# Training Optimization for Hybrid MIMO Communication Systems

Chengwen Xing, *Member, IEEE*, Dekang Liu, Shiqi Gong, *Student Member, IEEE*, Wei Xu, *Senior Member, IEEE*, Sheng Chen, *Fellow, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*— Channel estimation is conceived for hybrid multiple-input multiple-output (MIMO) communication systems. Both mean square error minimization and mutual information maximization are used as our performance metrics and a pair of low-complexity channel estimation schemes are proposed. In each scheme, the training sequence and the analog matrices of the transmitter and receiver are jointly optimized. We commence by designing the optimal training sequences and analog matrices for the first scheme. Upon relying on the resultant optimal structures, the training optimization problems are substantially simplified and the nonconvexity resulting from the analog matrices can be overcome. In the second scheme, the channel estimation and data transmission share the same analog matrices, which beneficially reduces the overhead of optimizing the associated analog matrices. Therefore, a composite channel matrix is estimated instead of the true channel matrix. By exploiting the statistical optimization framework advocated, the analog matrices can be designed independently of the training sequence. Based on the resultant analog matrices, the training sequence can then be efficiently designed according to diverse channel statistics and performance metrics. Finally, we conclude by quantifying the performance benefits of the proposed estimation schemes.**

***Index Terms*—Hybrid MIMO communications, analog matrices, channel estimation, training optimization.**

## I. Introduction

MULTIPLE antenna aided multiple-input multiple-output (MIMO) techniques [1]–[4] are recognized as one of the important pillars of next-generation wireless networks. In the quest for increased degrees of freedom, the scale of MIMO systems has been increasing as they have evolved. Massive MIMO or large-scale MIMO schemes constitute an important enabling technology, which have in fact already been widely used in certain communication applications. For example, on satellites the number of antennas may be as high as 300. Furthermore, ships and other large-bodied vehicles have also used large-scale antenna arrays. However, the potential performance gains are critically dependent on the availability and accuracy or the absence of channel state information (CSI). Therefore, channel estimation becomes a critical part of various MIMO communication systems [5]–[12].

Generally, a classic channel estimation task consists of two components, training optimization and channel estimator design. There is a rich body of literature on channel estimation. When the channel statistics are unknown, the least squares (LS) channel estimator is preferred and random training sequences having a white spectrum constitute the optimal choice. By contrast, when the channel statistics are known, the maximum likelihood (ML) or the linear minimum mean square error (LMMSE) channel estimator [12]–[14] are the natural choices. However, in this case the training optimization is more complicated. Moreover, when specific constraints have to be taken into account, such as the per-antenna power constraints, the training optimization becomes even more complex [15].

The channel estimation problem of massive MIMO systems has been extensively studied in the literature [11], [16]–[20]. The key behind these contributions is how to exploit the specific structure of massive MIMO channels for reducing the escalating overhead for channel estimation [11], [20]–[25]. For millimeter-wave (mmWave) based massive MIMO systems, the sparsity of the mmWave channel impulse responses (CIRs) has been exploited in [11], [22]. By reconsidering channel estimation with special emphasis on the angular domain [21]–[23], the resultant channel estimation algorithms become capable of improving the estimation accuracy, despite their reduced overhead.

For large-scale MIMO systems the popular family of hybrid structure constitutes an economic way of striking a

Manuscript received August 29, 2019; revised February 15, 2020; accepted May 3, 2020. Date of publication May 18, 2020; date of current version August 12, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61722104, Grant 61671058, and Grant 61620106001, and in part by the Ericsson. The work of Lajos Hanzo was supported in part by the Engineering and Physical Sciences Research Council projects EP/N004558/1, EP/P034284/1, EP/P034284/1, and EP/P003990/1 (COALESCE), in part by the Royal Society's Global Challenges Research Fund Grant, and in part by the European Research Council's Advanced Fellow Grant QuantCom. The associate editor coordinating the review of this article and approving it for publication was I. Guvenc. *(Corresponding author: Chengwen Xing.)*

Chengwen Xing, Dekang Liu, and Shiqi Gong are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: xingchengwen@gmail.com; dkliu17@126.com; gsqyx@163.com).

Wei Xu is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: wxu@seu.edu.cn).

Sheng Chen is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K., and also with the Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia (e-mail: sqc@ecs.soton.ac.uk).

Lajos Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.u).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2020.2993694

performance vs. complexity trade-off [26], [27].Explicitly, a hybrid MIMO structure is composed of an analog part and a digital part [26]. In the analog part, only signal phases are adjusted by analogue phase-shifters. Although these hybrid structures strike a compelling performance vs. complexity trade-off for large-scale MIMOs, they impose new challenges. In particular, because of the nonconvex nature of optimizing the analog part [26], the holistic optimization of hybrid MIMO systems is more challenging than that of their fully digital counterparts [27]. Since the knowledge of the CIR is required, channel estimation is an essential prerequisite for hybrid transceiver designs [27]. It is worth noting that channel estimation must be performed in the digital domain. Explicitly, channel estimation conceived for hybrid MIMO channels is much more challenging than for its fully digital counterpart.

Moreover, many hybrid MIMO communication systems operate at microwave frequencies [28], [29], which are lower than mmWave carriers. The challenge is that their CIRs are typically not sparse. In this case we have to estimate large-scale MIMO channels, when conceiving hybrid MIMO systems. This is more challenging than traditional channel estimation owing to the constantmodulus constraints imposed on the associated analog matrices. Hence the conventional designs are unsuitable for striking an attractive performance vs. complexity trade-off for large-scale hybrid MIMO systems.

Against this backdrop, we conceive efficient channel estimation techniques for hybrid MIMO systems in which both the transmitter and receiver are equipped with hybrid antenna arrays. Given our specific power and constant modulus constraints, the training sequences and analog matrices are optimized based on the sum-mean square error (MSE) minimization and mutual information maximization. The main contributions of the proposed channel estimation schemes are as follows:

- Compared to the existing channel estimation schemes conceived for large-scale MIMO systems, the training optimization framework proposed in this paper is not limited to sparse mmWave or THz CIRs. Explicitly, it can also be readily applied to micro-wave channels in a wide range of application scenarios. In contrast to the fully digital systems of [30], in the current paper we investigate a hybrid MIMO system which consists of an analog and a digital part. Hybrid structured antenna arrays strike an attractive tradeoff between the system performance and hardware cost. As for the analog parts, all the elements of the analog matrix have the same constant modulus. In other words, only the phase of each element in the analog matrix is adjustable. Therefore, the training optimization of the sum-MSE minimization and mutual information maximization conceived for hybrid MIMO systems is much more challenging than that of their counterparts in fully digital MIMO systems.
- Explicitly, both sum MSE minimization and mutual information maximization are considered for training optimization. Our performance metrics highlight the trade-offs, when estimating the different elements of the MIMO channel matrix. Hence this treatise fills a gap, given the paucity of mutual information maximization based training optimization solutions for hybrid MIMO systems.
- In order to avoid high dimensional computations, a pair of low-complexity channel estimation schemes are proposed for hybrid MIMO systems. In the first scheme, the optimal structures of training sequence and analog matrices are derived, based on which the joint optimization problem can be substantially simplified. In the second scheme, in order to reduce the overheads, in the channel estimation and data transmission the same analog matrices are used at both the transmitter and receiver. Then a composite channel is estimated instead of the true channel matrix. A statistical optimization framework is proposed, based on which the analog matrices can be designed. Then the training sequence can be efficiently optimized according to the different statistical parameters of the channel.

The whole paper is organized as follows. Section II gives the signal model for channel estimation in hybrid MIMO systems. In Section III, left-right channel estimation schemes are investigated. Then a composite channel estimation is presented in Section IV. The simulation results are given in Section V to assess the performance of the proposed channel estimation algorithms. Our conclusions are drawn in Section VI.

*Notation:* The matrix $\boldsymbol{Z}^{\frac{1}{2}}$ is the Hermitian square root of positive semidefinite $\boldsymbol{Z}$. The identity matrix of appropriate dimension is denoted by $\boldsymbol{I}$, and $\otimes$ denotes Kronecker product. The expressions $\boldsymbol{\Lambda} \searrow$ and $\boldsymbol{\Lambda} \nearrow$ represent a rectangular or square diagonal matrix with the diagonal elements in descending order and ascending order, respectively. In addition, $\|\cdot\|_F$ is the matrix Frobenius norm. $[\boldsymbol{Y}]_{:,N}$ consists of the first $N$ columns of $\boldsymbol{Y}$, $[\boldsymbol{Y}]_{[1:N_1;1:N_2]}$ is the sub-matrix consisting of the first $N_1$ rows and the first $N_2$ columns of $\boldsymbol{Y}$. For further clarification, some important math definitions in the following derivations are summarized in Table I.

## II. Signal Model for Channel Estimation

In the system, the transmitter and the receiver are both equipped with hybrid structured antenna arrays. At both the transmitter and the receiver, the number of radio frequency (RF) chains is smaller than that of antennas. Due to the hybrid structure, it is impossible to estimate the channel matrix using a constant analog matrix at the transmitter or receiver. To clarify this fact, a brief discussion is first given.

Consider that a training sequence $\boldsymbol{X}_1$ is multiplied by an analog matrix $\boldsymbol{F}_{\mathrm{A},1}$ before being transmitted to the destination. At the destination, the received signal $\boldsymbol{Y}_{\mathrm{F},1}$ is given by

$$\boldsymbol{Y}_{\mathrm{F},1} = \boldsymbol{H}\boldsymbol{F}_{\mathrm{A},1}\boldsymbol{X}_1 + \boldsymbol{N}_1, \tag{1}$$

where $\boldsymbol{H}$ is the channel matrix to be estimated and $\boldsymbol{N}_1$ is the additive noise matrix at the destination. The channel matrix satisfies the following Kronecker product structure [31]

$$\boldsymbol{H} = \boldsymbol{\Sigma}_{\mathrm{H}}^{\frac{1}{2}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}_{\mathrm{H}}^{\frac{1}{2}}, \tag{2}$$

where each element of $\boldsymbol{H}_{\mathrm{W}}$ is an independently identically distributed (i.i.d.) Gaussian variable with zero mean and unit variate. The matrices $\boldsymbol{\Sigma}_{\mathrm{H}}$ and $\boldsymbol{\Psi}_{\mathrm{H}}$ are the receive and transmit correlation matrices of the MIMO channel, respectively. Since channel estimation is performed in digital domain, the received training signal is first multiplied by an analog matrix $\boldsymbol{G}_{\mathrm{A},1}$

TABLE I
THE LIST OF MATH NOTATIONS

| Variable | Definition |
|---|---|
| $\boldsymbol{\Sigma}_{\mathrm{H}}$, $\boldsymbol{\Psi}_{\mathrm{H}}$ | The receive and transmit correlation matrices of the MIMO channel $\boldsymbol{H}$. |
| $\boldsymbol{\Sigma}_{\mathrm{N}}$, $\boldsymbol{\Psi}_{\mathrm{N}}$ | The receive and time-domain correlation matrices of the additive noise $\boldsymbol{N}$. |
| $\boldsymbol{X}$ | The training sequence to be optimized at the transmitter. |
| $\boldsymbol{G}_{\mathrm{A}}$, $\boldsymbol{F}_{\mathrm{A}}$ | The receive and transmit analog matrices. |
| $\boldsymbol{G}_{\mathrm{D,L}}$, $\boldsymbol{G}_{\mathrm{D,R}}$ | The left and right channel estimators in Section III. |
| $\boldsymbol{U}_{\mathrm{DFT}}$, | The DFT matrix. |
| $\alpha_1, \beta_1$ | The auxiliary variables in the channel estimation MSE matrix $\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X})$. |
| $P$ | The maximum transmit power |
| $\boldsymbol{G}_{\mathrm{D}}$ | The MMSE channel estimator in Section IV. |
| $\alpha_2$ | The auxiliary variable in the composite channel estimation MSE matrix $\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X}, \boldsymbol{G}_{\mathrm{A}}, \boldsymbol{G}_{\mathrm{D}})$. |
| $\mathrm{vec}(\boldsymbol{Y})$ | The vectorization operation which stacks the columns of $\boldsymbol{Y}$ to construct a bigger column vector. |
| $\mathrm{Tr}(\boldsymbol{Y})$ | The trace of $\boldsymbol{Y}$. |
| $\mathbb{E}\{\cdot\}$ | The mathematical expectation. |

