{-1}$ in (18), the following recursion formula of can be obtained:

$$\mathbf{\Sigma}_{t+1}=\mathbf{\Sigma}_t-\mathbf{\Sigma}_t\mathbf{X}_{t+1}\left(\mathbf{X}_{t+1}^H\mathbf{\Sigma}_t\mathbf{X}_{t+1}+\sigma^2\mathbf{I}_{N_{\mathrm{RF}}}\right)^{-1}\mathbf{X}_{t+1}^H\mathbf{\Sigma}_t, \tag{58}$$

$$\begin{aligned}\mathbf{\Sigma}_{\mathbf{h}}=\mathsf{E}\left(\mathbf{h}\mathbf{h}^H\right)&=\mathsf{E}\left(\left(\left(\mathbf{R}_{\mathrm{tx}}^{1/2}\mathbf{C}_{\mathrm{tx}}^{1/2}\right)^T\otimes\left(\mathbf{C}_{\mathrm{rx}}^{1/2}\mathbf{R}_{\mathrm{rx}}^{1/2}\right)\right)\mathrm{vec}\left(\mathbf{H}_{\mathrm{iid}}\right)\left(\mathrm{vec}\left(\mathbf{H}_{\mathrm{iid}}\right)\right)^H\left(\left(\mathbf{R}_{\mathrm{tx}}^{1/2}\mathbf{C}_{\mathrm{tx}}^{1/2}\right)^*\otimes\left(\mathbf{C}_{\mathrm{rx}}^{1/2}\mathbf{R}_{\mathrm{rx}}^{1/2}\right)^H\right)\right)\\&\overset{(a)}{=}\left(\left(\mathbf{R}_{\mathrm{tx}}^{1/2}\mathbf{C}_{\mathrm{tx}}^{1/2}\right)^T\otimes\left(\mathbf{C}_{\mathrm{rx}}^{1/2}\mathbf{R}_{\mathrm{rx}}^{1/2}\right)\right)\left(\left(\mathbf{R}_{\mathrm{tx}}^{1/2}\mathbf{C}_{\mathrm{tx}}^{1/2}\right)^*\otimes\left(\mathbf{C}_{\mathrm{rx}}^{1/2}\mathbf{R}_{\mathrm{rx}}^{1/2}\right)^H\right)\\&\overset{(b)}{=}\left(\left(\mathbf{C}_{\mathrm{tx}}^{1/2}\right)^T\mathbf{R}_{\mathrm{tx}}^*\left(\mathbf{C}_{\mathrm{tx}}^{1/2}\right)^*\right)\otimes\left(\mathbf{C}_{\mathrm{rx}}^{1/2}\mathbf{R}_{\mathrm{rx}}\left(\mathbf{C}_{\mathrm{rx}}^{1/2}\right)^H\right)\overset{(c)}{=}\mathbf{\Sigma}_{\mathrm{T}}\otimes\mathbf{\Sigma}_{\mathrm{R}}.\end{aligned} \tag{52}$$

$$\begin{aligned}I(\mathbf{y};\mathbf{h})&\overset{(a)}{=}\log_2\det\Bigg(\mathbf{I}_{N_{\mathrm{R}}N_{\mathrm{T}}}\\&\quad+\frac{1}{\sigma^2}\left[\mathbf{v}_1^*\otimes\mathbf{W}_1,\cdots,\mathbf{v}_Q^*\otimes\mathbf{W}_Q\right]\text{blkdiag}\left(\left(\mathbf{W}_1^H\mathbf{W}_1\right)^{-1},\cdots,\left(\mathbf{W}_Q^H\mathbf{W}_Q\right)^{-1}\right)\left[\mathbf{v}_1^*\otimes\mathbf{W}_1,\cdots,\mathbf{v}_Q^*\otimes\mathbf{W}_Q\right]^H\mathbf{\Sigma}_{\mathbf{h}}\Bigg)\\&\overset{(b)}{=}\log_2\det\left(\mathbf{I}_{N_{\mathrm{R}}N_{\mathrm{T}}}+\frac{1}{\sigma^2}\sum_{q=1}^{Q}\left(\left(\mathbf{v}_q^*\mathbf{v}_q^T\right)\otimes\left(\mathbf{W}_q\left(\mathbf{W}_q^H\mathbf{W}_q\right)^{-1}\mathbf{W}_q^H\right)\right)\mathbf{\Sigma}_{\mathbf{h}}\right).\end{aligned} \tag{53}$$

