# Ice-Filling: Near-Optimal Channel Estimation for Dense Array Systems

Mingyao Cui, *Graduate Student Member, IEEE*, Zijian Zhang, *Graduate Student Member, IEEE*, Linglong Dai, *Fellow, IEEE*, and Kaibin Huang, *Fellow, IEEE*

***Abstract*—By deploying a large number of antennas with sub-half-wavelength spacing in a compact space, dense array systems (DASs) can fully unleash the multiplexing and diversity gains of limited apertures. To acquire these gains, accurate channel state information acquisition is necessary but challenging due to the large antenna numbers. To overcome this obstacle, this paper reveals that designing the observation matrix to exploit the high spatial correlation of DAS channels is crucial for realizing near-optimal Bayesian channel estimation. Specifically, we prove that the observation matrix design for channel estimation is equivalent to a time-domain duality of point-to-point multiple-input multiple-output precoding, except for the change in the total power constraint on the precoding matrix to the pilot-wise discrete power constraint on the observation matrix. Inspired by Bayesian regression, a novel ice-filling algorithm is proposed to design amplitude-and-phase controllable observation matrices, and a majorization-minimization algorithm is proposed to address the phase-only controllable case. Particularly, we prove that the ice-filling algorithm can be interpreted as a "quantized" water-filling algorithm, wherein the latter's continuous power-allocation process is converted into the former's discrete pilot-assignment process. To support the near-optimality of the proposed designs, we provide comprehensive analyses on the achievable mean square errors and their asymptotic expressions. Finally, numerical results confirm that our proposed designs achieve the near-optimal channel estimation performance and outperform existing approaches significantly.**

***Index Terms*—Estimation theory, mutual-information maximization, dense array systems (DAS), Bayesian regression.**

## I. Introduction

FROM 3G to 5G, antenna array systems play an irreplaceable role in wireless communications [1], [2]. By strategically configuring multiple antennas, an antenna array can constructively manipulate the radiated signals, allowing for enhanced wireless coverage, improved spectral efficiency, and reduced power consumption [3], [4], [5]. The performance of arrays continuously improves with the number of antennas [6], [7]. However, in practical scenarios, the space available for antenna deployment is usually limited, which restricts the performance gains endowed by antenna arrays [8]. To achieve a better performance with a spatially-limited aperture, dense array systems (DASs) have attracted extensive attentions in recent years [9].

Generally, DASs represent a series of array technologies that massive sub-wavelength antennas are densely arranged within a compact space. Unlike the conventional arrays with half-wavelength antenna spacing $\lambda/2$, the antenna spacing of DASs is much smaller, such as $\lambda/6$ [10], $\lambda/8$ [11], $\lambda/10$ [12], or even $\lambda/23$ [13]. With an increased number of sub-channels, DASs promise to achieve high array gains and fully exploit the multiplexing and diversity gains of limited apertures [9]. Besides, DASs can also reduce the grating lobes and provide high performance for large values of oblique angles of incidence [14]. The typical DAS realizations include holographic multiple-input multiple-output (MIMO) [15], reconfigurable intelligent surfaces (RISs) [16], fluid antenna systems (FASs) [17], [18], graphene-based nano-antenna arrays [19], [20], radio-frequency lens [21], and so on. For example, in [10], massive sub-wavelength patches of $\lambda/6$-spacing are densely printed on a holographic MIMO surface to generate multiple beams flexibly. In [22], meta-elements sized of $\lambda/2.5\times\lambda/17.5$ are seamlessly integrated onto a feeding microstrip to achieve reconfigurable holographic surface (RHS) beam steering. In [13], a RIS composed of antennas sized of $\lambda/23\times\lambda/23$ is designed and fabricated for terahertz communications.

The performance gains of antenna arrays are realized via the constructive beamformers enabled by their phase-shifters (PSs) and radio frequency (RF) chains. To realize effective beamforming, the acquisition of accurate channel state information (CSI) is essential, particularly in situations where the number of RF chains is smaller than the number of massive antennas. Up to now, many estimators have been proposed to acquire the CSI of large antenna arrays, and most of them can be adopted in DASs [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37]. Their fundamental principle involves receiving pilot signals using observation matrices and recovering the channel by advanced estimation algorithms. For example, when the available pilot length is larger than the antenna number, the classical least square (LS) estimator is

Received 2 October 2024; revised 13 January 2025; accepted 15 March 2025. Date of publication 14 May 2025; date of current version 14 October 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 624B2123. The associate editor coordinating the review of this article and approving it for publication was H. Tabassum. *(Corresponding author: Zijian Zhang.)*

Mingyao Cui and Kaibin Huang are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: mycui@eee.hku.hk; huangkb@eee.hku.hk).

Zijian Zhang is with the Department of Electronic Engineering and the State Key Laboratory of Space Network and Communications, Tsinghua University, Beijing 100084, China (e-mail: zhangzj20@mails.tsinghua.edu.cn).

Linglong Dai is with the Department of Electronic Engineering and the State Key laboratory of Space Network and Communications, Tsinghua University, Beijing 100084, China, and also with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: daill@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TWC.2025.3567797

usually adopted. Leveraging the property of channel sparsity, compressed sensing (CS)-based channel estimators are widely studied to improve the estimation accuracy and reduce the pilot overhead [23], [24], such as the orthogonal matching pursuit (OMP)-based estimator [25], [26], the message passing (MP)-based estimator [27], [28], [29], and the gridless sparse signal reconstructor [30]. By training neural networks with a large amounts of channel data, the deep learning approaches are also utilized to realize both data-driven and model-driven channel estimators [31], [32], [33], [37]. Besides, beam alignment techniques [34], [35], [36], including beam sweeping and hierarchical beam training, have also been widely explored to acquire the implicit CSI with low pilot overhead.

Although most of existing channel estimators can be adopted in DASs, they fail to fully exploit the high spatial correlation of DAS channels, thereby leaving a remarkable performance gap from the optimal estimator. Specifically, this spatial correlation is attributed to the fact that, the extremely-dense deployment of DAS antennas significantly increases the similarity of radio waves impinging on antenna ports and aggravates the mutual-coupling effect between adjacent antenna circuits. This fact makes the covariance matrices of DAS channels no longer diagonal but highly structured. For diagonal covariance matrices, the observation matrices for receiving pilot signals are either generated randomly or set as predefined codebooks, such as the Discrete Fourier Transform (DFT) matrix [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], which are sufficient to achieve the optimality of channel estimation [38], [39]. In contrast, since the correlation matrix of DAS channels are full of structural features, it is believed that their observation matrices can be tightly aligned with these features in pilot transmission. This structured pilot-sensing process has a high potential to can remarkably boost the channel estimation accuracy in DASs [40], [41], [42]. For example, in [18], the structured channel covariance is utilized to determine the positions of fluid antennas during the pilot reception. Due to the strong channel correlation of FAS channels, the estimator in [18] can achieve higher accuracy than CS methods. Authors in [43] further extend the method in [18] to sparse channel scenarios. In [44], aiming at efficient CSI acquisition in holographic MIMO systems, the electromagnetic channel covariance is modeled and utilized to enable a Bayesian channel estimator. However, these existing estimators often rely on specific array structures or parametric channel models, which can neither be used to design the general observation matrix in DASs nor guarantee the performance to be optimal.

To fill in this blank, our work represents the first attempt on designing and analyzing the near-optimal observation matrices in the DAS channel estimation. The classical Bayesian regression method is adopted as the channel estimator, i.e., the minimum-mean-square-error (MMSE) method, while our key contributions lie in the designed observation matrices to enhance Bayesian regression schemes, as summarized below.

- **Mutual-information-maximization (MIM) based framework**: From the perspective of MIM, we propose to design the DAS's observation matrix by minimizing the information uncertainty between the received pilots and wireless channels. The formulated MIM problem for observation matrix design is shown to be a time-domain duality of point-to-point MIMO precoding design. In particular, the multi-RF-chain resources for MIMO precoding is transformed to the multi-timeslot-pilot resources for channel estimation, and the total power constraint on the precoding matrix is switched to the pilot-wise discrete power constraint on the observation matrix. Motivated by this finding, we show that the ideal (but unachievable in practice) observation matrix could be constructed by the water-filling principle, which involves the use of the eigenvectors of channel covariance matrix as well as optimal power allocation.
- **Ice-filling enabled observation matrix design**: For practical DAS channel estimations, the continuous power-allocation by water-filling is not applicable as the pilot length is discrete and each pilot transmission has independent power constraint. To deal with this issue, we propose an ice-filling algorithm to design the amplitude-and-phase controllable observation matrices. Inspired by Bayesian regression, this algorithm sequentially generates the optimal blocks of the observation matrix by maximizing the mutual information (MI) increment between two adjacent pilot transmissions. An important insight is that the ice-filling algorithm converts the *power-allocation-process* of the ideal water-filling algorithm into an *eigenvector-assignment-process*. We rigorously prove that, the continuous powers allocated by water-filling are quantized as the number of times that *each eigenvector of the channel covariance is assigned for pilot transmission* by ice-filling, with a quantization error smaller than one. The proven quantization nature thereby ensures the near-optimality of the proposed ice-filling algorithm.
- **Majorization minimization (MM) enabled observation matrix design**: We then apply our framework for DAS channel estimation to the situation when only the phases of the observation matrix's weights are controllable while their amplitudes are fixed. To address this challenge, we propose a MM algorithm for designing the observation matrix. Its novelty lies in replacing the primal non-convex MIM problem with a series of tractable approximate sub-problems having analytical solutions, which are solved in an alternating optimization manner. Numerical simulations demonstrate that its channel estimation performance is comparable to the ice-filling algorithm.
- **Performance analysis**: Comprehensive analyses on the achievable mean square errors (MSEs) are provided to validate the effectiveness of the proposed designs. We prove that, compared to the random observation matrix design, the ice-filling algorithm can significantly improve the estimation accuracy by the order of the ratio between the number of antennas and the rank of channel covariance. Furthermore, based on the derived quantization nature of ice-filling, the MSE gap between water-filling and ice-filling is proved to decay quadratically with the pilot length, demonstrating the near-optimality of the ice-filling algorithm. In addition, the close-form MSEs under imperfect channel covariance are also derived. The superiority of the proposed algorithms is further validated using extensive numerical results.

The remainder of the paper is organized as follows. The system model is presented in Section II. The proposed

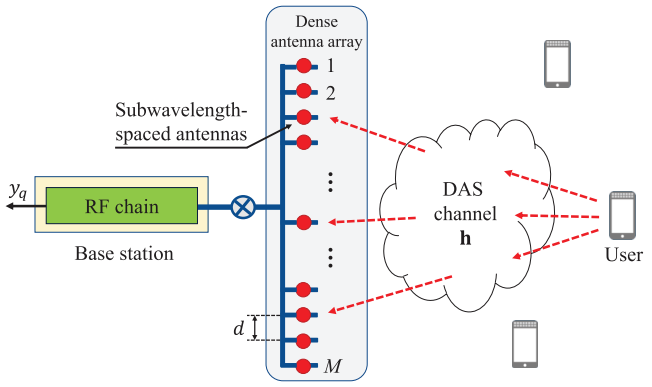


Fig. 1. An illustration of uplink channel estimation for a DAS.

MIM-based framework for observation matrix design is given in Section III. The ice-filling and MM algorithms are elaborated in Section IV. Then, the MSEs are analyzed in Section V. Simulation results are carried out in Section VI, and finally the conclusions are drawn in Section VII.

*Notation:* Lower-case and upper-case boldface letters represent vectors and matrices, respectively. $[\cdot]^{-1}$, $[\cdot]^{\dagger}$, $[\cdot]^{*}$, $[\cdot]^{\mathrm{T}}$, and $[\cdot]^{\mathrm{H}}$ denote the inverse, pseudo-inverse, conjugate, transpose, and conjugate-transpose operations, respectively; $\|\cdot\|_2$ denotes the $l_2$-norm of the argument; $\|\cdot\|_F$ denotes the Frobenius norm of the argument; $|\cdot|$ denotes the element-wise amplitude of its argument; $\mathsf{Tr}(\cdot)$ denotes the trace operator; $\mathsf{E}(\cdot)$ is the expectation operator; $\Re\{\cdot\}$ denotes the real part of the argument; $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue 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 $L$-dimensional all-one vector; and $\mathbf{0}_L$ is an all-zero vector or matrix with dimension $L$.

## II. System Model

As illustrated in Fig. 1, we consider the narrowband channel estimation of an uplink single-input multiple-output (SIMO) system. The base station (BS) employs an $M$-antenna DAS, which comprises an analog combining structure supported by one RF chain, to receive the pilots from a single-antenna user.

Let $\mathbf{h} \in \mathbb{C}^{M\times 1}$ be the channel vector and $Q$ the number of transmit pilots within a coherence-time frame. The received signal $y_q \in \mathbb{C}$ at the BS in timeslot $q$ is modeled as

$$y_q = \mathbf{w}_q^H \mathbf{h} s_q + \mathbf{w}_q^H \mathbf{z}_q, \tag{1}$$

where $\mathbf{w}_q \in \mathbb{C}^{M\times 1}$ is the observation vector at the BS, $s_q$ the pilot transmitted by the user, and $\mathbf{z}_q \sim \mathcal{CN}(\mathbf{0}_M, \sigma^2\mathbf{I}_M)$ the additive white Gaussian noise. As the observation vector $\mathbf{w}_q$ affects both the desired signal $\mathbf{h}s_q$ and the noise $\mathbf{z}_q$, its power has no impact on the estimation accuracy. Therefore, without any loss of generality, we can normalize $\mathbf{w}_q$ to $\|\mathbf{w}_q\|_2^2 = 1$, which reshapes the noise distribution to $\mathbf{w}_q^H\mathbf{z}_q \sim \mathcal{CN}(0, \sigma^2)$. Two practical implementations of the observation vector $\mathbf{w}_q$ will be discussed in our research: the amplitude-and-phase controllable combiner and the phase-only controllable combiner. The former connects each antenna to the RF chain using one PS and one low noise amplifier (LNA), enabling adjustment of both the amplitude and phase of the coefficients of $\mathbf{w}_q$. The latter deploys a single PS between each antenna and the RF chain. This implementations allows for only phase adjustment of the coefficients of $\mathbf{w}_q$, which enforces a constant modulus constraint on the observation vector, i.e., $|w_{q,m}| = \frac{1}{\sqrt{M}}$ for $m \in \{1, 2, \cdots, M\}$ [23].