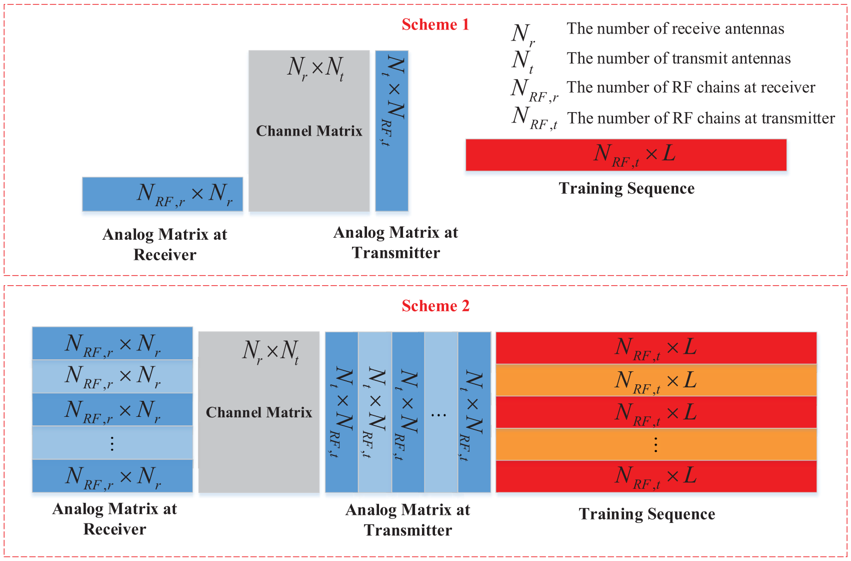

Fig. 1. Two potential approaches for channel estimation for hybrid MIMO systems.

before being processed by the fully digital channel estimator. Therefore, for channel estimation, the received signal in the digital domain is given by

$$\boldsymbol{Y}_{\mathrm{D},1} = \boldsymbol{G}_{\mathrm{A},1}\boldsymbol{H}\boldsymbol{F}_{\mathrm{A},1}\boldsymbol{X}_1 + \boldsymbol{G}_{\mathrm{A},1}\boldsymbol{N}_1. \tag{3}$$

Naturally, channel estimation aims to recover the desired channel matrix $\boldsymbol{H}$ from the received signal $\boldsymbol{Y}_{\mathrm{D},1}$. Unfortunately, as illustrated by Scheme 1 in Fig. 1, this task is impossible based on (3). This is because for hybrid MIMO systems, $\boldsymbol{G}_{\mathrm{A},1}$ is a fat matrix and $\boldsymbol{F}_{\mathrm{A},1}$ is a tall matrix. Thus based on the previous signal model, no matter how long the column of $\boldsymbol{X}_1$, the matrix $\boldsymbol{F}_{\mathrm{A},1}\boldsymbol{X}_1$ is usually a matrix that is rank deficient, and thus it is impossible to estimate the channel matrix accurately. It is worth noting that in some special cases, there may exist some special structures that can be exploited so that channel estimation based on (3) is achievable. For example, for mmWave channels, the channel matrices exhibit sparsity structures, which means that the number of parameters to estimate, i.e., the true dimension of the MIMO channel matrix, is much smaller than the size of the channel matrix. Then with aid of compressed sensing techniques, channel estimation is achievable. However, this result does not hold for the general channel matrix, e.g., macro-wave channels. In this work, we focus on the channel estimation for a general large dimensional channel matrix, where sparsity property does not exist.

Thus to overcome the problem imposed by dimension constraints, in the training transmission, the analog matrices at both the transmitter and receiver should be adjusted. The idea is similar to beam scan algorithms. Based on this idea as shown by Scheme 2 in Fig. 1 for channel estimation, the corresponding signal model is expressed by

$$\boldsymbol{Y}_{\mathrm{D}} = \boldsymbol{G}_{\mathrm{A}}\boldsymbol{H}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X} + \boldsymbol{G}_{\mathrm{A}}\boldsymbol{N}, \tag{4}$$

where $\boldsymbol{N}$ is the corresponding noise matrix, and

$$\begin{aligned} \boldsymbol{G}_{\mathrm{A}} &= \left[\boldsymbol{G}_{\mathrm{A},1}^{\mathrm{T}} \cdots \boldsymbol{G}_{\mathrm{A,K}}^{\mathrm{T}}\right]^{\mathrm{T}}, \\ \boldsymbol{F}_{\mathrm{A}} &= \left[\boldsymbol{F}_{\mathrm{A},1} \cdots \boldsymbol{F}_{\mathrm{A,K}}\right], \\ \boldsymbol{X} &= \left[\boldsymbol{X}_1^{\mathrm{T}} \cdots \boldsymbol{X}_K^{\mathrm{T}}\right]^{\mathrm{T}}. \end{aligned} \tag{5}$$

$\boldsymbol{F}_{\mathrm{A}}$ should be row full rank and $\boldsymbol{G}_{\mathrm{A}}$ should be column full rank. The composite noise term $\boldsymbol{N}$ in (4) also satisfies the following Kronecker product structure [35]

$$\boldsymbol{N} = \boldsymbol{\Sigma}_{\mathrm{N}}^{\frac{1}{2}} \boldsymbol{N}_{\mathrm{W}} \boldsymbol{\Psi}_{\mathrm{N}}^{\frac{1}{2}}, \tag{6}$$

where each element of $\boldsymbol{N}_{\mathrm{W}}$ is an i.i.d. Gaussian random variable with zero mean and unit variance. It is worth highlighting that within the coherence time the channel matrix is considered to be time-invariant. Based on (4), the estimation of the channel matrix $\boldsymbol{H}$ becomes achievable.

The remaining task of is to recover $\boldsymbol{H}$ from $\boldsymbol{Y}_{\mathrm{D}}$. Following the idea in traditional full digital MIMO channel estimation, the signal model (4) may first be transferred into a vector form

$$\mathrm{vec}(\boldsymbol{Y}_{\mathrm{D}}) = \Big(\big(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{F}_{\mathrm{A}}^{\mathrm{T}}\big) \otimes \boldsymbol{G}_{\mathrm{A}}\Big) \mathrm{vec}(\boldsymbol{H}) + \big(\boldsymbol{I} \otimes \boldsymbol{G}_{\mathrm{A}}\big) \mathrm{vec}(\boldsymbol{N}). \tag{7}$$

In the full digital MIMO case without the analog matrices $\boldsymbol{F}_{\mathrm{A}}$ and $\boldsymbol{G}_{\mathrm{A}}$, this model is of course extensively used. In hybrid MIMO, however, the existence of the analog matrices $\boldsymbol{F}_{\mathrm{A}}$ and $\boldsymbol{G}_{\mathrm{A}}$ makes the channel estimation more challenging.

Moreover, when $\boldsymbol{H}$ is a very large matrix, the widely used channel estimators, such as the LS and LMMSE estimators, suffer from prohibitively high complexity. To see why vectorization based channel estimation is not preferred in practical implementations, we examinze its complexity. It is well-known that the computation complexity of matrix inverse is on the order of $N^3$, denoted as $\mathsf{O}\big(N^3\big)$, where $N$ is the matrix dimension. Then for a channel matrix of dimension $128 \times 128$, the complexity of the LS or LMMSE channel estimator is $\mathsf{O}\big((128 \times 128)^3\big)$, which is excessive at the time of writing.

Therefore, the key task of training optimization for hybrid MIMO systems is how to design a low-complexity channel estimator under various physical constraints, such as power constraints at the transmitter and the constant modulus constraints on the analog matrices, while maintaining excellent estimation accuracy. This motivates our work.

*Remark:* We would like to point out that deep learning (DL) algorithms can also be chosen for estimating the channel matrix of the hybrid massive MIMO systems. This approach is different from our signal processing based perspective. Generally speaking, DL algorithms view the estimation procedure as a black box, while signal processing techniques usually aim for revealing some physical insights. DL algorithms are more suited to the problems having no signal models. This impediment may be circumvented at the cost of high computational complexity and a large amount of training overhead.

## III. Left-Right Channel Estimation

Based on the signal model (4), in order to control the dimension of channel estimator, the following channel estimator is proposed, in which the channel matrix is estimated as

$$\widehat{\boldsymbol{H}} = \boldsymbol{G}_{\mathrm{D,L}} \boldsymbol{Y}_{\mathrm{D}} \boldsymbol{G}_{\mathrm{D,R}}, \tag{8}$$

where $\boldsymbol{G}_{\mathrm{D,L}}$ and $\boldsymbol{G}_{\mathrm{D,R}}$ are the left and right channel estimators, respectively. Different from the traditional vectorization based channel estimator, here the high dimensional channel estimator is replaced by two low-dimensional channel estimators, and the corresponding channel estimation complexity is significantly reduced. We now investigate in depth how to find the left and right channel estimators in the both scenarios of with and without channel statistical information.

### A. Without Channel Statistics

We may recover $\boldsymbol{H}$ using the idea of ‘zero-forcing’, that is, using the following $\boldsymbol{G}_{\mathrm{D,L}}$ and $\boldsymbol{G}_{\mathrm{D,R}}$

$$\begin{aligned} \boldsymbol{G}_{\mathrm{D,L}} &= \big(\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}} \boldsymbol{G}_{\mathrm{A}}\big)^{-1} \boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}, \\ \boldsymbol{G}_{\mathrm{D,R}} &= \boldsymbol{X}^{\mathrm{H}} \boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}} \big(\boldsymbol{F}_{\mathrm{A}} \boldsymbol{X} \boldsymbol{X}^{\mathrm{H}} \boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)^{-1}. \end{aligned} \tag{9}$$

Clearly, if the following two equalities hold

$$\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}} \boldsymbol{G}_{\mathrm{A}} \propto \boldsymbol{I}, \quad \boldsymbol{F}_{\mathrm{A}} \boldsymbol{X} \boldsymbol{X}^{\mathrm{H}} \boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}} \propto \boldsymbol{I}, \tag{10}$$

then channel estimation is greatly simplified. By using the white sequence, we have $\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}} \propto \boldsymbol{I}$, and (10) becomes

$$\boldsymbol{F}_{\mathrm{A}} \boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}} \propto \boldsymbol{I}. \tag{11}$$

Therefore, we can design the analog matrices $\boldsymbol{F}_{\mathrm{A}}$ and $\boldsymbol{G}_{\mathrm{A}}$ appropriately to achieve (11). Each element of $\boldsymbol{G}_{\mathrm{A}}$ and $\boldsymbol{F}_{\mathrm{A}}$ has constant modulus. To satisfy these constraints, the most natural choice is to construct $\boldsymbol{G}_{\mathrm{A}}$ and $\boldsymbol{F}_{\mathrm{A}}$ based on discrete Fourier transform (DFT) matrix.