$$\begin{aligned}f\left(\mathbf{v}_{t+1},\mathbf{W}_{t+1}\right)&\overset{(a)}{=}\log_2\det\left(\mathbf{I}_{N_{\mathrm{RF}}}+\frac{1}{\sigma^2}\sum_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}}\sum_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}}\lambda_{t,n_{\mathrm{T}},n_{\mathrm{R}}}\left(\mathbf{v}_{t+1}^T\otimes\mathbf{W}_{t+1}^H\right)\left(\mathbf{a}_{n_{\mathrm{T}}}\otimes\mathbf{b}_{n_{\mathrm{R}}}\right)\left(\mathbf{a}_{n_{\mathrm{T}}}^H\otimes\mathbf{b}_{n_{\mathrm{R}}}^H\right)\left(\mathbf{v}_{t+1}^*\otimes\mathbf{W}_{t+1}\right)\right)\\&\overset{(b)}{=}\log_2\det\left(\mathbf{I}_{N_{\mathrm{RF}}}+\frac{1}{\sigma^2}\sum_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}}\sum_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}}\lambda_{t,n_{\mathrm{T}},n_{\mathrm{R}}}\left|\mathbf{a}_{n_{\mathrm{T}}}^H\mathbf{v}_{t+1}^*\right|^2\mathbf{W}_{t+1}^H\mathbf{b}_{n_{\mathrm{R}}}\mathbf{b}_{n_{\mathrm{R}}}^H\mathbf{W}_{t+1}\right).\end{aligned} \tag{56}$$

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

$$\mathbf{\Sigma}_{t+1}=\mathbf{U}_t\mathbf{\Lambda}_t\mathbf{U}_t^H-\mathbf{X}_{t+1}\mathrm{diag}\left(\frac{\lambda_1^2\left(\mathbf{\Sigma}_t\right)}{P\lambda_1\left(\mathbf{\Sigma}_t\right)+\sigma^2},\cdots,\frac{\lambda_{N_{\mathrm{RF}}}^2\left(\mathbf{\Sigma}_t\right)}{P\lambda_{N_{\mathrm{RF}}}\left(\mathbf{\Sigma}_t\right)+\sigma^2}\right)\mathbf{X}_{t+1}^H. \tag{59}$$

Given that $\mathbf{X}_{t+1}\mathrm{diag}\left(\frac{\lambda_1^2(\mathbf{\Sigma}_t)}{P\lambda_1(\mathbf{\Sigma}_t)+\sigma^2},\cdots,\frac{\lambda_{N_{\mathrm{RF}}}^2(\mathbf{\Sigma}_t)}{P\lambda_{N_{\mathrm{RF}}}(\mathbf{\Sigma}_t)+\sigma^2}\right)\mathbf{X}_{t+1}^H=\mathbf{U}_t\mathrm{diag}\left(\frac{P\lambda_1^2(\mathbf{\Sigma}_t)}{P\lambda_1(\mathbf{\Sigma}_t)+\sigma^2},\cdots,\frac{P\lambda_{N_{\mathrm{RF}}}^2(\mathbf{\Sigma}_t)}{P\lambda_{N_{\mathrm{RF}}}(\mathbf{\Sigma}_t)+\sigma^2},\underbrace{0,\cdots,0}_{N_{\mathrm{R}}N_{\mathrm{T}}-N_{\mathrm{RF}}}\right)\mathbf{U}_t^H$ and $\mathbf{\Sigma}_t=\mathbf{U}_t\mathbf{\Lambda}_t\mathbf{U}_t^H$, the equality in (20) can be derived from (59), 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 $\mathbf{\Sigma_h}$ can be decomposed as