Considering the total $Q$-timeslot pilot transmission, the received signal vector $\mathbf{y} := [y_1, \cdots, y_Q]^{\mathrm{T}}$ is expressed as

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

where $s_q$ is assumed to be 1 for all $q \in \{1, \cdots, Q\}$, $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_Q]$ stands for the observation matrix, and $\mathbf{z} := [\mathbf{w}_1^H\mathbf{z}_1, \cdots, \mathbf{w}_Q^H\mathbf{z}_Q]^{\mathrm{T}}$. As the power of $\mathbf{w}_q$ is normalized to 1, we have $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}_Q, \sigma^2\mathbf{I}_Q)$. The objective of channel estimation is to recover the channel $\mathbf{h}$ from the received pilot $\mathbf{y}$. As a DAS possesses highly correlated channels across different antenna ports, we adopt Bayesian regression to recover $\mathbf{h}$. Suppose the channel is generated from a Gaussian process $\mathcal{CN}(\mathbf{0}_M, \boldsymbol{\Sigma}_\mathbf{h})$, where the covariance matrix $\boldsymbol{\Sigma}_\mathbf{h} \in \mathbb{C}^{M\times M}$, also called *prior kernel*, characterizes our prior knowledge to the channel correlation. The kernel reflects the long-term statistical information of the covariance of channels, i.e., $\boldsymbol{\Sigma} = \mathsf{E}(\mathbf{h}\mathbf{h}^H)$, which can keep unchanged for a long time. To obtain the prior kernel in practical MIMO systems, an ideal approach is to leverage the channel realizations generated by some standard models or real-world datasets. In addition, some online second-moment learning methods proposed in [45] and [46] can also be used to obtain the prior kernel efficiently. Given the prior kernel, the joint probability distribution of $[\mathbf{h}^T, \mathbf{y}^T]^T$ satisfies

$$\mathcal{CN}\left(\begin{bmatrix}\mathbf{0}_M \\ \mathbf{0}_Q\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_\mathbf{h} & \boldsymbol{\Sigma}_\mathbf{h}\mathbf{W} \\ \mathbf{W}^H\boldsymbol{\Sigma}_\mathbf{h} & \mathbf{W}^H\boldsymbol{\Sigma}_\mathbf{h}\mathbf{W} + \sigma^2\mathbf{I}_Q\end{bmatrix}\right). \tag{3}$$

By invoking the property of the linear transform of Gaussian random vectors, the posterior mean and posterior covariance of $\mathbf{h}$ given observation $\mathbf{y}$ can be calculated as:

$$\boldsymbol{\mu}_{\mathbf{h}|\mathbf{y}} = \boldsymbol{\Sigma}_\mathbf{h}\mathbf{W}\left(\mathbf{W}^H\boldsymbol{\Sigma}_\mathbf{h}\mathbf{W} + \sigma^2\mathbf{I}_Q\right)^{-1}\mathbf{y}, \tag{4}$$

$$\boldsymbol{\Sigma}_{\mathbf{h}|\mathbf{y}} = \boldsymbol{\Sigma}_\mathbf{h} - \boldsymbol{\Sigma}_\mathbf{h}\mathbf{W}\left(\mathbf{W}^H\boldsymbol{\Sigma}_\mathbf{h}\mathbf{W} + \sigma^2\mathbf{I}_Q\right)^{-1}\mathbf{W}^H\boldsymbol{\Sigma}_\mathbf{h}. \tag{5}$$

The optimal Bayesian estimation to the channel vector is exactly the posterior mean, i.e., $\hat{\mathbf{h}} = \boldsymbol{\mu}_{\mathbf{h}|\mathbf{y}}$, which is also known as the MMSE estimator. Notably, the posterior covariance, $\boldsymbol{\Sigma}_{\mathbf{h}|\mathbf{y}}$, also called *posterior kernel*, characterizes the channel estimation error and is a function of the prior kernel $\boldsymbol{\Sigma}_\mathbf{h}$ and the observation matrix $\mathbf{W}$. Due to the extremely high spatial correlation exhibited by DAS channels, the prior kernel $\boldsymbol{\Sigma}_\mathbf{h}$ would remarkably deviate from the identity matrix. This deviation indicates that aligning the observation matrix $\mathbf{W}$ with the kernel's subspace could be greatly helpful in reducing the estimation error. Motivated by this fact, this paper concentrates on the design of observation matrix $\mathbf{W}$, by considering both the amplitude-and-phase controllable and phase-only controllable analog combiners, to boost the channel estimation performance.

## III. Mutual Information Maximization for Channel Estimation in DASs

In this section, we first present a MIM based framework for designing observation matrices. Then, a water-filling-inspired

observation-matrix-design is investigated as an ideal case of the proposed framework.

### A. MIM Based Framework for Observation Matrix Design

We adopt the MI between the received signal $\mathbf{y}$ and the channel $\mathbf{h}$ as the metric to evaluate the quality of the designed observation matrix. There are two reasons for using the MI as a metric. Firstly, MI tells the amount of information about the unknown channel $\mathbf{h}$ we can gain from the received signal $\mathbf{y}$. The second reason is that, according to the information-theoretic properties of Bayesian statistics from [47], the maximization of MI is asymptotically the minimization of Cramer Rao Bound (CRB).

Given the distributions $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}_M, \mathbf{\Sigma_h})$ and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}_Q, \sigma^2\mathbf{I}_Q)$, the MI is formulated as

$$\max_{\mathbf{W}\in\mathcal{W}} I(\mathbf{y};\mathbf{h}) = \log\det\left(\mathbf{I}_Q + \frac{1}{\sigma^2}\mathbf{W}^H\mathbf{\Sigma_h}\mathbf{W}\right), \quad (6)$$

where $I(\cdot;\cdot)$ denotes the MI, $\det(\cdot)$ is the determinant of its argument, and $\mathcal{W}$ represents the feasible set of $\mathbf{W}$ for both amplitude-and-phase controllable and phase-only controllable analog combiners. It is notable that the observation matrix design problem in (6) resembles the well-know point-to-point MIMO precoding problem. Comprehensive comparisons between these two problems are provided below.

- **Category of unknown data**: Both MIMO precoding and observation matrix design in (6) aim to reduce the uncertainty between the received data, $\mathbf{y}$, and the unknown data. The former regards the transmitted symbol vector, denoted by $\mathbf{x}$, as the unknown data. To facilitate MIMO precoding, *multiple RF chains* are employed to support an equal number of data streams. On the other hand, the unknown data to the latter is the channel, $\mathbf{h}$. Our considered uplink channel estimation requires *transmitting pilot signals over multiple timeslots* to recover $\mathbf{h}$. In comparison, one can discover that the multiple RF chains in MIMO precoding are converted to the multi-timeslot pilot resources in channel estimation. Thereby, the design of the observation matrix in (6) can be viewed as a time-domain duality of point-to-point MIMO precoding.
- **Constraint on the observation matrix**: The mathematical difference between MIMO precoding and observation matrix design for uplink channel estimation is attributed to the constraints of $\mathbf{W}$. For MIMO precoding, each column of $\mathbf{W}$ refers to a precoding vector associated with one dedicated RF chain. The optimization of precoding vectors, leveraging multiple RF chains, typically enforces the *total-power-constraint* on $\mathbf{W}$, i.e., $\|\mathbf{W}\|_F^2 = Q$. In terms of the observation matrix design, each column of $\mathbf{W}$ signifies an observation vector used for receiving one pilot signal. As discussed in Section II, since each observation vector $\mathbf{w}_q$ amplifies both the desired signal and the noise, it is subject to the *pilot-wise power constraint*, i.e., $\|\mathbf{w}_q\|_2^2 = 1$ or $|w_{q,m}| = \frac{1}{\sqrt{M}}$. The distinction in the observation matrix constraint differentiates the MIM in (6) from the classical MIMO precoding.

### B. Water-Filling Inspired Ideal Observation Matrix Design

Prior to addressing problem (6), we would like to consider an ideal (but practically unachievable) situation to establish an upper bound of (6). Specifically, we temporarily make the ideal assumption that the noise vector $\mathbf{z}$ in (2) is independent of the observation matrix $\mathbf{W}$ and its distribution is always $\mathcal{CN}(\mathbf{0}_Q, \sigma^2\mathbf{I}_Q)$ regardless of the power of $\mathbf{w}_q$. In this context, the pilot-wise power constraint $\|\mathbf{w}_q\|_2^2 = 1$ can be relaxed to the total power constraint $\|\mathbf{W}\|_F^2 = \sum_{q=1}^{Q}\|\mathbf{w}_q\|_2^2 = Q$, and problem (6) becomes identical to the point-to-point MIMO precoding problem, which can be solved by water-filling [48].

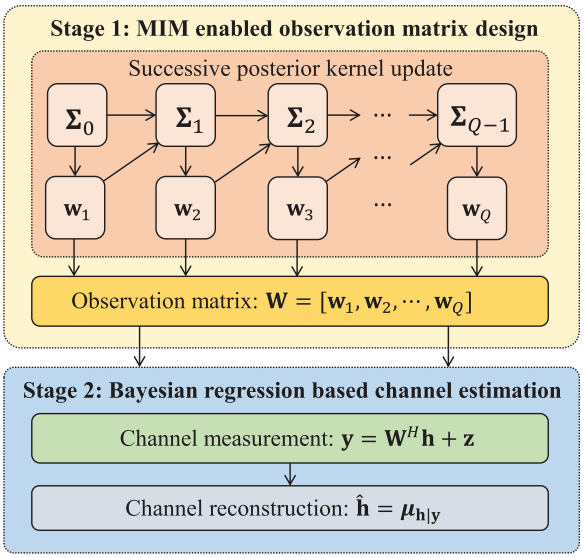


Fig. 2. Framework of the proposed observation matrix design.

Let the eigenvalue decomposition (EVD) of $\mathbf{\Sigma_h}$ be $\mathbf{U}_K\mathbf{\Lambda}_K\mathbf{U}_K^H$, where $K$ is the rank of $\mathbf{\Sigma_h}$, $\mathbf{U}_K = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_K]$, and $\mathbf{\Lambda}_K = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_K\}$ with $\lambda_1 \geq \cdots \geq \lambda_K > 0$. The ideal observation matrix $\mathbf{W}^{\mathrm{Ideal}}$ is given as $\mathbf{W}^{\mathrm{Ideal}} = \mathbf{U}_K\mathbf{P}$. The power allocation matrix $\mathbf{P} \in \mathbb{R}^{K\times Q}$ is expressed as $\mathbf{P} = [\mathbf{P}_K, \mathbf{0}_{K\times(Q-K)}]$ with $\mathbf{P}_K = \mathrm{diag}\{\sqrt{p_1}, \cdots, \sqrt{p_K}\}$. The power $p_k$ allocated to each eigenvector is calculated by the water-filling principle:

$$p_k = (\beta - \sigma^2/\lambda_k)^+, \forall k \in \{1, 2, \cdots, K\}, \quad (7)$$

where $(x)^+ \triangleq \max\{x, 0\}$, and the water-level $\beta$ is properly selected to meet the power constraint: $\sum_{k=1}^{K} p_k = Q$.

## IV. Practical Observation Matrix Design

This section elaborates on the observation matrix design in practical scenarios. Firstly, a greedy-method based principle is leveraged to solve the MIM problem in (6). Building on this principle, we propose an ice-filling algorithm for designing the amplitude-and-phase controllable observation matrices. Additionally, we conduct a comprehensive analysis on the relationship between ice-filling and water-filling. Subsequently, a MM-based algorithm is proposed for the design of phase-only controllable observation matrices. Last, the choice of prior kernel is also discussed.

### A. Observation Matrix Design Using Greedy Method

In practical uplink channel estimation where the noise vector $\mathbf{z}$ is amplified by the observation vector $\mathbf{w}_q$, the pilot-wise power constraint should be considered. In this case, it is

intractable to find the optimal solution to (6). To overcome this challenge, we harness a greedy method to obtain the observation vectors in a pilot-by-pilot manner, as shown in Fig. 2. Define $\mathbf{W}_t = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_t]$ as the observation matrix for timeslots $1 \sim t$, where $t < Q$, and denote $\mathbf{y}_t = \mathbf{W}_t^H \mathbf{h} + \mathbf{z}_t$ as the corresponding received signal. Given the current observation matrix $\mathbf{W}_t$, our greedy method aims to determine the next observation vector, $\mathbf{w}_{t+1}$, by maximizing the MI increment from timeslot $t$ to $t+1$, i.e., $\max_{\mathbf{w}_{t+1}} I(\mathbf{y}_{t+1}; \mathbf{h}) - I(\mathbf{y}_t; \mathbf{h})$.