It can be seen that for large-scale MIMO channel estimation without any channel statistic information, the traditional white sequence is still an effective candidate. It is worth highlighting that this ‘zero-forcing’ channel estimation algorithm suffers from a serious performance loss in low signal-to-noise ratio (SNR) regime. Specifically, the noise power may be significantly enhanced by this channel estimator [36].

### B. LMMSE Channel Estimator

Given the channel statistics, specifically, the second order statistics of the channel matrix and the noise covariance matrix, the LMMSE estimator generally offers much better estimation accuracy than the previous ‘zero-forcing’ design. Hence we focus on the LMMSE channel estimation for hybrid MIMO systems with the aid of channel statistics. It is worth emphasizing that because of $\boldsymbol{G}_{\mathrm{A}}$, it is challenging to directly estimate $\boldsymbol{H}$ in a linear form. In our work, we adopt two steps to realize channel estimation. At the first step, $\boldsymbol{G}_{\mathrm{D,L}} \boldsymbol{G}_{\mathrm{A}} \boldsymbol{H}$ is estimated instead of $\boldsymbol{H}$. The corresponding estimation MSE matrix is derived to be (12) as shown at the bottom of the next page.

In the signal model, the training sequence is on the right-hand side of the channel matrix and hence the right digital channel estimator will play a more important role in the channel estimation. Note that there is no constraint imposed on the digital estimator $\boldsymbol{G}_{\mathrm{D,R}}$ at the destination. The sum MSE

$\mathrm{Tr}\left(\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X}, \boldsymbol{G}_{\mathrm{A}}, \boldsymbol{G}_{\mathrm{D,L}}, \boldsymbol{G}_{\mathrm{D,R}})\right)$ is a quadratic function with respect to $\boldsymbol{G}_{\mathrm{D,R}}$ Then based on complex matrix derivatives the optimal $\boldsymbol{G}_{\mathrm{D,R}}$ can be derived in closed-form as

$$\begin{aligned}\boldsymbol{G}_{\mathrm{D,R}}^{\mathrm{opt}} = \Big(&\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big)\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X} \\ &+ \boldsymbol{\Psi}_{\mathrm{N}}\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\big)\Big)^{-1}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}} \\ &\times \mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big). \qquad (13)\end{aligned}$$

Based on the properties of positive semi-definite matrices, the following matrix inequality holds for the right digital channel estimator [36]

$$\begin{aligned}&\boldsymbol{\Phi}_{\mathrm{MSE}}\big(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X}, \boldsymbol{G}_{\mathrm{A}}, \boldsymbol{G}_{\mathrm{D,L}}, \boldsymbol{G}_{\mathrm{D,R}}^{\mathrm{opt}}\big) \\ &\qquad \preceq \boldsymbol{\Phi}_{\mathrm{MSE}}\big(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X}, \boldsymbol{G}_{\mathrm{A}}, \boldsymbol{G}_{\mathrm{D,L}}, \boldsymbol{G}_{\mathrm{D,R}}\big). \qquad (14)\end{aligned}$$

In other words, the optimal right digital channel estimator minimizes the channel estimation MSE matrix in the positive semi-definite matrix domain.

Substituting $\boldsymbol{G}_{\mathrm{D,R}}^{\mathrm{opt}}$ into (12), the channel estimation MSE matrix (12) is reformulated as (15) (shown at the bottom of this page). It is worth highlighting that from (15), the term $\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big)$ can be interpreted as a scaling factor which does not affect the system performance. If the scaling factor is taken into account in optimization, there will be a trivial conclusion that the optimal $\boldsymbol{G}_{\mathrm{D,L}}$ should be an all-zero matrix as in this case the channel estimation MSE will be zero. This is definitely wrong because when $\boldsymbol{G}_{\mathrm{D,L}}$ is zero, the task for $\boldsymbol{G}_{\mathrm{D,R}}$ becomes to estimate a channel matrix with zero covariance matrix. In this case, the corresponding MSE is zero but this result is meaningless. If we focus on the normalized MSE, this term will be removed directly.

The optimal $\boldsymbol{G}_{\mathrm{A}}$ and $\boldsymbol{G}_{\mathrm{D,L}}$ should minimize the following cost function to ensure that the estimation is accurate

$$\min\ \mathbb{E}\big\{\|\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H} - \boldsymbol{H}\|_F^2\big\}, \qquad (16)$$

based on which the optimal $\boldsymbol{G}_{\mathrm{D,L}}$ is given by

$$\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{opt}} = \big(\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}\big)^{-1}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}. \qquad (17)$$

In order to guarantee the channel estimation is feasible the analog estimator $\boldsymbol{G}_{\mathrm{A}}$ should be column full rank. The simplest method is to choose corresponding columns of the DFT matrix $\boldsymbol{U}_{\mathrm{DFE}}$ to construct $\boldsymbol{G}_{\mathrm{A}}$. Therefore, the optimal analog equalizer at the destination is

$$\boldsymbol{G}_{\mathrm{A}} = [\boldsymbol{U}_{\mathrm{DFT}}]_{:,N}, \qquad (18)$$

where $N$ is the number of data streams. In addition, $\boldsymbol{G}_{\mathrm{D,L}}$ equals to the left inverse of $\boldsymbol{G}_{\mathrm{A}}$, i.e.,

$$\boldsymbol{G}_{\mathrm{D,L}} = [\boldsymbol{U}_{\mathrm{DFT}}]_{:,N}^{\mathrm{H}}. \qquad (19)$$

Based on the results given above, a simplified channel estimation MSE matrix is defined as

$$\begin{aligned}&\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X}, \boldsymbol{G}_{\mathrm{A}}, \boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{opt}}, \boldsymbol{G}_{\mathrm{D,R}}^{\mathrm{opt}}) \\ &\quad = \beta_1\left(\boldsymbol{\Psi}_{\mathrm{H}}^{-1} + \alpha_1\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{\Psi}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right)^{-1} \\ &\quad \triangleq \boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X}), \qquad (20)\end{aligned}$$

where $\beta_1$ and $\alpha_1$ are defined respectively as

$$\begin{aligned}\beta_1 &= \mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big) = \mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\big), \\ \alpha_1 &= \frac{\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big)}{\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big)} = \frac{\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\big)}{\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{N}}\big)}. \qquad (21)\end{aligned}$$

Generally speaking, the training optimization for hybrid MIMO systems aims at optimizing a matrix monotonic function with respect to $\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X})$ [31], formulated as

$$\begin{aligned}&\min_{\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X}}\ f\big(\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X})\big) \\ &\ \mathrm{s.t.}\ \mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big) \le P, \quad \boldsymbol{F}_{\mathrm{A}} \in \mathcal{F}, \qquad (22)\end{aligned}$$

where $f(\cdot)$ is a matrix monotonically increasing function with respect to $\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X})$, and $P$ is the maximum transmit power constraint for the training design, while the set $\mathcal{F}$ denotes the analog matrix set with proper dimensions, in which each element of a analog matrix has constant modulus.

In the following, we investigate in depth the two training optimizations with specific performance metrics, namely, the MSE minimization and mutual information maximization.

### *C. Sum MSE Minimization*

First, we focus on sum MSE minimization and the corresponding objective is the sum of the MSE of each channel matrix element. In other words, it is the sum of the diagonal

$$\begin{aligned}\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X}, \boldsymbol{G}_{\mathrm{A}}, \boldsymbol{G}_{\mathrm{D,L}}, \boldsymbol{G}_{\mathrm{D,R}}) &= \mathbb{E}\big\{(\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{Y}_{\mathrm{D}}\boldsymbol{G}_{\mathrm{D,R}} - \boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H})^{\mathrm{H}}(\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{Y}_{\mathrm{D}}\boldsymbol{G}_{\mathrm{D,R}} - \boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H})\big\} \\ &= (\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{G}_{\mathrm{D,R}} - \boldsymbol{I})^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big)(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{G}_{\mathrm{D,R}} - \boldsymbol{I}) \\ &\quad + \boldsymbol{G}_{\mathrm{D,R}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{D,R}}\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\big). \qquad (12)\end{aligned}$$

$$\begin{aligned}&\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}}, \boldsymbol{X}, \boldsymbol{G}_{\mathrm{A}}, \boldsymbol{G}_{\mathrm{D,L}}, \boldsymbol{G}_{\mathrm{D,R}}^{\mathrm{opt}}) \\ &= \boldsymbol{\Psi}_{\mathrm{H}}\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big) - \boldsymbol{\Psi}_{\mathrm{H}}\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big)\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X} \\ &\quad \times \big(\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big)\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X} + \boldsymbol{\Psi}_{\mathrm{N}}\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\big)\big)^{-1}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big) \\ &= \mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big)\left(\boldsymbol{\Psi}_{\mathrm{H}}^{-1} + \frac{\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big)\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{\Psi}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}}{\mathrm{Tr}\big(\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{D,L}}\boldsymbol{G}_{\mathrm{A}}\big)}\right)^{-1}. \qquad (15)\end{aligned}$$

elements of the MSE matrix. For the sum MSE minimization, the training optimization problem is equivalent to

$$\begin{aligned} \min_{\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X}} \ & \mathrm{Tr}\left(\left(\boldsymbol{\Psi}_{\mathrm{H}}^{-1}+\alpha_1\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{\Psi}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right)^{-1}\right) \\ \mathrm{s.t.} \ & \mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)\leq P, \quad \boldsymbol{F}_{\mathrm{A}}\in\mathcal{F}, \end{aligned} \tag{23}$$

where $\boldsymbol{F}_{\mathrm{A}}$ is the analog matrix at the source, which has row full rank. In other words, $\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}$ is full rank.

Defining the following matrix variable

$$\widetilde{\boldsymbol{X}}=\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)^{\frac{1}{2}}\boldsymbol{X}, \tag{24}$$

the optimization problem (23) is equivalent to the optimization problem (25) as shown at the bottom of this page. Given the optimal solution of $\boldsymbol{F}_{\mathrm{A}}$ and based on the following singular value decomposition (SVD)

$$\boldsymbol{F}_{\mathrm{A}}=\boldsymbol{U}_{\boldsymbol{F}_{\mathrm{A}}}\boldsymbol{\Lambda}_{\boldsymbol{F}_{\mathrm{A}}}\boldsymbol{V}_{\boldsymbol{F}_{\mathrm{A}}}^{\mathrm{H}} \quad \text{with } \boldsymbol{\Lambda}_{\boldsymbol{F}_{\mathrm{A}}}\searrow, \tag{26}$$

the optimization problem (25) can be transferred into the optimization problem (27) as shown at the bottom of this page. Where $N_R$ is the number of rows in $\boldsymbol{F}_{\mathrm{A}}$.

In order to derive the optimal solution, we first quote the following inequality [37]

$$\mathrm{Tr}(\boldsymbol{A}+\boldsymbol{B})^{-1}\geq\sum_i 1/(\lambda_{\boldsymbol{A},i}+\lambda_{\boldsymbol{B},N-i+1}), \tag{28}$$

where $\lambda_{\boldsymbol{Z},i}$ denotes the $i$th largest eigenvalue of $\boldsymbol{Z}$. The equality holds when the unitary matrices of the eigenvalue decompositions (EVDs) of $\boldsymbol{A}$ and $\boldsymbol{B}$ satisfy the following relationship

$$\boldsymbol{U}_{\boldsymbol{A}}=\bar{\boldsymbol{U}}_{\boldsymbol{B}}, \tag{29}$$

where the unitary matrices $\boldsymbol{U}_{\boldsymbol{A}}$ and $\bar{\boldsymbol{U}}_{\boldsymbol{B}}$ are defined based on the following EVDs

$$\begin{aligned} \boldsymbol{A}&=\boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{U}_{\boldsymbol{A}}^{\mathrm{H}} \quad \text{with } \boldsymbol{\Lambda}_{\boldsymbol{A}}\searrow, \\ \boldsymbol{B}&=\bar{\boldsymbol{U}}_{\boldsymbol{B}}\bar{\boldsymbol{\Lambda}}_{\boldsymbol{B}}\bar{\boldsymbol{U}}_{\boldsymbol{B}}^{\mathrm{H}} \quad \text{with } \bar{\boldsymbol{\Lambda}}_{\boldsymbol{B}}\nearrow. \end{aligned} \tag{30}$$

From (28) and (27), the following conclusion can be obtained using the matrix-monotonic optimization framework [31].