$$\begin{aligned}\mathbf{\Sigma_h}&=\left(\mathbf{U}_{\mathrm{T}}\mathbf{\Lambda}_{\mathrm{T}}{\mathbf{U}_{\mathrm{T}}}^H\right)\otimes\left(\mathbf{U}_{\mathrm{T}}\mathbf{\Lambda}_{\mathrm{T}}{\mathbf{U}_{\mathrm{T}}}^H\right)\\&=\underbrace{\left(\mathbf{U}_{\mathrm{T}}\otimes\mathbf{U}_{\mathrm{R}}\right)}_{\mathbf{U}_0}\underbrace{\left(\mathbf{\Lambda}_{\mathrm{T}}\otimes\mathbf{\Lambda}_{\mathrm{R}}\right)}_{\text{Eigenvalue matrix}}\left(\mathbf{U}_{\mathrm{T}}^H\otimes\mathbf{U}_{\mathrm{R}}^H\right)\\&=\sum_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}}\sum_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}}\alpha_{n_{\mathrm{T}}}\beta_{n_{\mathrm{R}}}\left(\mathbf{a}_{n_{\mathrm{T}}}\otimes\mathbf{b}_{n_{\mathrm{R}}}\right)\left(\mathbf{a}_{n_{\mathrm{T}}}^H\otimes\mathbf{b}_{n_{\mathrm{R}}}^H\right),\end{aligned} \tag{60}$$

Based on (60), one can verify without difficulty that (21) is exactly the eigenvalue decomposition of $\mathbf{\Sigma_h}$, 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 (19) can be reorganized as

$$\begin{aligned}\max_{\mathbf{v}_{t+1},\mathbf{W}_{t+1}}\ & f\left(\mathbf{v}_{t+1},\mathbf{W}_{t+1}\right)\\\text{s.t.}\ & \mathbf{v}_{t+1}\in\left\{\sqrt{P}\mathbf{a}_{n_{\mathrm{T}}}^*\right\}_{n_{\mathrm{T}}=1}^{N_{\mathrm{T}}},\\&\mathbf{w}_{t+1,k}\in\{\mathbf{b}_{n_{\mathrm{R}}}\}_{n_{\mathrm{R}}=1}^{N_{\mathrm{R}}},\forall k\in\{1,\cdots,N_{\mathrm{RF}}\},\\&\mathbf{w}_{t+1,k}\neq\mathbf{w}_{t+1,k'},\forall k\neq k',\end{aligned} \tag{61}$$

where the objective function is given in (56), Shown at bottom of the pervious page, in which (a) holds according to the definition in (22) 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 (61) ensures the orthogonality of $\mathbf{W}_{t+1}$. Observing (61), 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 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

$$\mathbf{v}_{t+1}^{\mathrm{opt}}=\sqrt{P}\mathbf{a}_{n_{\mathrm{T}}^{\mathrm{opt}}}^*\text{ and }\mathbf{W}_{t+1}^{\mathrm{opt}}=\left[\mathbf{b}_{n_{\mathrm{R},1}^{\mathrm{opt}}},\cdots,\mathbf{b}_{n_{\mathrm{R},N_{\mathrm{RF}}}^{\mathrm{opt}}}\right], \tag{62}$$

respectively. Then, we have

$$\mathbf{a}_{n_{\mathrm{T}}}^H(\mathbf{v}_{t+1}^{\mathrm{opt}})^*=\begin{cases}\sqrt{P}, & n_{\mathrm{T}}=n_{\mathrm{T}}^{\mathrm{opt}}\\0, & \text{else},\end{cases} \tag{63a}$$

$$\mathbf{b}_{n_{\mathrm{R}}}^H\mathbf{W}_{t+1}^{\mathrm{opt}}=\begin{cases}\mathbf{e}_{n_R}^T, & n_{\mathrm{R}}\in\{n_{\mathrm{R},k}^{\mathrm{opt}}\}_{k=1}^{N_{\mathrm{RF}}}\\\mathbf{0}_{N_{\mathrm{RF}}}^T, & \text{else},\end{cases} \tag{63b}$$