As proved in [49], the MI increment is expressed as

$$I(\mathbf{y}_{t+1}; \mathbf{h}) - I(\mathbf{y}_t; \mathbf{h}) = \log_2 \left(1 + \frac{1}{\sigma^2} \mathbf{w}_{t+1}^H \mathbf{\Sigma}_t \mathbf{w}_{t+1}\right), \quad (8)$$

where $\mathbf{\Sigma}_t$ represents the posterior kernel of $\mathbf{h}$ given the observation $\mathbf{y}_t$. Using the block matrix inverse, we can explicitly write $\mathbf{\Sigma}_t$ as a function of $\mathbf{\Sigma}_{t-1}$ and $\mathbf{w}_t$ (Detailed proof can be found in [50, Appendix A]

$$\mathbf{\Sigma}_t = \mathbf{\Sigma}_{t-1} - \frac{\mathbf{\Sigma}_{t-1} \mathbf{w}_t \mathbf{w}_t^H \mathbf{\Sigma}_{t-1}}{\mathbf{w}_t^H \mathbf{\Sigma}_{t-1} \mathbf{w}_t + \sigma^2}. \quad (9)$$

Accordingly, the optimal $\mathbf{w}_{t+1}$ can be attained by solving the quadratic problem:

$$\mathbf{w}_{t+1} = \underset{\mathbf{w} \in \mathcal{W}}{\operatorname{argmax}}\, \mathbf{w}^H \mathbf{\Sigma}_t \mathbf{w}. \quad (10)$$

As a result, equations (8) to (10) inspire us to propose the framework in Fig. 2 for designing observation matrices. It begins by setting $\mathbf{\Sigma}_0$ as the prior kernel, $\mathbf{\Sigma}_\mathbf{h}$. Then, the first observation vector $\mathbf{w}_1$ is obtained by solving problem (10). In subsequent timeslots $t \in \{1, 2, \cdots, Q-1\}$, the posterior kernels $\mathbf{\Sigma}_t$ are updated from $\mathbf{\Sigma}_{t-1}$ and $\mathbf{w}_t$ using (9), while the observation vectors $\mathbf{w}_{t+1}$ are updated from $\mathbf{\Sigma}_t$ by solving problem (10). These two updating steps are repeated until all $Q$ observation vectors are generated for performing the ensuing Bayesian regression.

In the sequel, we will elaborate on solving problem (10) under the amplitude-and-phase controllable and phase-only controllable analog combiners, respectively.

### *B. Ice-Filling for Amplitude-and-Phase Controllable Observation Matrix*

For amplitude-and-phase controllable observation matrices, problem (10) is the well-known Rayleigh quotient problem:

$$\mathbf{w}_{t+1} = \underset{\|\mathbf{w}\|_2^2 = 1}{\operatorname{argmax}}\, \mathbf{w}^H \mathbf{\Sigma}_t \mathbf{w}. \quad (11)$$

The principal eigenvector of $\mathbf{\Sigma}_t$ yields the optimal solution to $\mathbf{w}_{t+1}$, i.e., $\mathbf{\Sigma}_t \mathbf{w}_{t+1} = \lambda_{\max}(\mathbf{\Sigma}_t) \mathbf{w}_{t+1}$. By leveraging the updating strategy of $\mathbf{\Sigma}_t$ described in (9) and the property of eigenvector, we can prove **Theorem 1**, which lays the foundation for our ice-filling algorithm.

*Theorem 1:* If $\mathbf{w}_t$ is the principal eigenvector of $\mathbf{\Sigma}_{t-1}$, then $\mathbf{\Sigma}_t$ can be rewritten as

$$\mathbf{\Sigma}_t = \mathbf{\Sigma}_{t-1} - \frac{\lambda_{\max}^2(\mathbf{\Sigma}_{t-1})}{\lambda_{\max}(\mathbf{\Sigma}_{t-1}) + \sigma^2} \mathbf{w}_t \mathbf{w}_t^H. \quad (12)$$

*Proof:* Applying the property of the principal eigenvector $\mathbf{\Sigma}_{t-1}\mathbf{w}_t = \lambda_{\max}(\mathbf{\Sigma}_{t-1})\mathbf{w}_t$, we get $\mathbf{w}_t^H \mathbf{\Sigma}_{t-1} \mathbf{w}_t = \lambda_{\max}(\mathbf{\Sigma}_{t-1})$ and $\mathbf{\Sigma}_{t-1}\mathbf{w}_t\mathbf{w}_t^H\mathbf{\Sigma}_{t-1} = \lambda_{\max}^2(\mathbf{\Sigma}_{t-1})\mathbf{w}_t\mathbf{w}_t^H$, which together with (9) give rise to (12). ■

*Remark 1: Given that $\mathbf{w}_t$ is the principal eigenvector of $\mathbf{\Sigma}_{t-1}$, **Theorem 1** implies that the posterior kernels, $\mathbf{\Sigma}_t$ and $\mathbf{\Sigma}_{t-1}$, share the identical eigenspace. This fact immediately leads to the result that the eigenvectors, $\mathbf{U}_K = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_K]$, of the prior kernel, $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_\mathbf{h}$, are inherited by all subsequent posterior kernels, $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \cdots$, and $\mathbf{\Sigma}_{Q-1}$. Thereby, we can draw the conclusion that all observation vectors, $\{\mathbf{w}_1\}_{t=1}^Q$, are selected from the eigenvectors of $\mathbf{\Sigma}_\mathbf{h}$.*

*Remark 2: Consider the eigenvalues of $\mathbf{\Sigma}_{t-1}$ and $\mathbf{\Sigma}_t$. We denote the EVD of $\mathbf{\Sigma}_{t-1}$ as $\mathbf{\Sigma}_{t-1} = \mathbf{U}_K \operatorname{diag}\{\lambda_1^{t-1}, \lambda_2^{t-1}, \cdots, \lambda_K^{t-1}\} \mathbf{U}_K^H$, where $\{\lambda_k^{t-1}\}_{k=1}^K$ refers to the eigenvalues of $\{\mathbf{u}_k\}_{k=1}^K$, respectively. We use $k_{t-1}$ to index the principal eigenvalue of $\mathbf{\Sigma}_{t-1}$. Accordingly, the equations $\mathbf{w}_t = \mathbf{u}_{k_{t-1}}$ and $\lambda_{\max}(\mathbf{\Sigma}_{t-1}) = \lambda_{k_{t-1}}^{t-1}$ hold, and **Theorem 1** can be rewritten as:*

$$\begin{aligned} \mathbf{\Sigma}_t &= \mathbf{U}_K \operatorname{diag}\{\lambda_1^{t-1}, \cdots, \lambda_{k_{t-1}}^{t-1}, \cdots, \lambda_K^{t-1}\} \mathbf{U}_K^H \\ &\quad - \frac{(\lambda_{k_{t-1}}^{t-1})^2}{\lambda_{k_{t-1}}^{t-1} + \sigma^2} \mathbf{u}_{k_{t-1}} \mathbf{u}_{k_{t-1}}^H \\ &= \mathbf{U}_K \operatorname{diag}\left\{\lambda_1^{t-1}, \cdots, \frac{\lambda_{k_{t-1}}^{t-1} \sigma^2}{\lambda_{k_{t-1}}^{t-1} + \sigma^2}, \cdots, \lambda_K^{t-1}\right\} \mathbf{U}_K^H. \end{aligned} \quad (13)$$

*Equation (13) reveals that only the principal eigenvalue, $\lambda_{k_{t-1}}^{t-1}$, of $\mathbf{\Sigma}_{t-1}$ is squeezed to an eigenvalue of $\mathbf{\Sigma}_t$ given by $\frac{\lambda_{k_{t-1}}^{t-1} \sigma^2}{\lambda_{k_{t-1}}^{t-1} + \sigma^2}$, while the remaining eigenvalues, $\lambda_k^{t-1}, \forall k \in \{1, 2, \cdots, K\} / \{k_{t-1}\}$, keep unchanged.*

**Algorithm 1** Ice-Filling-Based Observation Matrix Design

**Input:** Number of pilots $Q$, kernel $\mathbf{\Sigma}_\mathbf{h}$.
**Output:** Designed observation matrix $\mathbf{W}$.

1 Find the eigenvectors $[\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_K]$ and the corresponding eigenvalues $[\lambda_1, \lambda_2, \cdots, \lambda_K]$ of $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_\mathbf{h}$
2 Initialize $[\lambda_1^0, \lambda_2^0, \cdots, \lambda_K^0] = [\lambda_1, \lambda_2, \cdots, \lambda_K]$
3 **for** $t = 0, \cdots, Q-1$ **do**
4  $k_t = \arg\max_{k \in \{1,2,\cdots,K\}} \{\lambda_k^t\}$
5  Eigenvector-assignment: $\mathbf{w}_{t+1} = \mathbf{u}_{k_t}$
6  Eigenvalue-update: $\lambda_{k_t}^{t+1} = \frac{\lambda_{k_t}^t \sigma^2}{\lambda_{k_t}^t + \sigma^2}$
7  Eigenvalue-preserve: $\lambda_k^{t+1} = \lambda_k^t$ for $k \neq k_t$
8 **end for**
9 Construct observation matrix: $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_Q]$
10 **return:** Designed observation matrix $\mathbf{W}$

Motivated by **Remark 1** and **Remark 2**, our ice-filling algorithm is intrinsically an *eigenvector-assignment-process*, as presented in **Algorithm 1**. To elaborate, the eigenvalues $\{\lambda_1^t, \lambda_2^t, \cdots, \lambda_K^t\}$ of the posterior kernel $\mathbf{\Sigma}_t$ are initialized as the eigenvalues of the prior kernel $\mathbf{\Sigma}_\mathbf{h}$ at $t = 0$ in Step 2. In each timeslot, we find the largest eigenvalue from $\{\lambda_1^t, \lambda_2^t, \cdots, \lambda_K^t\}$ and index it by $k_t$ in Step 4. Then, the corresponding eigenvector $\mathbf{u}_{k_t}$ of $\mathbf{\Sigma}_h$, which is equivalent to the principal eigenvector of $\mathbf{\Sigma}_t$, is assigned to $\mathbf{w}_{t+1}$ in Step 5. Finally, according to **Remark 2**, the selected eigenvalue $\lambda_{k_t}^t$ is squeezed to $\frac{\lambda_{k_t}^t \sigma^2}{\lambda_{k_t}^t + \sigma^2}$, while the other eigenvalues of $\mathbf{\Sigma}_t$ are preserved, to attain the eigenvalues of the next kernel $\mathbf{\Sigma}_{t+1}$. These steps are repeatedly executed until all observation vectors are generated, which completes the ice-filling algorithm.

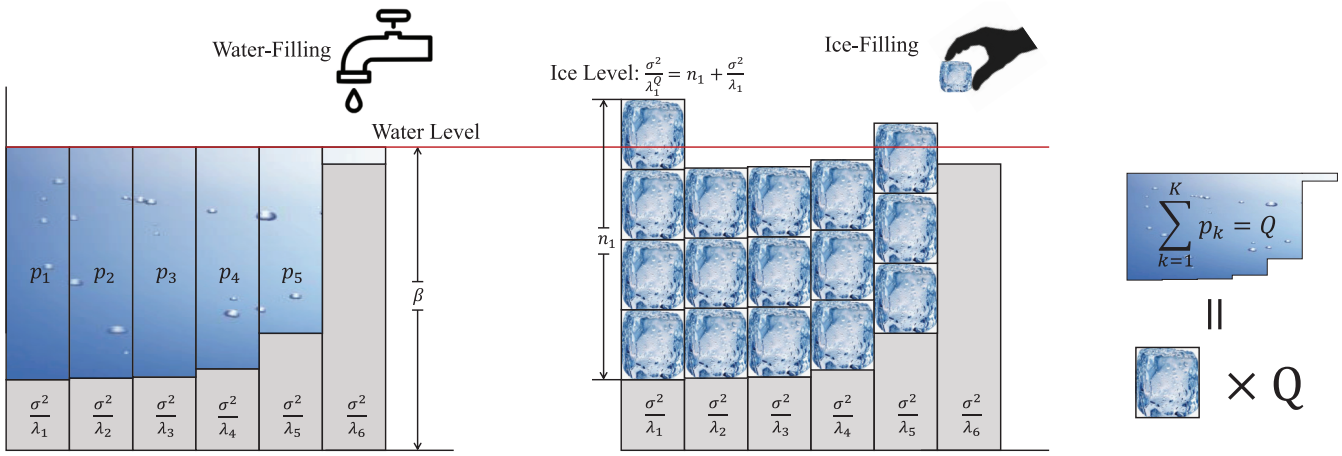


Fig. 3. Comparison between water-filling and ice-filling. The rank of the prior kernel, the number of antennas, and the total pilot length are set as $K = 6$, $M = 128$, and $Q = 16$, respectively. The eigenvalues of the prior kernel, $\mathbf{\Sigma_h}$, are in a descending order, i.e., $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5 > \lambda_6$.

### C. Ice-Filling Versus Water-Filling

In this subsection, we put forward a more insightful interpretation to the ice-filling algorithm that establishes its connection with the water-filling principle. Generally speaking, the proposed ice-filling algorithm can be viewed as a quantization of the water-filling principle. Specifically, it is clear that the observation matrices generated by both the water-filling and the ice-filling algorithms fall within the eigenspace of $\mathbf{\Sigma_h}$. The major distinction is attributed to their "power allocation strategies". The water-filling principle allocates $Q$ units of power to the $K$ eigenvectors using (7). The ice-filling algorithm transforms this power-allocation process into an *eigenvector-assignment process*, where the $K$ eigenvectors are repeatably assigned to the observation vectors in $Q$ timeslots.

To be more precise, recall that $p_k$ in (7) denotes the optimal power allocated to the $k$-th eigenvector, satisfying $\sum_{k=1}^{K} p_k = Q$. For a clear comparison, we introduce the definition of "pilot reuse frequency" as follows.

*Definition 1:* The pilot reuse frequency $n_k^t \in \mathbb{Z}^+$, with $k \in \{1, 2, \cdots, K\}$ and $t \in \{1, 2, \cdots, Q\}$, is defined as the number of times that the $k$-th eigenvector, $\mathbf{u}_k$, of the prior kernel is selected as the observation vector by the ice-filling algorithm during timeslots $1 \sim t$. The pilot reuse frequencies satisfy $\sum_{k=1}^{K} n_k^t = t$. For ease of expression, we define $n_k := n_k^Q$ as the pilot reuse frequency during the total $Q$ timeslots.