*Conclusion 1: The optimal* $\widetilde{\boldsymbol{X}}$ *for the optimization problem (25) satisfies the following optimal structure*

$$\widetilde{\boldsymbol{X}}=\boldsymbol{V}_{\boldsymbol{F}_{\mathrm{A}}}\mathrm{diag}\big\{\boldsymbol{U}_{\boldsymbol{F}_{\mathrm{A}}}^{\mathrm{H}}\boldsymbol{U}_{\boldsymbol{\Psi}_{\mathrm{H}}}\boldsymbol{\Lambda}_{\boldsymbol{X}},\boldsymbol{0}\big\}\bar{\boldsymbol{U}}_{\boldsymbol{\Psi}_{\mathrm{N}}}^{\mathrm{H}}, \tag{31}$$

*where* $\boldsymbol{\Lambda}_{\boldsymbol{X}}$ *is a diagonal matrix, while the unitary matrices* $\boldsymbol{U}_{\boldsymbol{\Psi}_{\mathrm{H}}}$ *and* $\bar{\boldsymbol{U}}_{\boldsymbol{\Psi}_{\mathrm{N}}}$ *are defined based on the following EVDs*

$$\begin{aligned} \boldsymbol{\Psi}_{\mathrm{H}}&=\boldsymbol{U}_{\boldsymbol{\Psi}_{\mathrm{H}}}\boldsymbol{\Lambda}_{\boldsymbol{\Psi}_{\mathrm{H}}}\boldsymbol{U}_{\boldsymbol{\Psi}_{\mathrm{H}}}^{\mathrm{H}} \text{ with } \boldsymbol{\Lambda}_{\boldsymbol{\Psi}_{\mathrm{H}}}\searrow, \\ \boldsymbol{\Psi}_{\mathrm{N}}&=\bar{\boldsymbol{U}}_{\boldsymbol{\Psi}_{\mathrm{N}}}\bar{\boldsymbol{\Lambda}}_{\boldsymbol{\Psi}_{\mathrm{N}}}\bar{\boldsymbol{U}}_{\boldsymbol{\Psi}_{\mathrm{N}}}^{\mathrm{H}} \text{ with } \bar{\boldsymbol{\Lambda}}_{\boldsymbol{\Psi}_{\mathrm{N}}}\nearrow. \end{aligned} \tag{32}$$

Using (31) in (25), $\boldsymbol{F}_{\mathrm{A}}$ can be removed, i.e., it does not affect the optimization, and we have the following conclusion.

*Conclusion 2: For the optimization problem (25), the optimal analog matrix* $\boldsymbol{F}_{\mathrm{A}}$ *is an arbitrary row full rank matrix with constant modulus elements.*

Based on the optimal structure given in Conclusion 1, the original sum-MSE minimization problem can be simplified into the following optimization problem

$$\begin{aligned} \min_{\{f_i^2\}} \ & \sum_i \frac{1}{\frac{1}{\lambda_{\boldsymbol{\Psi}_{\mathrm{H}},i}}+\frac{\alpha_1 f_i^2}{\lambda_{\boldsymbol{\Psi}_{\mathrm{N}},i}}} \\ \mathrm{s.t.} \ & \sum_i f_i^2\leq P, \end{aligned} \tag{33}$$

where $f_i=[\boldsymbol{\Lambda}_{\boldsymbol{X}}]_{i,i}$, $\lambda_{\boldsymbol{\Psi}_{\mathrm{H}},i}=[\boldsymbol{\Lambda}_{\boldsymbol{\Psi}_{\mathrm{H}}}]_{i,i}$, and $\lambda_{\boldsymbol{\Psi}_{\mathrm{N}},i}=[\bar{\boldsymbol{\Lambda}}_{\boldsymbol{\Psi}_{\mathrm{N}}}]_{i,i}$. The optimal solution of (33) is a standard water-filling solution given by [38]

$$f_i^2=\left(\sqrt{\frac{\lambda_{\boldsymbol{\Psi}_{\mathrm{N}},i}}{\alpha_1\mu}}-\frac{\lambda_{\boldsymbol{\Psi}_{\mathrm{N}},i}}{\alpha_1\lambda_{\boldsymbol{\Psi}_{\mathrm{H}},i}}\right)^{+}, \tag{34}$$

where $\mu$ is the Lagrange multiplier for the power constraint in (33).

### D. Mutual Information Maximization

The mutual information between the true channel and the estimated channel is another important performance metric for training optimization [14]. Specifically, the mutual information maximization based training optimization for hybrid MIMO communications is equivalent to

$$\begin{aligned} \max_{\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X}} \ & \log\left|\boldsymbol{\Psi}_{\mathrm{H}}^{-1}+\alpha_1\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{\Psi}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right| \\ \mathrm{s.t.} \ & \mathrm{Tr}(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}})\leq P, \quad \boldsymbol{F}_{\mathrm{A}}\in\mathcal{F}. \end{aligned} \tag{35}$$

Similarly, by defining the following auxiliary variable

$$\widetilde{\boldsymbol{X}}=\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)^{\frac{1}{2}}\boldsymbol{X}, \tag{36}$$

the training optimization (35) is equivalent to (37) as shown at the bottom of the next page. Similar to the MSE minimization,

$$\begin{aligned} \min_{\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X}} \ & \mathrm{Tr}\left(\left(\boldsymbol{\Psi}_{\mathrm{H}}^{-1}+\alpha_1\boldsymbol{F}_{\mathrm{A}}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)^{-\frac{1}{2}}\widetilde{\boldsymbol{X}}\boldsymbol{\Psi}_{\mathrm{N}}^{-1}\widetilde{\boldsymbol{X}}^{\mathrm{H}}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)^{-\frac{1}{2}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right)^{-1}\right) \\ \mathrm{s.t.} \ & \mathrm{Tr}\left(\boldsymbol{F}_{\mathrm{A}}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)^{-\frac{1}{2}}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{H}}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)^{-\frac{1}{2}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right)\leq P, \ \boldsymbol{F}_{\mathrm{A}}\in\mathcal{F}. \end{aligned} \tag{25}$$

$$\begin{aligned} \min_{\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X}} \ & \mathrm{Tr}\left(\left(\boldsymbol{\Psi}_{\mathrm{H}}^{-1}+\alpha_1\boldsymbol{U}_{\boldsymbol{F}_{\mathrm{A}}}\left[\boldsymbol{V}_{\boldsymbol{F}_{\mathrm{A}}}^{\mathrm{H}}\widetilde{\boldsymbol{X}}\boldsymbol{\Psi}_{\mathrm{N}}^{-1}\widetilde{\boldsymbol{X}}^{\mathrm{H}}\boldsymbol{V}_{\boldsymbol{F}_{\mathrm{A}}}\right]_{[1:N_R;1:N_R]}\boldsymbol{U}_{\boldsymbol{F}_{\mathrm{A}}}^{\mathrm{H}}\right)^{-1}\right) \\ \mathrm{s.t.} \ & \mathrm{Tr}\left(\left[\boldsymbol{V}_{\boldsymbol{F}_{\mathrm{A}}}^{\mathrm{H}}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{H}}\boldsymbol{V}_{\boldsymbol{F}_{\mathrm{A}}}\right]_{[1:N_R;1:N_R]}\right)\leq P, \quad \boldsymbol{F}_{\mathrm{A}}\in\mathcal{F}, \end{aligned} \tag{27}$$

the optimization (37) can be transferred to (38) as shown at the bottom of this page. Further consider the inequality [37]

$$\log|\boldsymbol{A}+\boldsymbol{B}| \leq \sum\nolimits_i \log(\lambda_{\boldsymbol{A},i}+\lambda_{\boldsymbol{B},N-i+1}). \tag{39}$$

The equality holds when the unitary matrices $\boldsymbol{U}_{\boldsymbol{A}}$ and $\bar{\boldsymbol{U}}_{\boldsymbol{B}}$, defined in (30), satisfy $\boldsymbol{U}_{\boldsymbol{A}}=\bar{\boldsymbol{U}}_{\boldsymbol{B}}$. Based on (38) and (39), we have the following two conclusions using the matrix-monotonic optimization framework [31].

*Conclusion 3: The optimal $\widetilde{\boldsymbol{X}}$ of the problem (37) satisfies the following structure*

$$\widetilde{\boldsymbol{X}}=\boldsymbol{V}_{\boldsymbol{F}_{\mathrm{A}}}\mathrm{diag}\{\boldsymbol{U}_{\boldsymbol{F}_{\mathrm{A}}}^{\mathrm{H}}\boldsymbol{U}_{\boldsymbol{\Psi}_{\mathrm{H}}}\boldsymbol{\Lambda}_{\boldsymbol{X}},\boldsymbol{0}\}\boldsymbol{U}_{\boldsymbol{\Psi}_{\mathrm{N}}}^{\mathrm{H}}. \tag{40}$$

*Conclusion 4: For the optimization problem (37), the optimal analog matrix $\boldsymbol{F}_{\mathrm{A}}$ is an arbitrary row full rank matrix with constant modulus elements.*

Based on Conclusions 1, 2, 3 and 4, it can be stated that for our training optimization conceived for the sum-MSE and for the mutual information metrics, the analog matrices and training sequences have the same optimal structure. However it will be shown later in this section that for the sum-MSE and for the mutual information metrics, the optimal diagonal matrices in Conclusion 1 and 3 are different.

Exploiting the optimal structure given in Conclusion 3, the mutual information maximization problem is transferred into the following optimization problem

$$\begin{aligned}\max_{\{f_i^2\}} \ & \sum_i \log\left(\frac{1}{\lambda_{\boldsymbol{\Psi}_{\mathrm{H}},i}}+\frac{\alpha_1 f_i^2}{\lambda_{\boldsymbol{\Psi}_{\mathrm{N}},i}}\right)\\ \mathrm{s.t.} \ & \sum_i f_i^2 \leq P.\end{aligned} \tag{41}$$

The optimal water-filling solution of (41) is given by [38]

$$f_i^2=\left(\frac{1}{\mu}-\frac{\lambda_{\boldsymbol{\Psi}_{\mathrm{N}},i}}{\alpha_1\lambda_{\boldsymbol{\Psi}_{\mathrm{H}},i}}\right)^+. \tag{42}$$

## IV. Composite Channel Estimation

The idea of composite channel estimation is to estimate $\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H}\boldsymbol{F}_{\mathrm{A}}$ rather than $\boldsymbol{H}$. The advantage is that only a much smaller matrix is to be estimated. For this to work, however, the analog matrices used in channel estimation and data transmission must be the same. Generally, the analog matrices may be different in channel estimation and data transmission. Sometimes, however, the analog matrices are designed based on channel statistics instead of instantaneous CSI [32]–[34]. Therefore, the analog matrices used in data transmission are also available in channel estimation. Consequently, in these cases, composite channel estimation is applicable, and the premise of this kind of channel estimation is how to choose the analog matrices $\boldsymbol{G}_{\mathrm{A}}$ and $\boldsymbol{F}_{\mathrm{A}}$.