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 (63) into (61), the optimal MI increment $f\left(\mathbf{v}_{t+1}^{\mathrm{opt}},\mathbf{W}_{t+1}^{\mathrm{opt}}\right)$ can be expressed by

$$\begin{aligned}&f\left(\mathbf{v}_{t+1}^{\mathrm{opt}},\mathbf{W}_{t+1}^{\mathrm{opt}}\right)\\&=\log_2\det\left(\mathbf{I}_{N_{\mathrm{RF}}}+\frac{P}{\sigma^2}\sum_{n_{\mathrm{R}}}\lambda_{t,n_{\mathrm{T}}^{\mathrm{opt}},n_{\mathrm{R}}}(\mathbf{W}_{t+1}^{\mathrm{opt}})^H\mathbf{b}_{n_{\mathrm{R}}}\mathbf{b}_{n_{\mathrm{R}}}^H\mathbf{W}_{t+1}^{\mathrm{opt}}\right)\\&=\log_2\det\left(\mathbf{I}_{N_{\mathrm{RF}}}+\frac{P}{\sigma^2}\mathrm{diag}\left(\lambda_{t,n_{\mathrm{T}}^{\mathrm{opt}},n_{\mathrm{R},1}^{\mathrm{opt}}},\cdots,\lambda_{t,n_{\mathrm{T}}^{\mathrm{opt}},n_{\mathrm{R},N_{\mathrm{RF}}}^{\mathrm{opt}}}\right)\right)\\&=\sum_{k=1}^{N_{\mathrm{RF}}}\log_2\left(1+\frac{P\lambda_{t,n_{\mathrm{T}}^{\mathrm{opt}},n_{\mathrm{R},k}^{\mathrm{opt}}}}{\sigma^2}\right),\end{aligned} \tag{64}$$

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 (24). This completes the proof.

References

[1] Z. Zhang and M. Cui, "Two-dimensional ice filling based channel estimation in densifying MIMO systems," in *Proc. IEEE Int. Conf. Commun. (IEEE ICC'25)*, Jun. 2025, pp. 4750–4755.

[2] C. Huang et al., "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[3] R. Deng et al., "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.

[4] 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.

[5] I. Kanbaz, O. Yurduseven, and M. Matthaiou, "Super-directive antenna arrays: How many elements do we need?" in *Proc. IEEE Wireless Commun. Netw. Conf. (IEEE WCNC'24)*, 2024, pp. 1–6.

[6] 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.

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

[8] 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.

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

[10] 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.

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

[12] T. Gong et al., "Holographic MIMO communications: Theoretical foundations, enabling technologies, and future directions," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 1, pp. 196–257, 1st Quart. 2024.

[13] E. Basar et al., "Reconfigurable intelligent surfaces for 6G: Emerging hardware architectures, applications, and open challenges," *IEEE Veh. Technol. Mag.*, vol. 19, no. 3, pp. 27–47, Sep. 2024.

[14] L. Wei et al., "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.

[15] 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.

[16] 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.

[17] 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.

[18] M. Di Renzo, D. Dardari, and N. Decarli, "LoS MIMO-arrays vs. LoS MIMO-surfaces," in *Proc. 17th Eur. Conf. Antennas Propag. (EuCAP'23)*, 2023, pp. 1–5.

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

[20] 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.

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

[22] 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.

[23] T. Zheng, J. Zhu, Q. Yu, Y. Yan, and L. Dai, "Coded beam training," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 3, pp. 928–943, Mar. 2025.