Comparing the definitions of $p_k$ and $n_k$, we can intuitively interpret the $n_k$-timeslot reuse of the observation vector $\mathbf{u}_k$ for pilot transmission as allocating $n_k$ units of power to $\mathbf{u}_k$. More importantly, **Definition 1** enables us to derive the analytical expressions of the eigenvalues $\{\lambda_k^t\}_{k=1}^K$ in **Algorithm 1**. Note that the eigenvalue update rule of the ice-filling algorithm can be rewritten as a recursive formula:

$$\lambda_{k_t}^{t+1} = \frac{\lambda_{k_t}^t \sigma^2}{\lambda_{k_t}^t + \sigma^2} \Leftrightarrow \frac{\sigma^2}{\lambda_{k_t}^{t+1}} = 1 + \frac{\sigma^2}{\lambda_{k_t}^t}, \tag{14}$$

where $\lambda_{k_t}^t$ is the largest eigenvalue from $\{\lambda_1^t, \lambda_2^t, \cdots, \lambda_K^t\}$. This recursive formula indicates that whenever the $k$-th eigenvector is selected as an observation vector, the value of $\frac{\sigma^2}{\lambda_k^t}$ increases by 1. By further considering the definition of the pilot reuse frequency $n_k^t$, that the $k$-th eigenvector is selected by $n_k^t$ times during timeslots $1 \sim t$, we can naturally obtain the relationship between $n_k^t$ and $\frac{\sigma^2}{\lambda_k^t}$:

$$\frac{\sigma^2}{\lambda_k^t} = n_k^t + \frac{\sigma^2}{\lambda_k} \Leftrightarrow n_k^t = \frac{\sigma^2}{\lambda_k^t} - \frac{\sigma^2}{\lambda_k}, \ k \in \{1, \cdots, K\}. \tag{15}$$

The comparison between equations (15) and (7) allows us to interpret the ice-filling algorithm as a quantization of the water-filling algorithm, as elaborated below.

**Interpretation of ice-filling**: As depicted in Fig. 3, the water-filling principle allocates $Q$ units of water (power) to a vessel with $K$ channels, each having a unique base level $\frac{\sigma^2}{\lambda_k}$, $k \in \{1, 2, \cdots, K\}$. By controlling the *uniform* water level $\beta$, the optimal power $p_k$ is determined by the gap between the base level and the water level. In contrast, the ice-filling algorithm transforms the total $Q$ units of water into $Q$ ice blocks, each containing one unit of power and being used for one pilot transmission. Our algorithm starts from an empty vessel and fills one ice block onto the channel having the deepest base surface $\frac{\sigma^2}{\lambda_{k_0}}$. This operation is equivalent to finding the largest eigenvalue $\lambda_{k_0}$ from $\{\lambda_1, \cdots, \lambda_K\}$. Then, the eigenvector $\mathbf{u}_{k_0}$ is assigned to the first observation vector $\mathbf{w}_1$, and ice level of this channel increases from $\frac{\sigma^2}{\lambda_{k_0}}$ to $\frac{\sigma^2}{\lambda_{k_0}^1} = \frac{\sigma^2}{\lambda_{k_0}} + 1$. In the subsequent time slots, the remaining $Q-1$ ice blocks are filled onto the channels locating at the deepest base surfaces or ice levels, indexed by $k_t = \arg\min_k \left\{\frac{\sigma^2}{\lambda_k^t}\right\}_{k=1}^K$, one by one. The corresponding eigenvectors $\mathbf{u}_{k_t}$ are used for pilot transmission, and the ice levels increase according to (15). Consequently, the final pilot reuse frequencies $\{n_k\}_{k=1}^K$ are determined by the number of ice blocks on top of each channel, and the ice levels are given by $\left\{\frac{\sigma^2}{\lambda_k^Q}\right\}_{k=1}^K = \left\{n_k + \frac{\sigma^2}{\lambda_k}\right\}_{k=1}^K$.

The comparison between the ice-filling and water-filling in Fig. 3 indicates that the continuous powers $\{p_k\}_{k=1}^K$ are quantized into the discrete pilot reuse frequencies $\{n_k\}_{k=1}^K$. This quantization results in the non-uniform ice levels $\left\{\frac{\sigma^2}{\lambda_k^Q}\right\}_{k=1}^K$, as opposed to the uniform water level $\beta$. To see the quantization nature of ice-filling more clearly, we prove **Theorem 2** to upper bound the quantization error between $p_k$ and $n_k$.

*Theorem 2:* Considering the $k$-th pilot reuse frequency, $n_k$, and the $k$-th optimal power, $p_k$, we have

$$|n_k - p_k| < 1. \tag{16}$$

*Proof:* See [50, Appendix B].■

*Remark 3:* ***Theorem 2*** *rigorously proves that the deviation of $n_k$ from $p_k$ is less than 1, rendering the pilot reuse frequency a good approximation of the optimal power allocation. Particularly, when the total pilot length $Q$ is large, the relative error between $n_k$ and $p_k$ tends to zero because*

$$\frac{|n_k - p_k|}{p_k} < \frac{1}{p_k} \overset{Q\to+\infty}{=} 0. \tag{17}$$

*Thereafter, the proposed ice-filling algorithm features a discrete approximation to the ideal water-filling algorithm with the relative quantization error decaying rapidly.*

**Algorithm 2** MM-Based Observation Matrix Design

**Input:** Number of pilots $Q$, kernel $\mathbf{\Sigma_h}$.
**Output:** Designed observation matrix $\mathbf{W}$.

```
Initialization: W = ∅, Σ_0 = Σ_h.
for t = 0, ··· , Q − 1 do
    Initialize w_{t+1} randomly
    Update: X = Tr(λ_max(Σ_t) I_M − Σ_t) I_M
    while no convergence of w_{t+1}^H Σ_t w_{t+1} do
        Update: v = w_{t+1}
        Update the t + 1-th observation vector: w_{t+1} = (1/√M) exp(j∠(X − λ_max(Σ_t) I_M + Σ_t) v)
    end while
    Update kernel: Σ_{t+1} = Σ_t − (Σ_t w_{t+1} w_{t+1}^H Σ_t) / (w_{t+1}^H Σ_t w_{t+1} + σ^2)
end for
Construct observation matrix: W = [w_1, w_2, ··· , w_Q]
return: Designed observation matrix W
```

### D. Majorization Minimization for Phase-Only Controllable Combiner

Turn now to the phase-only controllable analog combiners where the amplitude of each element of $\mathbf{W}$ is fixed as a constant. The observation vector at the $(t+1)$-th timeslot can be obtained by

$$\mathbf{w}_{t+1} = \underset{|\mathbf{w}|=\frac{1}{\sqrt{M}}\mathbf{1}_M}{\operatorname{argmax}} \ \mathbf{w}^H \mathbf{\Sigma}_t \mathbf{w}. \tag{18}$$

Due to the unit modulus constraint, the problem (18) is a non-convex programming. To cope with this problem, we propose a MM-based observation matrix design, as presented in **Algorithm 2**. Its key idea is to reformulate the non-convex problem as a series of more tractable approximate subproblems, which can be solved in an alternating manner [51]. Specifically, the proposed MM-based method is illustrated as follows.

To begin with, given that adding a constant to the objective function does not affect the optimal solution of (18), problem (18) is equivalent to

$$\begin{aligned}\mathbf{w}_{t+1} &\overset{(a)}{=} \underset{|\mathbf{w}|=\frac{1}{\sqrt{M}}\mathbf{1}_M}{\operatorname{argmax}} \ \mathbf{w}^H \mathbf{\Sigma}_t \mathbf{w} - \lambda_{\max}(\mathbf{\Sigma}_t)\, \mathbf{w}^H \mathbf{w} \\ &= \underset{|\mathbf{w}|=\frac{1}{\sqrt{M}}\mathbf{1}_M}{\operatorname{argmin}} \ \mathbf{w}^H \left(\lambda_{\max}(\mathbf{\Sigma}_t)\, \mathbf{I}_M - \mathbf{\Sigma}_t\right) \mathbf{w}, \end{aligned} \tag{19}$$

where (a) holds because $\mathbf{w}^H\mathbf{w} = 1$. Then, we consider to solve problem (19) by iteratively minimizing the upper-bound of its objective function. To this end, we introduce an auxiliary variable $\mathbf{v} \in \mathbb{C}^{M\times 1}$ to obtain an upper-bound function $\bar{f}(\mathbf{w}, \mathbf{v})$, given by [52]:

$$\begin{aligned}&\mathbf{w}^H \left(\lambda_{\max}(\mathbf{\Sigma}_t)\, \mathbf{I}_M - \mathbf{\Sigma}_t\right) \mathbf{w} \le \bar{f}(\mathbf{w}, \mathbf{v}) \\ &= \mathbf{w}^H \mathbf{X} \mathbf{w} - 2\Re\left\{\mathbf{w}^H \left(\mathbf{X} - \lambda_{\max}(\mathbf{\Sigma}_t)\, \mathbf{I}_M + \mathbf{\Sigma}_t\right) \mathbf{v}\right\} \\ &\quad + \mathbf{v}^H \left(\mathbf{X} - \lambda_{\max}(\mathbf{\Sigma}_t)\, \mathbf{I}_M + \mathbf{\Sigma}_t\right) \mathbf{v}, \end{aligned} \tag{20}$$

where $\mathbf{X} = \mathsf{Tr}\left(\lambda_{\max}(\mathbf{\Sigma}_t)\, \mathbf{I}_M - \mathbf{\Sigma}_t\right) \mathbf{I}_M$ and the equality holds when $\mathbf{v} = \mathbf{w}$. In this way, $\mathbf{w}$ and $\mathbf{v}$ in the upper-bound function $\bar{f}(\mathbf{w}, \mathbf{v})$ can be alternatively optimized to approach a sub-optimal solution to problem (19) [51]. In particular, for a given $\mathbf{w}$, $\mathbf{v}$ can be updated by $\mathbf{v} = \mathbf{w}$. For a given $\mathbf{v}$, by utilizing $\mathbf{w}^H\mathbf{X}\mathbf{w} = \mathsf{Tr}\left(\lambda_{\max}(\mathbf{\Sigma}_t)\, \mathbf{I}_M - \mathbf{\Sigma}_t\right)$ and removing the unrelated part in $\bar{f}(\mathbf{w}, \mathbf{v})$ in (20), the subproblem of updating $\mathbf{w}$ can be rewritten as

$$\begin{aligned}&\mathbf{w}_{t+1} \\ &= \underset{|\mathbf{w}|=\frac{1}{\sqrt{M}}\mathbf{1}_M}{\operatorname{argmax}} \ \Re\left\{\mathbf{w}^H \left(\mathbf{X} - \lambda_{\max}(\mathbf{\Sigma}_t)\, \mathbf{I}_M + \mathbf{\Sigma}_t\right) \mathbf{v}\right\}. \end{aligned} \tag{21}$$

It can be easily proved that the optimal solution to (21) is given by $\mathbf{w}_{t+1} = \frac{1}{\sqrt{M}} \exp\left(j\angle\left(\mathbf{X} - \lambda_{\max}(\mathbf{\Sigma}_t)\, \mathbf{I}_M + \mathbf{\Sigma}_t\right) \mathbf{v}\right)$. This solution, associated with the updating approach of the posterior kernel in (9), completes the algorithm. It is worth noting that, since the updates of $\mathbf{v}$ and $\mathbf{w}_{t+1}$ both monotonously increase the objective function, the convergence of **Algorithm 2** is naturally guaranteed.

### E. Kernel Selection in Non-Ideal Case

In some non-ideal cases where the perfect prior kernel, $\mathbf{\Sigma_h}$, is difficult to acquire, the initialization stages of **Algorithms 1 and 2** may pose a challenge. To deal with this issue, we suggest two methods to obtain imperfect prior kernels.

*1) Statistical Kernel:* Firstly, the prior kernel can be selected as the statistical covariance of historical estimated channels, i.e., $\hat{\mathbf{\Sigma}}_\mathbf{h} = \mathsf{E}(\hat{\mathbf{h}}\hat{\mathbf{h}}^H)$. Due to the error of channel estimation, the historical channels $\hat{\mathbf{h}}$ deviate from the true channels $\mathbf{h}$. We use a Gaussian noise to model this deviation, and get $\hat{\mathbf{h}} = \mathbf{h} + \mathbf{z_h}$, wherein $\mathbf{z_h} \sim \mathcal{CN}\left(\mathbf{0}_M, \sigma_\mathbf{h}^2 \mathbf{I}_M\right)$ denotes the historical channel estimation error. For simplicity, we name $\sigma_\mathbf{h}^2$ as the *kernel estimation error*. Thereby, the statistical kernel is modeled as

$$\hat{\mathbf{\Sigma}}_\mathbf{h} = \mathbf{\Sigma_h} + \sigma_\mathbf{h}^2 \mathbf{I}_M. \tag{22}$$

*2) Artificial Kernels:* Another choice is to employ artificially designed kernels to mimic the perfect kernel. To this end, the artificial kernels should assign high similarity to nearby antennas while diminishing the correlation rapidly with inter-antenna distance. For example, here we consider a general uniform planer array (UPA) equipped with $M$ antennas with an antenna spacing of $d$. The numbers of its horizontal antennas and vertical antennas are $M_x$ and $M_y$, respectively. Striking a balance between complexity and practicality, we recommend two artificial kernels as follows.