### A. Analog Matrix Optimization

In data transmission phase, the signal model is

$$\widetilde{\boldsymbol{y}}=\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{D}}\boldsymbol{x}+\boldsymbol{G}_{\mathrm{A}}\boldsymbol{n}, \tag{43}$$

based on which the optimization problem of average capacity maximization is given in (44), shown at the bottom of this page, where $P_{\mathrm{T}}$ is the maximum transmit power in data transmission. It is worth highlighting that different from the channel estimation procedure, in data transmission, $\boldsymbol{F}_{\mathrm{A}}$ is column full rank and $\boldsymbol{G}_{\mathrm{A}}$ is row full rank. Based on (2) and defining

$$\widetilde{\boldsymbol{F}}_{\mathrm{D}}=\left(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\right)^{\frac{1}{2}}\boldsymbol{F}_{\mathrm{D}}, \tag{45}$$

the average capacity maximization (44) can be rewritten as (46), shown at the bottom of this page.

As proved in Appendix the optimal $\boldsymbol{G}_{\mathrm{A}}$ is the optimal solution of the following optimization problem

$$\max_{\boldsymbol{G}_{\mathrm{A}}} \boldsymbol{\lambda}\left(\boldsymbol{\Sigma}_{\mathrm{H}}^{\frac{1}{2}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\left(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\right)^{-1}\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{H}}^{\frac{1}{2}}\right), \tag{47}$$

where $\boldsymbol{\lambda}(\boldsymbol{Z})=\left[\lambda_{\boldsymbol{Z},1}\cdots\lambda_{\boldsymbol{Z},N}\right]^{\mathrm{T}}$, which is equivalent to the following optimization problem

$$\max_{\boldsymbol{G}_{\mathrm{A}}} \boldsymbol{\lambda}\left(\left(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\right)^{-1/2}\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\left(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\right)^{-1/2}\right). \tag{48}$$

$$\begin{aligned}\max_{\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X}} \ & \log\left|\boldsymbol{\Psi}_{\mathrm{H}}^{-1}+\alpha_1\boldsymbol{F}_{\mathrm{A}}\left(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right)^{-\frac{1}{2}}\widetilde{\boldsymbol{X}}\boldsymbol{\Psi}_{\mathrm{N}}^{-1}\widetilde{\boldsymbol{X}}^{\mathrm{H}}\left(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right)^{-\frac{1}{2}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right|\\ \mathrm{s.t.} \ & \mathrm{Tr}\left(\boldsymbol{F}_{\mathrm{A}}\left(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right)^{-\frac{1}{2}}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{H}}\left(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right)^{-\frac{1}{2}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right)\leq P, \quad \boldsymbol{F}_{\mathrm{A}}\in\mathcal{F}.\end{aligned} \tag{37}$$

$$\begin{aligned}\max_{\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X}} \ & \log\left|\boldsymbol{\Psi}_{\mathrm{H}}^{-1}+\alpha_1\boldsymbol{U}_{\boldsymbol{F}_{\mathrm{A}}}\left[\boldsymbol{V}_{\boldsymbol{F}_{\mathrm{A}}}^{\mathrm{H}}\widetilde{\boldsymbol{X}}\boldsymbol{\Psi}_{\mathrm{N}}^{-1}\widetilde{\boldsymbol{X}}^{\mathrm{H}}\boldsymbol{V}_{\boldsymbol{F}_{\mathrm{A}}}\right]_{1:N_R,1:N_R}\boldsymbol{U}_{\boldsymbol{F}_{\mathrm{A}}}^{\mathrm{H}}\right|\\ \mathrm{s.t.} \ & \mathrm{Tr}\left(\left[\boldsymbol{V}_{\boldsymbol{F}_{\mathrm{A}}}^{\mathrm{H}}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{H}}\boldsymbol{V}_{\boldsymbol{F}_{\mathrm{A}}}\right]_{1:N_R,1:N_R}\right)\leq P, \quad \boldsymbol{F}_{\mathrm{A}}\in\mathcal{F}.\end{aligned} \tag{38}$$

$$\begin{aligned}\max_{\boldsymbol{G}_{\mathrm{A}},\boldsymbol{F}_{\mathrm{A}},\boldsymbol{F}_{\mathrm{D}}} \ & \mathbb{E}\left\{\log\left|\boldsymbol{F}_{\mathrm{D}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\left(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\right)^{-1}\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{D}}+\boldsymbol{I}\right|\right\}\\ \mathrm{s.t.} \ & \mathrm{Tr}\left(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{F}_{\mathrm{D}}\boldsymbol{F}_{\mathrm{D}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\right)\leq P_{\mathrm{T}}.\end{aligned} \tag{44}$$

$$\begin{aligned}\max_{\boldsymbol{G}_{\mathrm{A}},\boldsymbol{F}_{\mathrm{A}},\widetilde{\boldsymbol{F}}_{\mathrm{D}}} \ & \mathbb{E}\{\log|\widetilde{\boldsymbol{F}}_{\mathrm{D}}^{\mathrm{H}}\left(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\right)^{-\frac{1}{2}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}^{\frac{1}{2}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{H}}^{\frac{1}{2}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\left(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\right)^{-1}\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{H}}^{\frac{1}{2}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}_{\mathrm{H}}^{\frac{1}{2}}\boldsymbol{F}_{\mathrm{A}}\left(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\right)^{-\frac{1}{2}}\widetilde{\boldsymbol{F}}_{\mathrm{D}}+\boldsymbol{I}|\}\\ \mathrm{s.t.} \ & \mathrm{Tr}\left(\widetilde{\boldsymbol{F}}_{\mathrm{D}}\widetilde{\boldsymbol{F}}_{\mathrm{D}}^{\mathrm{H}}\right)\leq P_{\mathrm{T}}.\end{aligned} \tag{46}$$

Moreover, the optimization problem (48) can further be rewritten as (49) shown at the bottom of this page. It is worth noting that the nonzero singular values of the matrix term $\boldsymbol{\Sigma}_{\mathrm{N}}^{\frac{1}{2}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)^{-1/2}$ are all ones. Then, based on the SVD it may be concluded that in (49) the term $\boldsymbol{\Sigma}_{\mathrm{N}}^{\frac{1}{2}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)^{-1/2}$ aims for selecting the first $N$ maximum eigenvalues of the inner matrix term $\boldsymbol{\Sigma}_{\mathrm{N}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{N}}^{-\frac{1}{2}}$. The solution of (48) is computed by solving the problem

$$\begin{aligned}\min_{\boldsymbol{\Lambda},\boldsymbol{Q},\boldsymbol{G}_{\mathrm{A}}} \quad & \big\|[\boldsymbol{U}_{\mathrm{NHN}}]_{:,N_r}\boldsymbol{\Lambda}\boldsymbol{Q}-\boldsymbol{\Sigma}_{\mathrm{N}}^{\frac{1}{2}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big\|_F^2\\ \text{s.t.}\quad & \boldsymbol{Q}\boldsymbol{Q}^{\mathrm{H}}=\boldsymbol{I},\quad \boldsymbol{G}_{\mathrm{A}}\in\mathcal{F},\end{aligned} \tag{50}$$

which can be transferred into

$$\begin{aligned}\min_{\boldsymbol{\Lambda},\boldsymbol{Q},\boldsymbol{G}_{\mathrm{A}}} \quad & \big\|\boldsymbol{\Sigma}_{\mathrm{N}}^{-\frac{1}{2}}[\boldsymbol{U}_{\mathrm{NHN}}]_{:,N_r}\boldsymbol{\Lambda}\boldsymbol{Q}-\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big\|_F^2\\ \text{s.t.}\quad & \boldsymbol{Q}\boldsymbol{Q}^{\mathrm{H}}=\boldsymbol{I},\quad \boldsymbol{G}_{\mathrm{A}}\in\mathcal{F},\end{aligned} \tag{51}$$

where the unitary matrix $\boldsymbol{U}_{\mathrm{NHN}}$ is defined in the EVD:

$$\boldsymbol{\Sigma}_{\mathrm{N}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{N}}^{-\frac{1}{2}}=\boldsymbol{U}_{\mathrm{NHN}}\boldsymbol{\Lambda}_{\mathrm{NHN}}\boldsymbol{U}_{\mathrm{NHN}}^{\mathrm{H}}. \tag{52}$$

The problem (51) can be solved by alternatively optimizing $\boldsymbol{\Lambda}$, $\boldsymbol{Q}$ and $\boldsymbol{G}_{\mathrm{A}}$. When $\boldsymbol{Q}$ and $\boldsymbol{G}_{\mathrm{A}}$ are fixed, the cost function in (51) is quadratic in $\boldsymbol{\Lambda}$, and the optimal $\boldsymbol{\Lambda}$ is derived in closed-form by setting the derivative of the cost function to zero. When $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}$ are fixed, the optimal $\boldsymbol{G}_{\mathrm{A}}$ is given by

$$\big[\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big]_{i,j}=e^{\jmath\arg\left(\left[\boldsymbol{\Sigma}_{\mathrm{N}}^{-\frac{1}{2}}[\boldsymbol{U}_{\mathrm{NHN}}]_{:,N_r}\boldsymbol{\Lambda}\boldsymbol{Q}\right]_{i,j}\right)}. \tag{53}$$

When $\boldsymbol{G}_{\mathrm{A}}$ and $\boldsymbol{\Lambda}$ are fixed, the optimal unitary matrix $\boldsymbol{Q}$ is

$$\boldsymbol{Q}=\boldsymbol{U}\boldsymbol{V}^{\mathrm{H}}, \tag{54}$$

with $\boldsymbol{V}$ and $\boldsymbol{U}$ defined by the following SVD

$$\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}^{-\frac{1}{2}}[\boldsymbol{U}_{\mathrm{NHN}}]_{:,N_r}\boldsymbol{\Lambda}=\boldsymbol{V}\boldsymbol{\Lambda}_3\boldsymbol{U}^{\mathrm{H}}\quad\text{with }\boldsymbol{\Lambda}_3\searrow. \tag{55}$$

As proved in Appendix the analog matrix $\boldsymbol{F}_{\mathrm{A}}$ of (46) is the solution of the following optimization problem

$$\max_{\boldsymbol{F}_{\mathrm{A}}}\ \boldsymbol{\lambda}\Big(\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-\frac{1}{2}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-\frac{1}{2}}\Big). \tag{56}$$

The optimal $\boldsymbol{F}_{\mathrm{A}}$ can be computed by solving the following optimization problem

$$\begin{aligned}\min_{\boldsymbol{\Lambda},\boldsymbol{Q},\boldsymbol{F}_{\mathrm{A}}} \quad & \big\|[\boldsymbol{U}_{\boldsymbol{\Psi}_{\mathrm{H}}}]_{:,N_r}\boldsymbol{\Lambda}\boldsymbol{Q}-\boldsymbol{F}_{\mathrm{A}}\big\|_F^2\\ \text{s.t.}\quad & \boldsymbol{Q}\boldsymbol{Q}^{\mathrm{H}}=\boldsymbol{I},\quad \boldsymbol{F}_{\mathrm{A}}\in\mathcal{F},\end{aligned} \tag{57}$$

where $\boldsymbol{U}_{\boldsymbol{\Psi}_{\mathrm{H}}}$ is defined in the EVD in (32). Similarly, an iterative optimization can be used to solve the optimization problem (57).