[24] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Englewood Cliff s, NJ, USA: Prentice-Hall, Inc., 1993.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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. 10325–10329, Oct. 2019.

[30] 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.

[31] X. Ma, Z. Gao, F. Gao, and M. Di Renzo, "Model-driven deep learning based channel estimation and feedback for millimeter-wave massive hybrid MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2388–2406, Aug. 2021.

[32] A. Pizzo and A. Lozano, "Mutual coupling in holographic MIMO: Physical modeling and information-theoretic analysis," *IEEE J. Sel. Areas Inf. Theory*, vol. 6, pp. 111–126, May 2025.

[33] C. Williams and C. Rasmussen, "Gaussian processes for regression," in *Adv. Neural Inf. Process. Syst.*, vol. 8, 1995, pp. 514–520.

[34] 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.*, vol. 24, no. 1, pp. 309–322, Jan. 2025.

[35] 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.

[36] M. Cui, Z. Zhang, L. Dai, and K. Huang, "Ice-filling: Near-optimal channel estimation for dense array systems," *IEEE Trans. Wireless Commun.*, vol. 24, no. 10, pp. 8551–8564, Oct. 2025.

[37] Y. Wu, J. P. Linnartz, J. Bergmans, and S. Attallah, "Effects of antenna mutual coupling on the performance of MIMO systems," in *Proc. 29th Symp. Inf. Theory in Benelux. Citeseer*, 2008, pp. 1–8.

[38] K.-H. Chen and J.-F. Kiang, "Effect of mutual coupling on the channel capacity of MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 398–403, Jan. 2016.

[39] N. Kolomvakis and E. Björnson, "Exploiting mutual coupling characteristics for channel estimation in holographic MIMO," in *Proc. IEEE Global Commun. Conf. (IEEE GLOBECOM'24)*, 2024, pp. 3570–3575.

[40] MathWorks, "Antenna toolbox," Natick, MA, 2024. [Online]. Available: https://www.mathworks.com/products/antenna.html

[41] A. Pizzo, L. Sanguinetti, and T. L. Marzetta, "Fourier plane-wave series expansion for holographic MIMO communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 6890–6905, Sep. 2022.

[42] I. Gupta and A. Ksienski, "Effect of mutual coupling on the performance of adaptive arrays," *IEEE Trans. Antennas Propag.*, vol. 31, no. 5, pp. 785–791, Sep. 1983.

[43] C. A. Balanis, *Antenna Theory: analysis and Design.* Hoboken, NJ, USA: John Wiley & Sons, 2016.

[44] 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.

[45] 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.

[46] 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.

[47] 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. (IEEE ICC'20)*, Jun. 2020, pp. 1–7.

[48] B. Clarke and A. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.

[49] E. Zhang and C. Huang, "On achieving optimal rate of digital precoder by rf-baseband codesign for MIMO systems," in *Proc. IEEE 80th Veh. Technol. Conf. (IEEE VTC'14-Fall)* 2014, pp. 1–5.

[50] E. Vlachos, A. Kaushik, Y. C. Eldar, and G. C. Alexandropoulos, "Time-domain channel estimation for extremely large MIMO THz communication systems under dual-wideband fading conditions,". 2023, *arXiv:2310.14745*.

[51] E. Vlachos, G. C. Alexandropoulos, and J. Thompson, "Wideband MIMO channel estimation for hybrid beamforming millimeter wave systems via random spatial sampling," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 1136–1150, Sep. 2019.

[52] 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. (IEEE ICC'24)*, Jun. 2024, pp. 4869–4874.

[53] 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.

[54] X. Zhang, S. Song, and K. B. Letaief, "Fundamental limits of non-centered non-separable channels and their application in holographic MIMO communications," *IEEE Trans. Inf. Theory*, vol. 71, no. 9, pp. 6870–6894, Sep. 2025.