- **Exponential kernel**: The exponential kernel $\mathbf{\Sigma}_{\text{exp}}$, is the most popular selection in Bayesian estimation. Let $\mathbf{m}_x = \left[-\frac{M_x-1}{2}, -\frac{M_x-3}{2}, \cdots, \frac{M_x-1}{2}\right]^T$ and $\mathbf{m}_y = \left[-\frac{M_y-1}{2}, -\frac{M_y-3}{2}, \cdots, \frac{M_y-1}{2}\right]^T$. Then, the exponential kernels for the two dimensions of UPA, $\mathbf{\Sigma}_{\text{exp},x} \in \mathbb{C}^{M_x \times M_x}$ and $\mathbf{\Sigma}_{\text{exp},y} \in \mathbb{C}^{M_y \times M_y}$, can be respectively expressed as

$$\mathbf{\Sigma}_{\text{exp},x} = \exp\left(-\eta_1^2 \frac{4\pi^2 d^2}{\lambda^2} \left|\mathbf{1}_{M_x}^T \otimes \mathbf{m}_x - \mathbf{m}_x^T \otimes \mathbf{1}_{M_x}\right|^{\odot 2}\right), \tag{23}$$

$$\mathbf{\Sigma}_{\text{exp},y} = \exp\left(-\eta_1^2 \frac{4\pi^2 d^2}{\lambda^2} \left|\mathbf{1}_{M_y}^T \otimes \mathbf{m}_y - \mathbf{m}_y^T \otimes \mathbf{1}_{M_y}\right|^{\odot 2}\right), \tag{24}$$

where $\eta_1$ is an adjustable hyper-parameter and $\mathbf{X}^{\odot 2}$ denotes the element-wise product of two matrices $\mathbf{X}$. Then, the overall kernel can be written as

$$\mathbf{\Sigma}_{\text{exp}} = \mathbf{\Sigma}_{\text{exp},x} \otimes \mathbf{\Sigma}_{\text{exp},y}. \tag{25}$$

The exponential kernel, $\mathbf{\Sigma}_{\text{exp}}$, is a heuristic kernel. It exhibits the highest correlation at the diagonal elements, with the channel correlation diminishing exponentially as the antenna spacing increases. Compared to other kernels, the exponential kernel shows reduced sensitivity to outliers, making it suitable for reconstructing channels that lack obvious regularity.

- **Bessel kernel**: The Bessel kernel, denoted as $\mathbf{\Sigma}_{\text{bes}}$, is well-suited for capturing and modeling complex-valued data exhibiting oscillatory patterns. It characterizes the covariance of array steering vector when the user distributes uniformly in the spatial domain. For the considered UPA, the Bessel kernel can be obtained as $\mathbf{\Sigma}_{\text{bes}} = \mathsf{E}(\mathbf{a}(\theta,\phi)\mathbf{a}^H(\theta,\phi))$, where $\mathbf{a}(\theta,\phi)$ is the array steering vector of UPA, and $\theta \sim \mathcal{U}(-\pi/2, \pi/2)$ and $\phi \sim \mathcal{U}(-\pi/2, \pi/2)$ denote the angle-of-arrivals. By calculating expectation, the Bessel kernel is expressed as

$$\mathbf{\Sigma}_{\text{bes},x} = J_0\left(\eta_2 \frac{2\pi d}{\lambda} \left|\mathbf{1}_{M_x}^T \otimes \mathbf{m}_x - \mathbf{m}_x^T \otimes \mathbf{1}_{M_x}\right|\right), \tag{26}$$

$$\mathbf{\Sigma}_{\text{bes},y} = J_0\left(\eta_2 \frac{2\pi d}{\lambda} \left|\mathbf{1}_{M_y}^T \otimes \mathbf{m}_y - \mathbf{m}_y^T \otimes \mathbf{1}_{M_y}\right|\right), \tag{27}$$

where $J_0$ is the zero-order Bessel function of the first kind and $\eta_2$ is a hyper-parameter. Thus, the overall kernel can be written as

$$\mathbf{\Sigma}_{\text{bes}} = \mathbf{\Sigma}_{\text{bes},x} \otimes \mathbf{\Sigma}_{\text{bes},y}. \tag{28}$$

## V. Performance Analysis

To evaluate the performance of the proposed MIM-based observation matrix design, this section provides comprehensive MSE analyses under different algorithmic conditions.

### A. Estimation Accuracy Under Perfect Kernel

When the perfect prior kernel $\mathbf{\Sigma}_{\mathbf{h}}$ is available, the MSEs achieved by the proposed algorithms can be determined by the trace of their posterior kernels, i.e.,

$$\mathsf{E}(\|\boldsymbol{\mu}_{\mathbf{h}|\mathbf{y}} - \mathbf{h}\|_2^2) = \mathsf{Tr}\left(\mathbf{\Sigma}_{\mathbf{h}|\mathbf{y}}\right). \tag{29}$$

The subsequent two lemmas give out the close-form MSEs and their asymptotic expressions of the water-filling-based and ice-filling-based channel estimators, respectively.

*Lemma 1:* If the perfect kernel $\mathbf{\Sigma}_{\mathbf{h}}$ is adopted, the MSE $\delta_{\text{wf}}$ of the water-filling algorithm is given by

$$\delta_{\text{wf}} = \sum_{k=1}^{K} \frac{\lambda_k \sigma^2}{p_k \lambda_k + \sigma^2} \overset{Q\to+\infty}{=} \mathcal{O}(K^2 Q^{-1}). \tag{30}$$

*Proof:* See [50, Appendix C].■

*Lemma 2:* If the perfect kernel $\mathbf{\Sigma}_{\mathbf{h}}$ is adopted, the MSE $\delta_{\text{if}}$ of the ice-filling algorithm is given by

$$\delta_{\text{if}} = \sum_{k=1}^{K} \frac{\lambda_k \sigma^2}{n_k \lambda_k + \sigma^2} \overset{Q\to+\infty}{=} \mathcal{O}(K^2 Q^{-1}). \tag{31}$$

*Proof:* See [50, Appendix D].■

We can gain two insights from **Lemma 1** and **Lemma 2**. Firstly, the MSEs achieved by the water-filling and ice-filling methods have identical asymptotic expressions, both of which decay at the rate of $Q^{-1}$. The only difference is owing to the quantization of the continuous powers $\{p_k\}$ in (30) to the pilot reuse frequencies $\{n_k\}$ in (31). Next, the MSEs decay quadratically as the rank of the channel kernel, $K$, decreases. This finding demonstrates the superiority of DASs in channel estimation. A denser antenna deployment increases the correlation between inter-antenna channels, decreases the rank of the channel kernel, and thereby improves the channel estimation accuracy.

Taking into account the quantization nature revealed in **Theorem 2**, the asymptotic achievability bound of the MSE of the ice-filling algorithm can be evaluated by the following theorem.

*Theorem 3:* As $Q \to +\infty$, the asymptotic MSE difference, $|\delta_{\text{wf}} - \delta_{\text{if}}|$, is given by

$$|\delta_{\text{wf}} - \delta_{\text{if}}| = \mathcal{O}(K^3 Q^{-2}). \tag{32}$$

*Proof:* See [50, Appendix E].■

*Remark 4: **Theorem 3** guarantees that the MSE gap between the ice-filling and the ideal water-filling algorithms decays at a rate of $K^3Q^{-2}$, which demonstrates the near-optimality of the proposed ice-filling channel estimator when $Q \to +\infty$ or $K \to 0$.*

To validate the superiority of the proposed design, it is of great interest to evaluate the MSE achieved by the Bayesian regression (4) using randomly generated observation matrices. For amplitude-and-phase controllable cases, we assume that all elements of $\mathbf{W}$ are independently generated from a Gaussian distribution $\mathcal{CN}(0, 1/M)$. For phase-only controllable cases, the angles of all elements of $\mathbf{W}$ are randomly selected from $[-\pi, +\pi]$. Then, we derive the asymptotic MSE of a random observation matrix in **Lemma 3**.

*Lemma 3:* Assume that the perfect kernel $\mathbf{\Sigma}_{\mathbf{h}}$ is adopted and $Q$ is sufficiently large. When $\mathbf{W}$ is randomly generated, the MSE $\delta_{\text{rnd}}$ achieved in both phase-and-amplitude and phase-only controllable cases can be approximated by

$$\delta_{\text{rnd}} \approx \sum_{k=1}^{K} \frac{\lambda_k \sigma^2}{\frac{Q}{M} \lambda_k + \sigma^2} \overset{Q\to+\infty}{=} \mathcal{O}(KMQ^{-1}), \tag{33}$$

where the approximate equality can be infinitely close to an equality as $Q \to \infty$.

*Proof:* See [50, Appendix F].■

It is evident from **Lemmas 1–3** that meticulously designed observation matrices can scale down the asymptotic MSE from $\mathcal{O}(KMQ^{-1})$ to $\mathcal{O}(K^2Q^{-1})$. This improvement is attributed to the fact that the observation matrices designed by water-filling and ice-filling approaches are tightly aligned with the kernel's eigenspace of dimension $K$. In contrast, the random observation matrix wastes a large amount of power in the kernel's null space of dimension $M-K$, which contains no useful information about the channel. This finding demonstrates the significant role played by the observation matrix to enhance channel estimation accuracy, particularly in DASs where $M \gg K$.

### B. Estimation Accuracy Under Imperfect Kernel

We now consider the situation where the perfect prior kernel $\mathbf{\Sigma_h}$ is not available. In this context, an imperfect prior kernel, such as the statistical kernel in (22) and the artificial kernels (25) and (28), has to be adopted in the proposed algorithms. We denote the adopted prior kernel as $\mathbf{\Sigma}$. Then, one can use the following lemma to evaluate the MSE performance analytically.

*Lemma 4:* Given a prior kernel $\mathbf{\Sigma}$ as the input of the proposed channel estimator, the MSE $\hat{\delta}$ of channel estimation is given by

$$\hat{\delta} = \mathsf{Tr}\left(\left(\mathbf{\Pi}^H\mathbf{W}^H - \mathbf{I}_M\right)\mathbf{\Sigma_h}(\mathbf{W\Pi} - \mathbf{I}_M)\right) + \sigma^2\mathsf{Tr}\left(\mathbf{\Pi}^H\mathbf{\Pi}\right), \tag{34}$$

where $\mathbf{\Pi} := \left(\mathbf{W}^H\mathbf{\Sigma W} + \sigma^2\mathbf{I}_Q\right)^{-1}\mathbf{W}^H\mathbf{\Sigma}$.

*Proof:* See [50, Appendix G].■

**Lemma 4** indicates that, the estimation error caused by the non-ideal input cannot be neglected. For example, if the input kernel $\mathbf{\Sigma}$ behaves far from the real kernel $\mathbf{\Sigma_h}$, the value of the MSE may become significantly large due to the existence of $\mathbf{\Pi}$ in (34). To obtain more insightful results, we use the statistical kernel, $\hat{\mathbf{\Sigma}}_\mathbf{h}$ in (22), as an example to analyze the achievable MSEs. To ease the derivation, we first prove a useful expression of the MSE in **Lemma 5**.

*Lemma 5:* The MSE $\hat{\delta}$ in (34) under the imperfect kernel $\hat{\mathbf{\Sigma}}_\mathbf{h}$ can be written as a function of $\mathbf{WW}^H$, given by

$$\begin{aligned}\hat{\delta} = {}& \mathsf{Tr}\left(\left(\hat{\mathbf{\Sigma}}_\mathbf{h}^H\mathbf{\Omega}^H - \mathbf{I}_M\right)\mathbf{\Sigma_h}\left(\mathbf{\Omega}\hat{\mathbf{\Sigma}}_\mathbf{h} - \mathbf{I}_M\right)\right) \\ & + \sigma^2\mathsf{Tr}\left(\hat{\mathbf{\Sigma}}_\mathbf{h}\mathbf{\Xi}\hat{\mathbf{\Sigma}}_\mathbf{h}\right),\end{aligned} \tag{35}$$

wherein $\mathbf{\Omega}$ and $\mathbf{\Xi}$ are subfunctions of $\mathbf{WW}^H$, written as

$$\mathbf{\Omega} = \frac{\mathbf{WW}^H}{\sigma^2} - \frac{\mathbf{WW}^H}{\sigma^4}\left(\hat{\mathbf{\Sigma}}_\mathbf{h}^{-1} + \frac{\mathbf{WW}^H}{\sigma^2}\right)^{-1}\mathbf{WW}^H, \tag{36}$$

$$\begin{aligned}\mathbf{\Xi} = {}& \frac{\mathbf{WW}^H}{\sigma^4} - 2\frac{\mathbf{WW}^H}{\sigma^6}\left(\hat{\mathbf{\Sigma}}_\mathbf{h}^{-1} + \frac{\mathbf{WW}^H}{\sigma^2}\right)^{-1}\mathbf{WW}^H \\ & + \frac{\mathbf{WW}^H}{\sigma^8}\left(\left(\hat{\mathbf{\Sigma}}_\mathbf{h}^{-1} + \frac{\mathbf{WW}^H}{\sigma^2}\right)^{-1}\mathbf{WW}^H\right)^2.\end{aligned} \tag{37}$$

*Proof:* See [50, Appendix H].■

Recall that the essential difference among the three estimators, i.e., water-filling algorithm, ice-filling algorithm, and the estimator with randomly generated $\mathbf{W}$, is the form of observation matrix $\mathbf{W}$. Thus, by utilizing **Lemma 5**, the MSE of different algorithms under the imperfect kernel $\hat{\mathbf{\Sigma}}_\mathbf{h}$ can be obtained by replacing $\mathbf{WW}^H$ in $\hat{\delta}$ with their corresponding analytical expressions. Aided by some matrix operations, we can thereby prove the following lemma.