Since the objective function in (57) is quadratic in $\boldsymbol{\Lambda}$, the optimal $\boldsymbol{\Lambda}$ is readily derived in closed-form when $\boldsymbol{Q}$ and $\boldsymbol{F}_{\mathrm{A}}$ are fixed. When $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}$ are fixed, the optimal $\boldsymbol{F}_{\mathrm{A}}$ is

$$[\boldsymbol{F}_{\mathrm{A}}]_{i,j}=e^{\jmath\arg\left(\left[[\boldsymbol{U}_{\boldsymbol{\Psi}_{\mathrm{H}}}]_{:,N_r}\boldsymbol{\Lambda}\boldsymbol{Q}\right]_{i,j}\right)}. \tag{58}$$

When $\boldsymbol{F}_{\mathrm{A}}$ and $\boldsymbol{\Lambda}$ are fixed, the optimal unitary matrix $\boldsymbol{Q}$ is $\boldsymbol{Q}=\boldsymbol{U}\boldsymbol{V}^{\mathrm{H}}$, with $\boldsymbol{V}$ and $\boldsymbol{U}$ defined in the following SVD

$$\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big[\boldsymbol{U}_{\boldsymbol{\Psi}_{\mathrm{H}}}\big]_{:,N_t}\boldsymbol{\Lambda}=\boldsymbol{V}\boldsymbol{\Lambda}_4\boldsymbol{U}^{\mathrm{H}}\quad\text{with }\boldsymbol{\Lambda}_4\searrow. \tag{59}$$

### *B. Sum MSE Minimization Based Training Optimization*

Given the known analog matrices $\boldsymbol{G}_{\mathrm{A}}$ and $\boldsymbol{F}_{\mathrm{A}}$, the signal model in channel estimation is given by

$$\widetilde{\boldsymbol{Y}}=\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}+\boldsymbol{G}_{\mathrm{A}}\boldsymbol{N}, \tag{60}$$

where $\boldsymbol{G}_{\mathrm{A}}$ and $\boldsymbol{F}_{\mathrm{A}}$ are both full rank squared matrices. Based on (60), we estimate the composite channel matrix $\boldsymbol{\mathcal{H}}=\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H}\boldsymbol{F}_{\mathrm{A}}$, instead of $\boldsymbol{H}$, as

$$\widehat{\boldsymbol{\mathcal{H}}}=\widetilde{\boldsymbol{Y}}\boldsymbol{G}_{\mathrm{D}}, \tag{61}$$

with $\boldsymbol{G}_{\mathrm{D}}$ as the channel estimator. It is worth recapping that for this scheme to work, the analog matrices in channel estimation must be the same as in data transmission. The corresponding channel estimation MSE matrix is given by

$$\begin{aligned}&\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X},\boldsymbol{G}_{\mathrm{A}},\boldsymbol{G}_{\mathrm{D}})\\ &\quad=\mathbb{E}\big\{\big(\widetilde{\boldsymbol{Y}}\boldsymbol{G}_{\mathrm{D}}-\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H}\boldsymbol{F}_{\mathrm{A}}\big)^{\mathrm{H}}\big(\widetilde{\boldsymbol{Y}}\boldsymbol{G}_{\mathrm{D}}-\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H}\boldsymbol{F}_{\mathrm{A}}\big)\big\}\\ &\quad=\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}(\boldsymbol{X}\boldsymbol{G}_{\mathrm{D}}-\boldsymbol{I})(\boldsymbol{X}\boldsymbol{G}_{\mathrm{D}}-\boldsymbol{I})^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\big)\\ &\qquad+\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{D}}\boldsymbol{G}_{\mathrm{D}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{N}}\big),\end{aligned} \tag{62}$$

based on which the sum MSE can be expressed as

$$\begin{aligned}&\mathrm{Tr}\big(\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X},\boldsymbol{G}_{\mathrm{A}},\boldsymbol{G}_{\mathrm{D}})\big)\\ &\quad=\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)\mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}(\boldsymbol{X}\boldsymbol{G}_{\mathrm{D}}-\boldsymbol{I})(\boldsymbol{X}\boldsymbol{G}_{\mathrm{D}}-\boldsymbol{I})^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\big)\\ &\qquad+\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{D}}\boldsymbol{G}_{\mathrm{D}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{N}}\big).\end{aligned} \tag{63}$$

The term $\mathrm{Tr}\big(\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X},\boldsymbol{G}_{\mathrm{A}},\boldsymbol{G}_{\mathrm{D}})\big)$ is a quadratic function for $\boldsymbol{G}_{\mathrm{D}}$. Thus the optimal $\boldsymbol{G}_{\mathrm{D}}$ satisfies the following equality

$$\begin{aligned}&\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{G}_{\mathrm{D}}\\ &\quad+\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)\boldsymbol{\Psi}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{D}}\\ &\quad=\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}},\end{aligned} \tag{64}$$

and it can be derived as

$$\begin{aligned}\boldsymbol{G}_{\mathrm{D}}=&\Big(\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\\ &\quad+\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)\boldsymbol{\Psi}_{\mathrm{N}}\Big)^{-1}\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)\\ &\quad\times\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}.\end{aligned} \tag{65}$$

With the computed analog matrices $\boldsymbol{G}_{\mathrm{A}}$ and $\boldsymbol{F}_{\mathrm{A}}$, in the following we will investigate the corresponding training optimizations for different performance metrics and different available channel statistics.

*1) General Case of $\Psi_N \not\propto I$:* When $\boldsymbol{G}_{\mathrm{D}}$ is fixed, the sum MSE minimization based training optimization becomes

$$\begin{aligned}\min_{\boldsymbol{X}}\quad & \mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}(\boldsymbol{X}\boldsymbol{G}_{\mathrm{D}}-\boldsymbol{I})(\boldsymbol{X}\boldsymbol{G}_{\mathrm{D}}-\boldsymbol{I})^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\big)\\ \text{s.t.}\quad & \mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)\le P.\end{aligned} \tag{66}$$

$$\max_{\boldsymbol{G}_{\mathrm{A}}}\ \boldsymbol{\lambda}\Big(\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)^{-1/2}\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\mathrm{N}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{N}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathrm{N}}^{\frac{1}{2}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)^{-1/2}\Big). \tag{49}$$

The corresponding Lagrange function is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{X},\mu) = \mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}(\boldsymbol{X}\boldsymbol{G}_{\mathrm{D}}-\boldsymbol{I})(\boldsymbol{X}\boldsymbol{G}_{\mathrm{D}}-\boldsymbol{I})^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\big)\\ +\mu\Big(\mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)-P\Big),\end{aligned} \tag{67}$$

based on which the corresponding Karush-Kuhn-Tucker (KKT) conditions are given by [38]

$$\begin{cases}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{G}_{\mathrm{D}}\boldsymbol{G}_{\mathrm{D}}^{\mathrm{H}}+\mu\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}=\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{G}_{\mathrm{D}}^{\mathrm{H}},\\ \mu\Big(\mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)-P\Big)=0,\\ \mu\geq 0,\quad \mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)\leq P.\end{cases} \tag{68}$$

Based on the KKT conditions, the optimal training sequence $\boldsymbol{X}$ satisfies the following equality

$$\begin{aligned}\mathrm{vec}(\boldsymbol{X}) = \Big(\big(\boldsymbol{G}_{\mathrm{D}}\boldsymbol{G}_{\mathrm{D}}^{\mathrm{H}}\big)^{\mathrm{T}}\otimes\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}+\mu\boldsymbol{I}\otimes\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\Big)^{-1}\\ \times\,\mathrm{vec}\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\boldsymbol{G}_{\mathrm{D}}^{\mathrm{H}}\big).\end{aligned} \tag{69}$$

As the optimization problem (66) is convex for $\boldsymbol{X}$, the KKT conditions (68) are the necessary and sufficient conditions for the optimal solutions. Then $\mu$ can be computed using one dimension search, e.g., bisection search, to guarantee that the KKT conditions hold [38].

*2) Special Case of $\Psi_N = \alpha_N I$:* Substituting $\boldsymbol{G}_{\mathrm{D}}$ of (65) into (63) yields

$$\begin{aligned}&\mathrm{Tr}\big(\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X},\boldsymbol{G}_{\mathrm{A}},\boldsymbol{G}_{\mathrm{D}})\big)\\ &\quad=\mathrm{Tr}\left(\left(\frac{\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-1}}{\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}\big)}+\frac{\boldsymbol{X}\boldsymbol{\Psi}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}}{\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}\big)}\right)^{-1}\right),\end{aligned} \tag{70}$$

where for composite channel estimation, both $\boldsymbol{G}_{\mathrm{A}}$ and $\boldsymbol{F}_{\mathrm{A}}$ are full rank squared matrices. Similar to the previous discussion, based on the MSE matrix formulation, the optimization for the hybrid MIMO system is given by

$$\begin{aligned}&\min_{\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X}}\ \mathrm{Tr}\left(\left(\frac{\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-1}}{\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}\big)}+\frac{\boldsymbol{X}\boldsymbol{\Psi}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}}{\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}\big)}\right)^{-1}\right)\\ &\ \mathrm{s.t.}\ \mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)\leq P.\end{aligned} \tag{71}$$

Similarly, by defining the following auxiliary variable $\widetilde{\boldsymbol{X}}$

$$\widetilde{\boldsymbol{X}}=\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{\frac{1}{2}}\boldsymbol{X}, \tag{72}$$

the optimization problem (71) is rewritten as (73), shown at the bottom of this page, where

$$\alpha_2=\mathrm{Tr}(\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}})/[\alpha_{\mathrm{N}}\mathrm{Tr}(\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}})]. \tag{74}$$

To transfer this optimization problem into a convex one that can be efficiently solved, we introduce a new variable $\boldsymbol{Q}$ as

$$\boldsymbol{Q}=\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{H}}. \tag{75}$$

Since $\widetilde{\boldsymbol{X}}$ is a fat matrix, to guarantee that the channel matrix can be estimated, $\boldsymbol{Q}$ is a full rank matrix. Thus, in the optimization we do not need to consider the rank constraint on $\boldsymbol{Q}$. With the definition of $\boldsymbol{Q}$, (73) is equivalent to (76), shown at the bottom of this page [38], where $\boldsymbol{M}$ is a positive semidefinite matrix. Then using Schur-complement, the first constraint can be replaced by the following linear matrix inequality equivalently,

$$\begin{bmatrix}\boldsymbol{M} & \boldsymbol{I}\\ \boldsymbol{I} & \big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-1}+\alpha_2\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-\frac{1}{2}}\boldsymbol{Q}\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-\frac{1}{2}}\end{bmatrix}\succeq\boldsymbol{0}. \tag{77}$$

As a result, the optimization (73) can be transferred into the standard semidefinite programming (SDP) problem given as (78), shown at the bottom of this page, which can be solved efficiently using for example the CVX software toolbox [40].