*Lemma 6:* Under the imperfect kernel $\hat{\mathbf{\Sigma}}_\mathbf{h}$, the achievable MSE for the water-filling algorithm can be written as

$$\begin{aligned}\hat{\delta}_{\text{wf}} &= \sigma^2\sum_{k=1}^{K}\frac{\lambda_k\sigma^2 + \hat{p}_k\left(\lambda_k+\sigma_\mathbf{h}^2\right)^2}{\left(\hat{p}_k\left(\lambda_k+\sigma_\mathbf{h}^2\right)+\sigma^2\right)^2} + \sigma^2\sum_{m=K+1}^{M}\frac{\hat{p}_m\sigma_\mathbf{h}^4}{\left(\hat{p}_m\sigma_\mathbf{h}^2+\sigma^2\right)^2} \\ &\overset{Q\to+\infty}{=} \mathcal{O}\left(\sigma^2M^2Q^{-1}\right),\end{aligned} \tag{38}$$

wherein $\{\lambda_1,\cdots,\lambda_M\}$ represent the $M$ non-negative eigenvalues of $\mathbf{\Sigma_h}$ in a descending order.[1] The power $\hat{p}_m$ allocated to each eigenvector is calculated by the water-filling principle $\hat{p}_m = \left(\hat{\beta} - \frac{\sigma^2}{\lambda_m+\sigma_\mathbf{h}^2}\right)^+$ for $m \in \{1,\cdots,M\}$. The water-level $\hat{\beta}$ can be determined via binary search such that $\sum_{m=1}^{M}\hat{p}_m = Q$.

Under the imperfect kernel $\hat{\mathbf{\Sigma}}_\mathbf{h}$, the achievable MSE for the ice-filling algorithm can be written as

$$\begin{aligned}\hat{\delta}_{\text{if}} &= \sigma^2\sum_{k=1}^{K}\frac{\lambda_k\sigma^2 + \hat{n}_k\left(\lambda_k+\sigma_\mathbf{h}^2\right)^2}{\left(\hat{n}_k\left(\lambda_k+\sigma_\mathbf{h}^2\right)+\sigma^2\right)^2} + \sigma^2\sum_{m=K+1}^{M}\frac{\hat{n}_m\sigma_\mathbf{h}^4}{\left(\hat{n}_m\sigma_\mathbf{h}^2+\sigma^2\right)^2} \\ &\overset{Q\to+\infty}{=} \mathcal{O}\left(\sigma^2M^2Q^{-1}\right),\end{aligned} \tag{39}$$

wherein the pilot reuse frequencies $\{\hat{n}_1,\cdots,\hat{n}_M\}$ can be obtained by **Algorithm 1** with $\hat{\mathbf{\Sigma}}_\mathbf{h}$ being the input kernel. In particular, $\sum_{m=1}^{M}\hat{n}_m = Q$ holds.

For the estimator enabled by a randomly generated $\mathbf{W}$ with sufficiently large $Q$, the achievable MSE under the imperfect kernel $\hat{\mathbf{\Sigma}}_\mathbf{h}$ can be written as

$$\begin{aligned}\hat{\delta}_{\text{rnd}} &\approx \sigma^2\sum_{k=1}^{K}\frac{\lambda_k\sigma^2 + \frac{Q}{M}\left(\lambda_k+\sigma_\mathbf{h}^2\right)^2}{\left(\frac{Q}{M}\left(\lambda_k+\sigma_\mathbf{h}^2\right)+\sigma^2\right)^2} + \sigma^2\sum_{m=K+1}^{M}\frac{\frac{Q}{M}\sigma_\mathbf{h}^4}{\left(\frac{Q}{M}\sigma_\mathbf{h}^2+\sigma^2\right)^2} \\ &\overset{Q\to+\infty}{=} \mathcal{O}\left(\sigma^2M^2Q^{-1}\right),\end{aligned} \tag{40}$$

wherein the approximate equality can be infinitely close to an equality as $Q \to \infty$.

*Proof:* See [50, Appendix I].■

From **Lemma 6**, it is easy to prove that the MSEs under the imperfect kernel $\hat{\mathbf{\Sigma}}_\mathbf{h}$ are increased by the kernel estimation error $\sigma_\mathbf{h}^2$. The reason is that, the estimation error makes the input kernel full-rank, i.e., $\hat{\mathbf{\Sigma}}_\mathbf{h} = \mathbf{\Sigma_h} + \sigma_\mathbf{h}^2\mathbf{I}_M$. Then, the power (or pilot reuse frequency) for estimation may be allocated to all eigenvectors of $\mathbf{\Sigma_h}$. However, allocating power to those eigenvectors associated with the eigenvalue $\sigma_\mathbf{h}^2$ contributes nearly no improvement to channel recovery accuracy. Besides, an imperfect kernel input also changes the weights of the posterior mean $\boldsymbol{\mu}_{\mathbf{h}|\mathbf{y}}$ in (4), which resembles a Bayesian estimator that underestimates the impact of noise.

*Remark 5: According to (38), (39), and (40), an efficient way to alleviate the effect of $\sigma_\mathbf{h}^2$ on MSEs is to introduce the prior knowledge that the real kernel $\mathbf{\Sigma_h}$ is rank-K. Then, the estimator only allocates power to the eigenvalues associated with the K-largest eigenvalues of $\hat{\mathbf{\Sigma}}_\mathbf{h}$, so that the second term*

[1] That is to say, if $\mathbf{\Sigma_h}$ is rank-$K$, we have $\lambda_m = 0$ for all $m > K$.

*in (38), (39), and (40) can be eliminated. For example, when $p_m$ for $m > K$ are forced to be zero, $\hat{\delta}_{\rm wf}$ in (38) becomes*

$$\hat{\delta}_{\rm wf} = \sigma^2 \sum_{k=1}^{K} \frac{\lambda_k \sigma^2 + p_k \left(\lambda_k + \sigma_{\mathbf{h}}^2\right)^2}{\left(p_k \left(\lambda_k + \sigma_{\mathbf{h}}^2\right) + \sigma^2\right)^2} \overset{Q\to+\infty}{=} \mathcal{O}\left(\sigma^2 K^2 Q^{-1}\right), \tag{41}$$

*wherein $p_k$ for $k \in \{1, \cdots, K\}$ is given in (7). For $\sigma_{\mathbf{h}}^2 > 0$, it is evident that $\hat{\delta}_{\rm wf}$ in (41) is lower than that in (38).*

Based on the results in **Lemma 6**, the asymptotic MSEs when the kernel estimation error $\sigma_{\mathbf{h}}^2$ is infinitely large are expressed as follows.

*Corollary 1:* When $\sigma_{\mathbf{h}}^2 \to +\infty$, the asymptotic MSEs $\hat{\delta}$ for the water-filling algorithm, ice-filling algorithm, and the estimator based on random $\mathbf{W}$ can be respectively written as

$$\hat{\delta}_{\rm wf} \overset{\sigma_{\mathbf{h}}^2\to+\infty}{=} \mathcal{O}\left(\sigma^2 \sum\nolimits_{k=1}^{L_{\rm wf}} \hat{p}_k^{-1}\right) = \mathcal{O}\left(\sigma^2 M^2 Q^{-1}\right), \tag{42}$$

$$\hat{\delta}_{\rm if} \overset{\sigma_{\mathbf{h}}^2\to+\infty}{=} \mathcal{O}\left(\sigma^2 \sum\nolimits_{k=1}^{L_{\rm if}} \hat{n}_k^{-1}\right) = \mathcal{O}\left(\sigma^2 M^2 Q^{-1}\right), \tag{43}$$

$$\hat{\delta}_{\rm rnd} \overset{\sigma_{\mathbf{h}}^2\to+\infty}{=} \mathcal{O}\left(\sigma^2 M^2 Q^{-1}\right), \tag{44}$$

where $L_{\rm wf}$ and $L_{\rm if}$ are the numbers of the positive values in $\{\hat{p}_m\}_{m=1}^M$ and $\{\hat{n}_m\}_{m=1}^M$, respectively.

*Proof:* The three asymptotic MSEs can be easily obtained by letting $\sigma_{\mathbf{h}}^2 \to +\infty$ for $\hat{\delta}_{\rm wf}$ in (38), $\hat{\delta}_{\rm if}$ in (39), and $\hat{\delta}_{\rm rnd}$ in (40), respectively. Since their derivations are similar, here we focus on the asymptotic $\hat{\delta}_{\rm wf}$ as an example. By letting $\sigma_{\mathbf{h}}^2 \to +\infty$ and removing the small-order components, we have $\hat{\delta}_{\rm wf} \overset{\sigma_{\mathbf{h}}^2\to+\infty}{=} \mathcal{O}\left(\sigma^2 \sum_{k=1}^{L_{\rm wf}} \hat{p}_k^{-1}\right)$. Since $\hat{p}_m = \left(\hat{\beta} - \frac{\sigma^2}{\lambda_m + \sigma_{\mathbf{h}}^2}\right)^+$ for $m \in \{1, \cdots, M\}$, we have $\hat{p}_m = \hat{\beta} = \frac{Q}{M}$ and $L_{\rm wf} = M$, which completes the proof.■

### C. Computational Complexity Analysis

The computational complexity of the proposed channel estimators is low in practical applications. Specifically, according to Fig. 2, the computational complexity can be divided into two components, i.e., observation matrix design in Stage 1 and Bayesian regression in Stage 2.

For the observation matrix design, the complexity of **Algorithm 1** is mainly caused by the eigenvalue decomposition and the principal eigenvalue selection, which is $\mathcal{O}(M^3 + QK)$. The complexity of **Algorithm 2** mainly depends on the search for the largest eigenvalue and the iterations required for the convergence of MM, which is $\mathcal{O}(M^3 + I_o M^2 Q)$ wherein $I_o$ is the number of iterations.

For the Bayesian regression, according to (4), we find that the recovered channel $\hat{\mathbf{h}}$ is exactly the weighted sum of received pilots $\mathbf{y}$. In particular, the weight is calculated by $\mathbf{\Sigma_h W}\left(\mathbf{W}^H \mathbf{\Sigma_h W} + \sigma^2 \mathbf{I}_Q\right)^{-1}$, thus the overall complexity of Stage 2 is $\mathcal{O}(M^3)$. It is notable that the weight only relies on the designed $\mathbf{W}$ and $\mathbf{\Sigma_h}$. It means that the weight for recovering $\mathbf{h}$ can also be calculated offline and then employed for online channel estimation.

## VI. Simulation Results

In this section, we present simulation results to verify the effectiveness of the proposed channel estimators for DASs.

TABLE I
SIMULATION PARAMETERS OF CHANNEL MODEL IN 3GPP TR 38.901

| Channel parameters | Values in [53] |
|---|---|
| Carrier frequency $f_c$ | 3.5 GHz |
| Number of clusters | 23 |
| Number of rays per cluster | 20 |
| Path gains | $\mathcal{CN}(0,1)$ |
| Incident angles | $\mathcal{U}(-90^\circ, +90^\circ)$ |
| Maximum angle spread | $\mathcal{U}(-5^\circ, +5^\circ)$ |
| Maximum delay spread | $\mathcal{U}(-30$ ns, $+30$ ns$)$ |

### A. Simulation Setup and Benchmarks

To account for a practical scenario, the 3GPP TR 38.901 channel model is used for simulations, whose key parameters are given in Table I. Otherwise specifically specified, we consider a UPA. The number of antennas is set to $M = 128$, and the numbers of horizontal antennas and vertical antennas are set to $M_x = 16$ and $M_y = 8$, respectively. The antenna spacing is set to $\frac{\lambda}{8}$. The signal-to-noise ratio (SNR) is defined as $\text{SNR} = \frac{\mathsf{E}\left(\|\mathbf{h}\|^2\right)}{\sigma^2}$. The performance is evaluated by the normalized mean square error (NMSE), which is defined as $\text{NMSE} = \mathsf{E}\left(\frac{\|\mathbf{h}-\hat{\mathbf{h}}\|^2}{\|\mathbf{h}\|^2}\right)$. 3000 Monte-Carlo experiments are carried out to plot each figure.

To demonstrate the effectiveness of the proposed ice-filling and MM algorithms, six observation matrices are considered for comparison:

- Random matrix $\mathbf{W}^{\rm Rand} \in \mathbb{C}^{M\times Q}$: The coefficients of $\mathbf{W}^{\rm Rand}$ are independently generated from the standard complex Gaussian distribution. Then, each column is normalized to satisfy the unit power constraint.
- Top-$Q$ matrix $\mathbf{W}^{{\rm Top}Q} \in \mathbb{C}^{M\times Q}$: The first $Q$ major eigenvectors of $\mathbf{\Sigma_h}$ are used to set the observation matrix $\mathbf{W}^{{\rm Top}Q}$.
- Ice-filling matrix $\mathbf{W}^{\rm IF} \in \mathbb{C}^{M\times Q}$: The proposed ice-filling algorithm is used to generate $\mathbf{W}^{\rm IF}$.
- MM matrix $\mathbf{W}^{\rm MM} \in \mathbb{C}^{M\times Q}$: The proposed MM algorithm is used to generate $\mathbf{W}^{\rm MM}$.
- Water-filling matrix $\mathbf{W}^{\rm WF} \in \mathbb{C}^{M\times Q}$: The water-filling scheme is employed to generate $\mathbf{W}^{\rm WF}$. To implement this ideal observation matrix, the noise vector, $\mathbf{z}$, in (2) is assumed to be independent of the observation matrix $\mathbf{W}^{\rm WF}$, and its distribution is fixed as $\mathcal{CN}(\mathbf{0}_Q, \sigma^2 \mathbf{I}_Q)$.
- DFT matrix $\mathbf{W}^{\rm DFT} \in \mathbb{C}^{M\times M}$: The DFT matrix requires the pilot length, $Q$, to be equal to the number of antennas, $M$. This matrix is used for performing LS algorithm.

Note that the pilot length for the DFT matrix $\mathbf{W}^{\rm DFT}$ is $Q = M = 128$, while that for the other matrices is $Q = 64$ by default. The above observation matrices are used in four channel estimators for comparison: The LS, vector approximate message passing (VAMP) [28], OMP, and MMSE algorithms. For the LS algorithm, it has to be implemented by the DFT matrix $\mathbf{W}^{\rm DFT}$ with a $Q = M$ pilot length. As for the OMP and VAMP algorithms, we conduct numerical experiments in advance to evaluate their performance under matrices, $\mathbf{W}^{\rm Rand}$, $\mathbf{W}^{{\rm Top}Q}$, $\mathbf{W}^{\rm MM}$ and $\mathbf{W}^{\rm IF}$, respectively. It is shown that the combinations of "VAMP +$\mathbf{W}^{\rm Rand}$" and "OMP

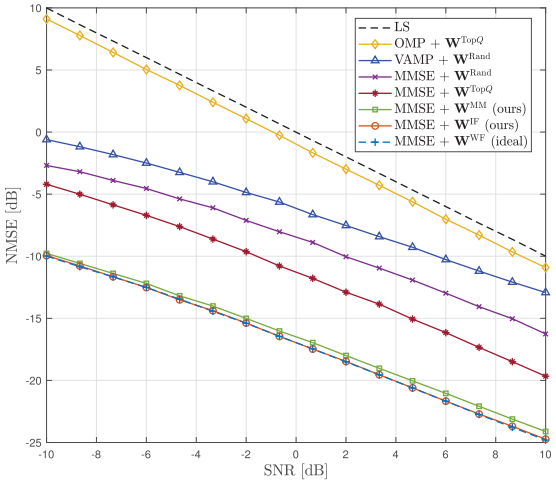


Fig. 4. The effect of SNR on NMSE performance under perfect kernel $\mathbf{\Sigma_h}$.

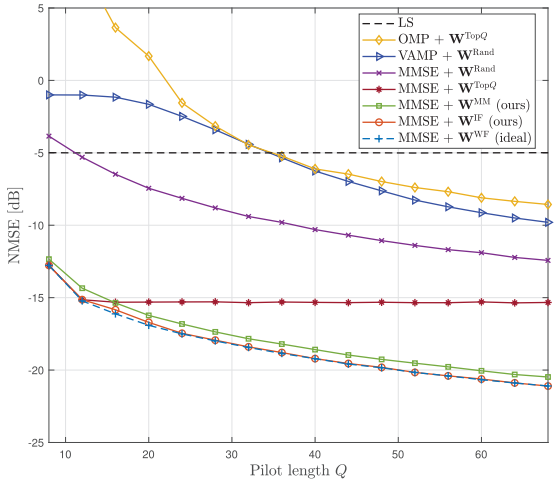


Fig. 5. The NMSE as a function of pilot length. The pilot length allocated to the LS method is fixed to 128, whereas the pilot length $Q$ for other schemes rises from 8 to 68.

$+\mathbf{W}^{\mathrm{Top}Q}$" exhibit the best channel estimation performance.[2] Thereby, the following simulations use the random matrix $\mathbf{W}^{\mathrm{Rand}}$ for VAMP and the top-$Q$ matrix $\mathbf{W}^{\mathrm{Top}Q}$ for OMP. The MMSE algorithm in (4) is the estimator adopted in this paper. We will evaluate its performance under matrices, $\mathbf{W}^{\mathrm{Rand}}$, $\mathbf{W}^{\mathrm{Top}Q}$, $\mathbf{W}^{\mathrm{MM}}$, $\mathbf{W}^{\mathrm{IF}}$, and $\mathbf{W}^{\mathrm{WF}}$, respectively, to show the superiority of the designed observation matrices.

### *B. Performance Evaluation Under Perfect Kernel*

In this subsection, we use the perfect kernel, $\mathbf{\Sigma_h}$, to trigger the water-filling, ice-filling, and MM algorithms, and then perform the Bayesian regression. The NMSE performance as a function of SNR is evaluated in Fig. 4. The SNR ranges from $-10$ dB $\sim 10$ dB. Thanks to the careful design of observation matrix, our schemes "MMSE $+\mathbf{W}^{\mathrm{MM}}$" and "MMSE $+\mathbf{W}^{\mathrm{IF}}$" remarkably outperform existing benchmarks. For example, a 10 dB NMSE gap between "MMSE $+\mathbf{W}^{\mathrm{IF}}$" and "VAMP $+\mathbf{W}^{\mathrm{Rand}}$ is visible. Moreover, the ice-filling matrix $\mathbf{W}^{\mathrm{IF}}$ can reduce the NMSE of MMSE algorithms by 5 dB compared to the top-$Q$ matrix $\mathbf{W}^{\mathrm{Top}Q}$. Notably, the NMSE performance of MMSE attained by the ice-filling matrix tightly aligns with that of the ideal water-filling matrix, given that the ice-filling intrinsically allocates $Q$ quantized units of power to the $Q$-timeslot pilots. Another finding of interest is that the NMSE gap between "MMSE $+\mathbf{W}^{\mathrm{IF}}$" and "MMSE $+\mathbf{W}^{\mathrm{MM}}$" is no higher than 0.5 dB, demonstrating the near-optimality of the proposed MM algorithm in phase-only-controllable design.

Next, we investigate the impact of pilot length, $Q$, on the NMSE performance in Fig. 5. The pilot length allocated to the LS method is fixed to 128, whereas the pilot length for the other algorithms ranges from 8 to 68. The SNR is -5 dB. It is evident from Fig. 5 that the proposed algorithms consume significantly lower pilot overhead than the other benchmarks to achieve the same NMSE level. For instance, to obtain a $-5$ dB NMSE, the pilot length required for the VAMP algorithm is larger than 30. In contrast, 8 pilots are sufficient for the proposed ice-filling and MM algorithms to attain a NMSE lower than - 10 dB. Additionally, it is notable that the NMSE performance of "MMSE $+\mathbf{W}^{\mathrm{Top}Q}$ no longer decreases when $Q > 12$. This is because the top-$Q$ matrix begins to allocate pilots to invalid eigenvectors with very small eigenvalues when $Q > 12$, while the proposed ice-filling and MM algorithms continue to select valid eigenvectors from the $1^{\mathrm{st}}$ to $12^{\mathrm{th}}$ eigenvalues. As a result, the performance gain attained by our algorithms continuously improve as the pilot length increases.

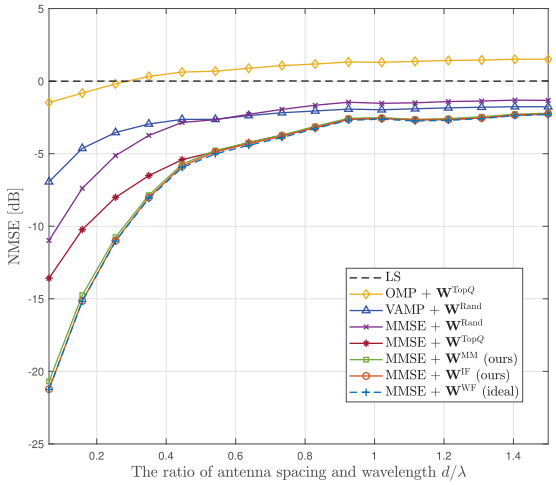


Fig. 6. NMSE versus the ratio of antenna spacing and wavelength.

In Fig. 6, we demonstrate the superiority of a DAS in precise channel estimation. The curves of the achieved NMSE performance versus the ratio of antenna spacing and wavelength, i.e., $\frac{d}{\lambda}$, are plotted. The SNR is 0 dB. We can observe that, as the normalized antenna spacing $\frac{d}{\lambda}$ decreases from $\frac{3}{2}$ to $\frac{1}{16}$, the NMSEs for all channel estimators excluding LS method declines rapidly. For example, the NMSE achieved by the proposed ice-filling algorithm is decreased by about 15 dB when the antenna spacing ranges from $\lambda/2$ to $\lambda/8$. This phenomenon is attributed to the fact that, a smaller antenna spacing leads to stronger spatial correlations of channels over different antennas, which results in a more informative

[2]This is because CS algorithms expect to make the column coherence of observation matrix as small as possible. As the designed matrix $\mathbf{W}^{\mathrm{IF}}$ might select the same eigenvectors in different time slots, it is not suitable for OMP and VAMP algorithms.

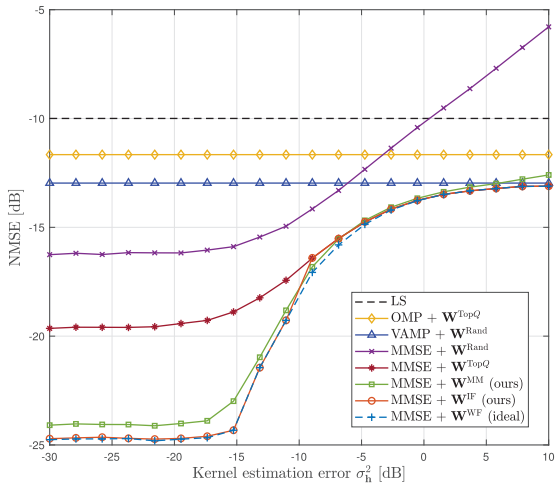


Fig. 7. The effect of kernel estimation error on NMSE performance under the imperfect kernel $\hat{\boldsymbol{\Sigma}}_{\mathbf{h}}$.

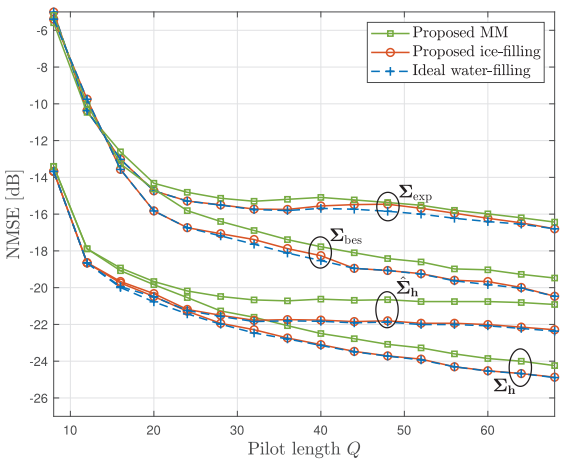


Fig. 9. The effect of pilot length on NMSE performance under the perfect kernel $\boldsymbol{\Sigma}_{\mathbf{h}}$, the Bessel kernel $\boldsymbol{\Sigma}_{\mathrm{bes}}$, and the exponential kernel $\boldsymbol{\Sigma}_{\mathrm{exp}}$.

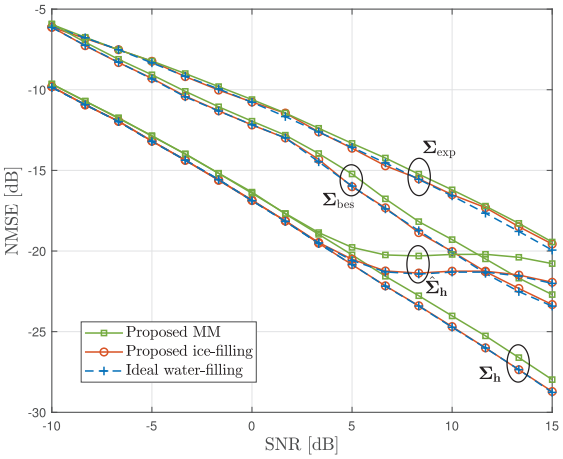


Fig. 8. The effect of SNR on NMSE performance under the perfect kernel $\boldsymbol{\Sigma}_{\mathbf{h}}$, the statistical kernel $\hat{\boldsymbol{\Sigma}}_{\mathbf{h}}$, the Bessel kernel $\boldsymbol{\Sigma}_{\mathrm{bes}}$, and the exponential kernel $\boldsymbol{\Sigma}_{\mathrm{exp}}$.

kernel $\boldsymbol{\Sigma}_{\mathbf{h}}$ for more precise channel reconstruction. Since the underpinning spatial information of wireless channels is embedded in the structured kernel, these Bayesian regression based channel estimators are enhanced under DASs.

### C. Performance Evaluation Under Imperfect Kernel

In some scenarios where the perfect prior kernel $\boldsymbol{\Sigma}_{\mathbf{h}}$ cannot be obtained, the statistical kernel and artificial kernels defined in Section IV-E can be utilized to replace the perfect kernel $\boldsymbol{\Sigma}_{\mathbf{h}}$ in Algorithms 1$\sim$ 2. We first consider the statistical kernel $\hat{\boldsymbol{\Sigma}}_{\mathbf{h}} = \boldsymbol{\Sigma}_{\mathbf{h}} + \sigma_{\mathbf{h}}^2 \mathbf{I}_M$. The NMSE as a function of the kernel estimation error $\sigma_{\mathbf{h}}^2$ is shown in Fig. 7. One can observe that all MMSE methods suffer from visible performance loss due to the imperfect kernel input. Fortunately, the proposed methods still hold the superiority among all schemes. In particular, the estimation accuracy of the proposed methods is almost not influenced when $\sigma_{\mathbf{h}}^2 < -15\,\mathrm{dB}$, which shows their high robustness to the kernel error.

We further evaluate the achievable NMSE performance under all considered kernels, including the perfect kernel $\boldsymbol{\Sigma}_{\mathbf{h}}$, the statistical kernel $\hat{\boldsymbol{\Sigma}}_{\mathbf{h}}$, and the artificial kernels $\boldsymbol{\Sigma}_{\mathrm{exp}}$ and $\boldsymbol{\Sigma}_{\mathrm{bes}}$. The estimation error of $\hat{\boldsymbol{\Sigma}}_{\mathbf{h}}$ is set as $\boldsymbol{\sigma}_{\mathbf{h}}^2 = -12\,\mathrm{dB}$. The hyper-parameters, $\eta_1$ and $\eta_2$, of $\boldsymbol{\Sigma}_{\mathrm{exp}}$ and $\boldsymbol{\Sigma}_{\mathrm{bes}}$ are determined by the binary search through several numerical experiments, which are set to $\eta_1 = 0.56$ and $\eta_2 = 0.85$ in the considered scenario. In this way, we obtain the NMSE versus the SNR and that versus the pilot length $Q$ in Fig. 8 and Fig. 9, respectively. For clarity, we only plot the curves of MMSE algorithms. From these two figures one find that, for a given kernel, the curves of "ideal water-filling", "proposed ice-filling", and "proposed MM" are very close, which indicates that these three schemes have similar robustness to imperfect kernels.