### *C. Mutual Information Maximization Based Training Optimization*

For the composite channel estimation, the training optimization problem based on mutual information maximization can be formulated in the following form

$$\begin{aligned}&\max_{\boldsymbol{X},\boldsymbol{G}_{\mathrm{D}}}\ -\log\left|\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{F}_{\mathrm{A}},\boldsymbol{X},\boldsymbol{G}_{\mathrm{A}},\boldsymbol{G}_{\mathrm{D}})\right|\\ &\ \mathrm{s.t.}\ \mathrm{Tr}\big(\boldsymbol{F}_{\mathrm{A}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\big)\leq P.\end{aligned} \tag{79}$$

$$\begin{aligned}&\min_{\boldsymbol{F}_{\mathrm{A}},\widetilde{\boldsymbol{X}}}\ \mathrm{Tr}\left(\left(\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-1}+\alpha_2\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-\frac{1}{2}}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{H}}\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-\frac{1}{2}}\right)^{-1}\right)\\ &\ \mathrm{s.t.}\ \mathrm{Tr}\big(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{H}}\big)\leq P.\end{aligned} \tag{73}$$

$$\begin{aligned}&\min_{\boldsymbol{M},\boldsymbol{Q}}\ \mathrm{Tr}(\boldsymbol{M})\\ &\ \mathrm{s.t.}\ \boldsymbol{M}\succeq\left(\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-1}+\alpha_2\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-\frac{1}{2}}\boldsymbol{Q}\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-\frac{1}{2}}\right)^{-1}\\ &\qquad \boldsymbol{M}\succeq\boldsymbol{0},\quad \boldsymbol{Q}\succeq\boldsymbol{0},\ \mathrm{Tr}(\boldsymbol{Q})\leq P.\end{aligned} \tag{76}$$

$$\begin{aligned}&\min_{\boldsymbol{M},\boldsymbol{Q}}\ \mathrm{Tr}(\boldsymbol{M})\\ &\ \mathrm{s.t.}\ \begin{bmatrix}\boldsymbol{M} & \boldsymbol{I}\\ \boldsymbol{I} & \big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{\Psi}_{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-1}+\alpha_2\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-\frac{1}{2}}\boldsymbol{Q}\big(\boldsymbol{F}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{A}}\big)^{-\frac{1}{2}}\end{bmatrix}\succeq\boldsymbol{0},\quad \boldsymbol{Q}\succeq\boldsymbol{0}.\end{aligned} \tag{78}$$

Substituting (65) into (62) and using the resultant (62) in (79), we have the following equivalent optimization problem

$$\begin{aligned}&\max_{\boldsymbol{X}} \ \log\left|\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{\Psi}_{\rm H}\boldsymbol{F}_{\rm A}\right)^{-1}+\alpha_2\boldsymbol{X}\boldsymbol{X}^{\rm H}\right| \\ &\text{s.t. } {\rm Tr}\left(\boldsymbol{F}_{\rm A}\boldsymbol{X}\boldsymbol{X}^{\rm H}\boldsymbol{F}_{\rm A}^{\rm H}\right)\leq P. \end{aligned} \tag{80}$$

In contrast to the sum MSE minimization, the mutual information maximization in nature minimizes the geometric mean of the diagonal elements of the channel estimation MSE matrix.

Similarly after defining $\widetilde{\boldsymbol{X}}=\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{F}_{\rm A}\right)^{\frac{1}{2}}\boldsymbol{X}$, we have

$$\begin{aligned}&\max_{\widetilde{\boldsymbol{X}}} \ \log\Big|\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{\Psi}_{\rm H}\boldsymbol{F}_{\rm A}\right)^{-1} \\ &\qquad +\alpha_2\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{F}_{\rm A}\right)^{-\frac{1}{2}}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\rm H}\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{F}_{\rm A}\right)^{-\frac{1}{2}}\Big| \\ &\text{s.t. } {\rm Tr}\left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\rm H}\right)\leq P, \end{aligned} \tag{81}$$

which is equivalent to the following optimization problem

$$\begin{aligned}&\max_{\widetilde{\boldsymbol{X}}} \ \log\left|\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{F}_{\rm A}\right)^{\frac{1}{2}}\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{\Psi}_{\rm H}\boldsymbol{F}_{\rm A}\right)^{-1}\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{F}_{\rm A}\right)^{\frac{1}{2}}+\alpha_2\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\rm H}\right| \\ &\text{s.t. } {\rm Tr}\left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\rm H}\right)\leq P. \end{aligned} \tag{82}$$

It is obvious that the optimization (82) has closed-form optimal solutions. Specifically, based on the inequality (39), the optimal $\widetilde{\boldsymbol{X}}$ satisfies

$$\widetilde{\boldsymbol{X}}=\boldsymbol{Q}_{\boldsymbol{F}}\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{X}}}\boldsymbol{U}_{\rm Arb}^{\rm H}, \tag{83}$$

where $\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{X}}}$ is a diagonal matrix, and $\boldsymbol{U}_{\rm Arb}$ is an arbitrary unitary matrix with proper dimension, while the unitary matrix $\boldsymbol{Q}_{\boldsymbol{F}}$ is defined based on the following EVD

$$\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{F}_{\rm A}\right)^{\frac{1}{2}}\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{\Psi}_{\rm H}\boldsymbol{F}_{\rm A}\right)^{-1}\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{F}_{\rm A}\right)^{\frac{1}{2}}=\boldsymbol{Q}_{\boldsymbol{F}}\boldsymbol{\Lambda}_{\boldsymbol{F}\boldsymbol{\Psi}\boldsymbol{F}}\boldsymbol{Q}_{\boldsymbol{F}}^{\rm H}$$

with

$$\boldsymbol{\Lambda}_{\boldsymbol{F}\boldsymbol{\Psi}\boldsymbol{F}}\searrow. \tag{84}$$

Based on the optimal structure (83), the original optimization problem is simplified into the following one

$$\begin{aligned}&\max_{\{x_i^2\}} \ \sum_i \log(\lambda_i+\alpha_2 x_i^2) \\ &\text{s.t. } \sum_i x_i^2\leq P. \end{aligned} \tag{85}$$

whose optimal solution is the water-filling solution [38]

$$x_i^2=(1/\mu-\lambda_i/\alpha_2)^+ \tag{86}$$

where $\lambda_i=[\boldsymbol{\Lambda}_{\boldsymbol{F}\boldsymbol{\Psi}\boldsymbol{F}}]_{i,i}$.

### D. Complexity Analysis

Firstly, it may be readily observed that both benchmarks have the same complexity order as the proposed LS estimator of Section III-A, since the same DFT based training sequence is adopted, which is dominated by the matrix inversion operation having a complexity order of $\mathcal{O}(N_t^3+N_r^3)$, where $N_t$ and $N_r$ denote the numbers of antennas at the transmitter and the receiver, respectively. However, in practice, both benchmarks have more complex signal processing operations than the LS estimator, since they need either partial or complete channel statistics for channel estimation. Additionally, considering that both the matrix inversion and eigenvalue decomposition (EVD) operations are involved in deriving the optimal training sequence for the proposed MMSE estimator of Section III-B, the complexity order is also $\mathcal{O}(N_t^3+N_r^3)$, whilst achieving a better MSE performance than the proposed LS estimator. In Section IV, where the composite channel estimation having drastically reduced dimensions of $N_{RF}\times N_{RF}$ $(N_{RF}\ll\min(N_t,N_r))$ is studied, the complexity order of the proposed MMSE-ITE estimator detailed in Section IV-B1 is $\mathcal{O}[I(N_{RF}L)^3]$, where $N_{RF}$ and $L$ denote the number of RF chains and the training sequence length, respectively, while $I$ represents the number of iterations required for solving the problem (66). Moreover, the proposed MMSE estimator of Section IV-B2 is obtained by solving the SDP problem (78), whose complexity order is $\mathcal{O}(N_t^{3.5})\log(1/\epsilon)$ with $\epsilon$ being the precision factor [A4]. Note that the above complexity analysis of all the proposed channel estimators is also applicable to the associated mutual information maximization counterparts.

## V. Simulation Results

In this section, without loss of generality the widely used Raleigh fading channel model is adopted. We numerically evaluate the normalized MSE with respect to the effective channel statistics ${\rm Tr}(\boldsymbol{R}_{\rm H})$ and the mutual information (MI) performance of the LS estimator in Section III-A, the MMSE, i.e., LMMSE, estimator in Section III-B and the MMSE, i.e., LMMSE, estimator in Section IV, respectively, under a wide range of SNRs, antenna setups, the number of RF chains, and different channel correlations. Note that the channel covariance matrix $\boldsymbol{R}_{\rm H}$ is different for different channel estimators. Specifically, for the LS estimator of Section III-A, $\boldsymbol{R}_{\rm H}=\mathbb{E}\{\boldsymbol{H}\boldsymbol{H}^{\rm H}\}={\rm Tr}(\boldsymbol{\Psi}_{\rm H})\boldsymbol{\Sigma}_{\rm H}$. For the MMSE estimator of Section III-B, $\boldsymbol{R}_{\rm H}=\mathbb{E}\{\boldsymbol{G}_{\rm A}\boldsymbol{H}\boldsymbol{H}^{\rm H}\boldsymbol{G}_{\rm A}^{\rm H}\}={\rm Tr}(\boldsymbol{\Psi}_{\rm H})\boldsymbol{G}_{\rm A}\boldsymbol{\Sigma}_{\rm H}\boldsymbol{G}_{A}^{\rm H}$, while for the MMSE estimator of Section IV, $\boldsymbol{R}_{\rm H}=\mathbb{E}\{\boldsymbol{G}_{\rm A}\boldsymbol{H}\boldsymbol{F}_{\rm A}\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{H}^{\rm H}\boldsymbol{G}_{\rm A}^{\rm H}\}={\rm Tr}\left(\boldsymbol{F}_{\rm A}^{\rm H}\boldsymbol{\Psi}_{\rm H}\boldsymbol{F}_{\rm A}\right)\boldsymbol{G}_{\rm A}\boldsymbol{\Sigma}_{\rm H}\boldsymbol{G}_{\rm A}^{\rm H}$. Moreover, the scaled LS (SLS) estimator [14] and the MMSE estimator with orthogonal probing based on DFT (MMSE-DFT) [13], [14] are adopted as two comparisons. Note that except the LS estimator, all the other estimators require the second-order statistics of the channel and noise at the receiver. Furthermore, for the optimal training design, the eigenspace of the channel and noise second-order statistics is assumed to be known at the transmitter. Unfortunately, these two requirements impose large computational cost and high feedback overhead.

Unless otherwise stated, the numbers of transmit and receive antennas are set to $N_t=32$ and $N_r=16$, respectively, and the number of RF chains at both the transmitter and receiver is $N_{RF}=4$. Moreover, the training length is set to $L=N_t=32$ for both the LS estimator of Section III-A and the MMSE estimator of Section III-B, and $L=N_{RF}=4$ for the MMSE estimator of Section IV, respectively. The received noise power is assumed to be unity and hence the training SNR is defined as ${\rm SNR}=P$. We also utilize the exponential model to construct the correlation matrices of the channel and noise as follows

$$\begin{cases}\left[\boldsymbol{\Sigma}_{\rm H}\right]_{n_1,m_1}=|\zeta_t|^{|n_1-m_1|}, & \left[\boldsymbol{\Sigma}_{\rm N}\right]_{n_1,m_1}=|\zeta_r|^{|n_1-m_1|}, \\ \left[\boldsymbol{\Psi}_{\rm H}\right]_{n_2,m_2}=|\epsilon_t|^{|n_2-m_2|}, & \left[\boldsymbol{\Psi}_{\rm N}\right]_{n_2,m_2}=|\epsilon_r|^{|n_2-m_2|}, \\ |\zeta_t|\leq 1, \quad |\zeta_r|\leq 1, \ |\epsilon_t|\leq 1, \ |\epsilon_r|\leq 1, \end{cases} \tag{87}$$

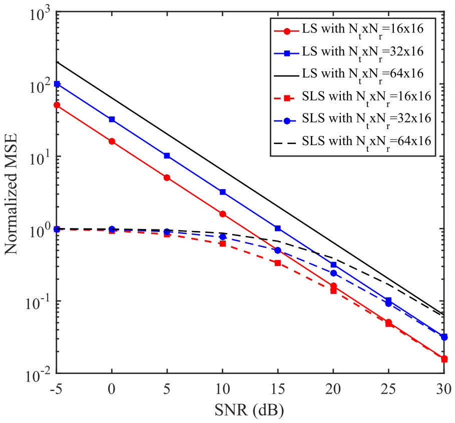

Fig. 2. Normalized MSE performance of the proposed LS and benchmark SLS estimators as functions of SNR under different antenna setups of $N_t \times N_r = 16 \times 16$, $32 \times 16$ and $64 \times 16$.