More importantly, one can find that the imperfect kernels lead to a decrease in estimation performances, while the CSI accuracy still holds considerable. Compared with the NMSE achieved by the perfect kernel, those achieved by the imperfect kernels experience an increase of about 5 dB. For example, when the SNR is set to 5 dB, the NMSEs of the proposed estimators for kernels $\boldsymbol{\Sigma}_{\mathbf{h}}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{h}}$, $\boldsymbol{\Sigma}_{\mathrm{bes}}$, and $\boldsymbol{\Sigma}_{\mathrm{exp}}$ are about $-20$ dB, $-20$ dB, $-15$ dB, and $-13$ dB, respectively. To achieve the NMSE of $-15$ dB, the required pilot lengths $Q$ for the four kernels are 10, 10, 18, and 20, respectively. Despite the performance losses, we find that the superiority of the proposed methods still hold while comparing to the baselines in Fig. 4 and Fig. 5. This observation is encouraging because it suggests that the proposed methods may not require real prior knowledge of channels. As an alternative, a virtual kernel composed of experiential parameters may be an ideal choice in practice, such that gains without pain can be achieved.

## VII. Conclusion

This paper incorporated the design of observation matrix into Bayesian channel estimation in DASs. The formulated MIM-based observation matrix design was shown to be a time-domain duality of classic MIMO precoding, which was ideally addressed by the water-filling principle. Targeting practical DAS realizations, we proposed a novel ice-filling and a MM enabled algorithms to design amplitude-and-phase controllable and phase-only controllable observation matrices, respectively. In particular, the ice-filling algorithm was proved to be a discrete approximation of water-filling, with the relative quantization error decaying rapidly. Comprehensive analyses and

numerical results validated the near-optimality of the proposed designs.

This work establishes a new understanding of channel estimation from the view of information theory. Several potential directions for extending our work are summarized as follows. In current communication systems, beam selection techniques, including beam training and beam tracking, are widely adopted CSI acquisition approaches. Designing new beam selection strategies using the idea of MIM will be interesting. Besides, the extension to more general systems, such as multi-user MIMO and wideband communications, and theoretically analyzing their channel estimation accuracy also deserve in-depth study.

## REFERENCES


[1] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[2] N. Shlezinger, G. C. Alexandropoulos, M. F. Imani, Y. C. Eldar, and D. R. Smith, "Dynamic metasurface antennas for 6G extreme massive MIMO communications," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 106–113, Apr. 2021.

[3] C. Han, L. Yan, and J. Yuan, "Hybrid beamforming for terahertz wireless communications: Challenges, architectures, and open problems," *IEEE Wireless Commun.*, vol. 28, no. 4, pp. 198–204, Aug. 2021.

[4] R. M. Dreifuerst and R. W. Heath Jr., "Massive MIMO in 5G: How beamforming, codebooks, and feedback enable larger arrays," *IEEE Commun. Mag.*, vol. 61, no. 12, pp. 18–23, Dec. 2023.

[5] K. Ying et al., "Reconfigurable massive MIMO: Harnessing the power of the electromagnetic domain for enhanced information transfer," *IEEE Wireless Commun.*, vol. 31, no. 3, pp. 125–132, Mar. 2023.

[6] 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.

[7] 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. 2023.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

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

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

[15] 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.

[16] 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.

[17] K.-K. Wong, A. Shojaeifard, K.-F. Tong, and Y. Zhang, "Fluid antenna systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1950–1962, Mar. 2021.

[18] Z. Zhang, J. Zhu, L. Dai, and R. W. Heath Jr., "Successive Bayesian reconstructor for channel estimation in fluid antenna systems," *IEEE Trans. Wireless Commun.*, vol. 24, no. 3, pp. 1992–2006, Mar. 2025.

[19] C. Han, J. M. Jornet, and I. Akyildiz, "Ultra-massive MIMO channel modeling for graphene-enabled terahertz-band communications," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Jun. 2018, pp. 1–5.

[20] C. Han and I. F. Akyildiz, "Three-dimensional end-to-end modeling and analysis for graphene-enabled terahertz band communications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 5626–5634, Jul. 2017.

[21] T. Kwon, Y.-G. Lim, B.-W. Min, and C.-B. Chae, "RF lens-embedded massive MIMO systems: Fabrication issues and codebook design," *IEEE Trans. Microw. Theory Techn.*, vol. 64, no. 7, pp. 2256–2271, Jul. 2016.

[22] O. Yurduseven, D. L. Marks, T. Fromenteze, and D. R. Smith, "Dynamically reconfigurable holographic metasurface aperture for a mills-cross monochromatic microwave camera," *Opt. Exp.*, vol. 26, pp. 5281–5291, Mar. 2018.

[23] Z. Gao, L. Dai, S. Han, Z. Wang, and L. Hanzo, "Compressive sensing techniques for next-generation wireless communications," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 144–153, Jun. 2018.

[24] 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.

[25] J. Lee, G.-T. Gil, and Y. H. Lee, "Channel estimation via orthogonal matching pursuit for hybrid MIMO systems in millimeter wave communications," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2370–2386, Jun. 2016.

[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, 2020.

[27] 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.

[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. Zhu, H. Guo, and V. K. N. Lau, "Bayesian channel estimation in multi-user massive MIMO with extremely large antenna array," *IEEE Trans. Signal Process.*, vol. 69, pp. 5463–5478, 2021.

[30] N. González-Prelcic, H. Xie, J. Palacios, and T. Shimizu, "Wideband channel tracking and hybrid precoding for mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2161–2174, Apr. 2021.

[31] X. Ma and Z. Gao, "Data-driven deep learning to design pilot and channel estimator for massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5677–5682, May 2020.

[32] Z. Hu, Y. Chen, and C. Han, "PRINCE: A pruned AMP integrated deep CNN method for efficient channel estimation of millimeter-wave and terahertz ultra-massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8066–8079, Nov. 2023.

[33] H. Lei, J. Zhang, H. Xiao, X. Zhang, B. Ai, and D. W. K. Ng, "Channel estimation for XL-MIMO systems with polar-domain multi-scale residual dense network," *IEEE Trans. Veh. Technol.*, vol. 73, no. 1, pp. 1479–1484, Jan. 2024.

[34] A. Alkhateeb, G. Leus, and R. W. Heath Jr., "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[35] Z. Xiao, T. He, P. Xia, and X. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.

[36] 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.

[37] 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.

[38] Y. Tsaig and D. L. Donoho, "Extensions of compressed sensing," *Signal Process.*, vol. 86, no. 3, pp. 549–571, Mar. 2006.

[39] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[40] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Proc. Adv. Neural Inf. Process. Syst.*, Nov. 1995, pp. 1–7.

[41] 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.

[42] E. Schulz, M. Speekenbrink, and A. Krause, "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions," *J. Math. Psychol.*, vol. 85, pp. 1–16, Aug. 2018.

[43] B. Xu, Y. Chen, Q. Cui, X. Tao, and K.-K. Wong, "Sparse Bayesian learning-based channel estimation for fluid antenna systems," *IEEE Wireless Commun. Lett.*, vol. 14, no. 2, pp. 325–329, Feb. 2025.
[44] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Channel modeling and channel estimation for holographic massive MIMO with planar arrays," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 997–1001, May 2022.
[45] M. B. Khalilsarai, T. Yang, S. Haghighatshoar, and G. Caire, "Structured channel covariance estimation from limited samples in massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–7.
[46] S. Park and R. W. Heath Jr., "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.
[47] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.
[48] T. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
[49] Z. Pi, "Optimal transmitter beamforming with per-antenna power constraints," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 3779–3784.
[50] M. Cui, Z. Zhang, L. Dai, and K. Huang, "Ice-filling: Near-optimal channel estimation for dense array systems," 2024, *arXiv:2404.06806*.
[51] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.
[52] J. Song, P. Babu, and D. P. Palomar, "Sequence design to minimize the weighted integrated and peak sidelobe levels," *IEEE Trans. Signal Process.*, vol. 64, no. 8, pp. 2051–2064, Apr. 2016.
[53] Study on Channel Model for Frequencies From 0.5 To 100 GHz, Standard TR 38.901, 3GPP, Dec. 2019.

**Mingyao Cui** (Graduate Student Member, IEEE) received the B.E. and M.S. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2020 and 2023, respectively. He is currently pursuing the Ph.D. degree with The University of Hong Kong. His research interests include Rydberg atomic receiver, mmWave/THz communications, and near-field MIMO communications. He was awarded the NSFC Young Student Basic Research Program (Ph.D. Candidate) in 2024, the HKPF Scholarship in 2023, the National Scholarship in 2022, and the IEEE ICC Outstanding Demo Award in 2022.

**Zijian Zhang** (Graduate Student Member, IEEE) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2020, where he is currently pursuing the Ph.D. degree in electronic engineering.

He is also an amateur in wireless localization and robotics. He has authored several articles for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE TRANSACTIONS ON COMMUNICATIONS. His research interests include massive MIMO, holographic MIMO (H-MIMO), reconfigurable intelligent surfaces (RISs), and fluid antenna systems (FASs). He received the National Scholarship in 2019 and 2024. In 2024, he won the Special Scholarship of Tsinghua University, which annually awards ten students out of over 40000 graduate students in Tsinghua University. He was listed in Stanford University's World's Top 2% Scientists in 2023.

**Linglong Dai** (Fellow, IEEE) received the B.S. degree from Zhejiang University, Hangzhou, China, in 2003, the M.S. degree (Hons.) from China Academy of Telecommunications Technology, Beijing, China, in 2006, and the Ph.D. degree (Hons.) from Tsinghua University, Beijing, in 2011. From 2011 to 2013, he was a Post-Doctoral Research Fellow with the Department of Electronic Engineering, Tsinghua University, where he was an Assistant Professor from 2013 to 2016, an Associate Professor from 2016 to 2022, and has been a Professor since 2022. He has co-authored the Book "*MmWave Massive MIMO: A Paradigm for 5G*" (Academic Press, 2016). He has authored or co-authored over 100 IEEE journal articles and over 60 IEEE conference papers. He also holds over 20 granted patents. His current research interests include massive MIMO, reconfigurable intelligent surface (RIS), millimeter-wave and terahertz communications, near-field communications, machine learning for wireless communications, and electromagnetic information theory. He received the five IEEE best paper awards at the IEEE ICC 2013, the IEEE ICC 2014, the IEEE ICC 2017, the IEEE VTC 2017-Fall, the IEEE ICC 2018, and the IEEE GLOBECOM 2023. He also received Tsinghua University Outstanding Ph.D. Graduate Award in 2011, Beijing Excellent Doctoral Dissertation Award in 2012, China National Excellent Doctoral Dissertation Nomination Award in 2013, the URSI Young Scientist Award in 2014, the IEEE Transactions on Broadcasting Best Paper Award in 2015, the Electronics Letters Best Paper Award in 2016, the National Natural Science Foundation of China for Outstanding Young Scholars in 2017, the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2017, the IEEE ComSoc Asia-Pacific Outstanding Paper Award in 2018, China Communications Best Paper Award in 2019, the IEEE Access Best Multimedia Award in 2020, the IEEE Communications Society Leonard G. Abraham Prize in 2020, the IEEE ComSoc Stephen O. Rice Prize in 2022, the IEEE ICC Outstanding Demo Award in 2022, and the National Science Foundation for Distinguished Young Scholars in 2023. He was listed as a Highly Cited Researcher by Clarivate Analytics from 2020 to 2024.

**Kaibin Huang** (Fellow, IEEE) received the B.Eng. and M.Eng. degrees from the National University of Singapore and the Ph.D. degree from The University of Texas at Austin, all in electrical engineering. He is currently a Professor and the Head of the Department of Electrical and Electronic Engineering, The University of Hong Kong (HKU), Hong Kong. He received the IEEE Communication Society's 2021 Best Survey Paper, the 2019 Best Tutorial Paper, the 2019 and 2023 Asia-Pacific Outstanding Paper, the 2015 Asia-Pacific Best Paper Award, and the best paper awards at IEEE GLOBECOM 2006 and IEEE/CIC ICCC 2018. He has been named as a Highly Cited Researcher by Clarivate from 2019 to 2023 and an AI 2000 Most Influential Scholar (Top 30 in Internet of Things) from 2023 to 2024. He was an IEEE Distinguished Lecturer of both the IEEE Communications Society and the IEEE Vehicular Technology Society. He is a member of the Engineering Panel of Hong Kong Research Grants Council (RGC) and a RGC Research Fellow (2021 Class). He received the Outstanding Teaching Award from Yonsei University, South Korea, in 2011. He served as the Lead Chair for the Wireless Communications Symposium of IEEE GLOBECOM 2017 and the Communication Theory Symposium of IEEE GLOBECOM 2023 and 2014 and the TPC Co-Chair for IEEE PIMRC 2017 and IEEE CTW 2023 and 2013. He is the Founding President of the HKU Chapter of National Academy of Inventors. He is an Area Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING, and IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING. Previously, he served on the editorial boards for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and IEEE WIRELESS COMMUNICATIONS LETTERS. He has guest edited special issues of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, and *IEEE Communications Magazine*, and IEEE NETWORK.