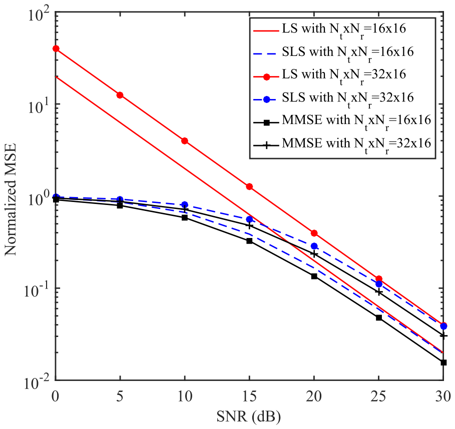

Fig. 3. Normalized MSE performance of the proposed LS, and MMSE of Section III-B as well as benchmark SLS estimators as functions of SNR under different antenna setups of $N_t \times N_r = 16 \times 16$ and $32 \times 16$.

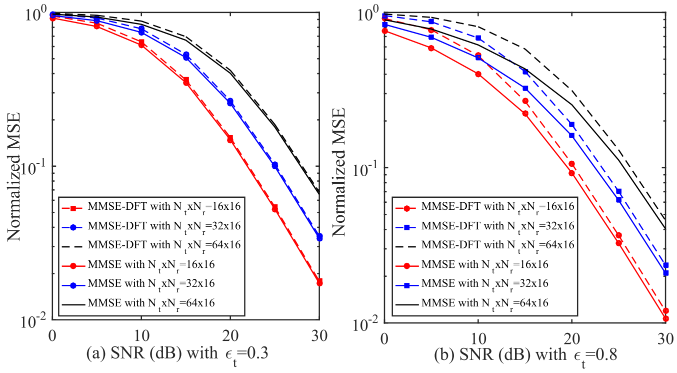

Fig. 4. Normalized MSE performance of the proposed MMSE of Section III-B and benchmark MMSE-DFT estimators as functions of SNR under different antenna setups of $N_t \times N_r = 16 \times 16$, $32 \times 16$ and $64 \times 16$ with (a) weak channel correlation $\epsilon_t = 0.3$, and (b) strong channel correlation $\epsilon_t = 0.8$.

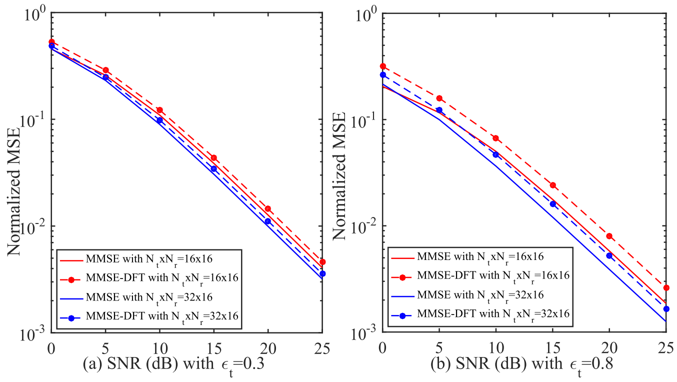

Fig. 5. Normalized MSE performance of the proposed MMSE (of Section IV-B.1) and benchmark MMSE-DFT estimators as functions of SNR under different antenna setups of $N_t \times N_r = 16 \times 16$ and $32 \times 16$ with (a) weak channel correlation $\epsilon_t = 0.3$, and (b) strong channel correlation $\epsilon_t = 0.8$.

where $\zeta_t$, $\zeta_r$, $\epsilon_t$ and $\epsilon_r$ are the correlation coefficients. Unless otherwise stated, we set $\zeta_t = \epsilon_t = 0.5$ and $\zeta_r = \epsilon_r = 0.4$.

### *A. Normalized MSE Performance*

Fig. 2 shows the normalized MSE performance of the LS and SLS estimators as functions of the SNR, where three different antenna setups of $N_t \times N_r = 16 \times 16$, $32 \times 16$ and $64 \times 16$ are considered. As expected, the normalized MSE performance of the both estimators improve as the SNR increases. In addition, the SLS estimator with the aid of the channel and noise second-order statistics naturally performs better than the LS estimator, especially at low-SNR region. However, this performance gain is realized at the expense of significantly higher computational cost.

Fig. 3 illustrates the normalized MSE performance of the LS, SLS, and MMSE of Section III-B estimators as functions of the SNR, under two different antenna setups of $N_t \times N_r = 16 \times 16$ and $32 \times 16$. Observe that for both these two antenna setups, the proposed MMSE estimator achieves better MSE performance than the benchmark SLS, since it directly minimize the channel estimation error by utilizing more channel and noise second-order statistics compared to the SLS estimator. The performance of the proposed LS estimator is the worst, as it does not utilize any seconder-statistics of the channel and noise and imposes the lowest complexity.

Fig. 4 depicts the normalized MSEs as functions of the SNR achieved by the MMSE estimator with optimal training of Section III-B, and the benchmark MMSE-DFT with orthogonal training, under three different antenna setups together with two different channel correlations of $\epsilon_t = 0.3$ and $\epsilon_t = 0.8$, respectively. It can be seen from Fig. 4 (a) that with a weakly correlated channel of $\epsilon_t = 0.3$, the performance of the MMSE-DFT is almost identical to that of our optimal MMSE estimator. This is because the orthogonal training is nearly optimal when the channel is asymptotically uncorrelated. However, observe from Fig. 4 (b) that with $\epsilon_t = 0.8$, the performance gap between the MMSE-DFT benchmark and our optimal MMSE estimator is clearly visible, particularly at low SNR region.

Similarly to Fig. 4, Fig. 5 compares the normalized MSEs of the proposed MMSE estimator of Section IV-B.1 and the MMSE-DFT benchmark by varying SNR, where two antenna setups and two cases of channel correlation are considered. Different from Section III-B, since the dimension of the estimated composite channel $\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H}\boldsymbol{F}_{\mathrm{A}}$ in Section IV does not depend on the number of transmit antennas, increasing $N_t$ actually provides the gain in the received SNR.

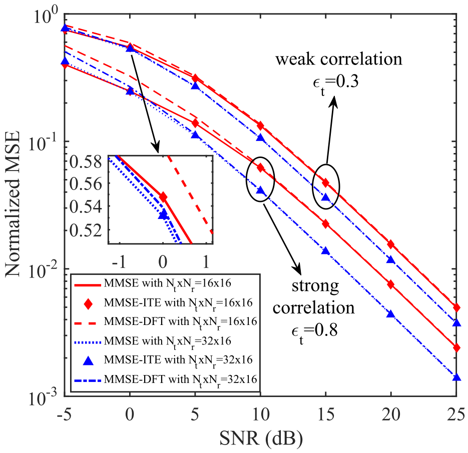

Fig. 6. Normalized MSE performance of the proposed MMSE-ITE estimator of Section IV-B.1 and MMSE estimator of Section IV-B.2 as well as benchmark MMSE-DFT as functions of SNR under different antenna setups of $N_t \times N_r = 16 \times 16$ and $32 \times 16$ as well as different channel correlations of $\epsilon_t = 0.3$, $\epsilon_t = 0.8$ and $\epsilon_r = 0$.

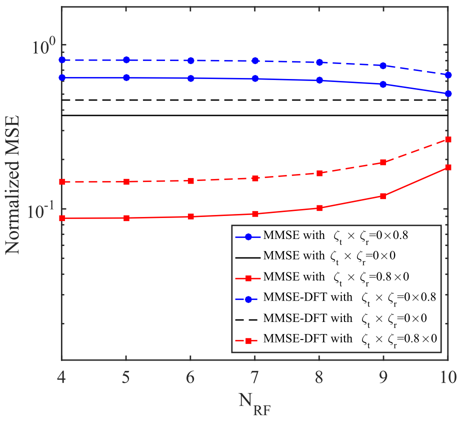

Fig. 7. Normalized MSE performance of the proposed MMSE (of Section III-B) and benchmark MMSE-DFT estimators as functions of the number of RF chains $N_{RF}$ under different channel and noise correlations of $\zeta_t \times \zeta_r = 0 \times 0$, $0 \times 0.8$ and $0.8 \times 0$, given SNR $= 15\,$dB.

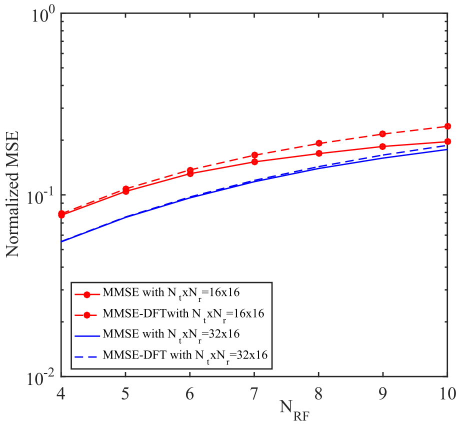

Fig. 8. Normalized MSEs of the MMSE estimator of Section IV-B.1 and benchmark MMSE-DFT versus the number of RF chains under different antenna setups of $N_t \times N_r = 16 \times 16$ and $32 \times 16$ with SNR $= 15\,$dB.

Consequently with the increase of $N_t$, it can be seen from Fig. 5 that the MSEs of the two estimators become better. In addition, the performance gap between the two estimators for the strong channel correlation is larger than that for the weak counterpart.

Fig. 6 further depicts the normalized MSEs achieved by the iterative MMSE (MMSE-ITE) estimator of Section IV-B.1, MMSE estimator of Section IV-B.2 and benchmark MMSE-DFT under the same conditions of Fig. 6 but with $\epsilon_r = 0$. In particular, we find that the proposed MMSE-ITE estimator of Section IV-B.1 can achieve nearly the same performance as the optimal MMSE estimator of Section IV-B.2 for both weak and strong channel correlations. As expected, the two proposed MMSE estimators outperform the MMSE-DFT benchmark, especially for the case of $N_t = 16$ and low SNR.

It is readily observed from Section III-B that the number of RF chains $N_{RF}$ actually has no influence on the proposed MMSE estimator, since the analog matrix $\boldsymbol{G}_A$ is assumed to be column full rank and thus the corresponding sum-MSE minimization in (23) is irrelevant to $N_{RF}$. However, we readily find that by setting $\boldsymbol{G}_{D,R} = \boldsymbol{I}_{N_{RF}}$, the proposed MMSE estimator of Section III-B can also be applied to estimate the effective channel $\boldsymbol{G}_A\boldsymbol{H}$ with much reduced channel dimension $N_{RF} \times N_t$. In this context, we can reexpress $\alpha_1$ in (21) as $\alpha_1 = \mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{H}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)/\mathrm{Tr}\big(\boldsymbol{G}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{N}}\boldsymbol{G}_{\mathrm{A}}^{\mathrm{H}}\big)$. Based on this setting, in Fig. 7, the normalized MSEs of the optimal training based MMSE estimator derived in Section III-B and the benchmark MMSE-DFT are compared by varying $N_{RF}$, given SNR $=$ 15 dB and different channel and noise correlations. It is easily inferred from (21) that the influence of $N_{RF}$ on the MSE is mainly dominated by the parameter $\alpha_1$. We consider three typical cases of $\alpha_1$ as: 1) $\zeta_t = \zeta_r = 0$, 2) $\zeta_t = 0$ and $\zeta_r = 0.8$ and 3) $\zeta_t = 0.8$ and $\zeta_r = 0$. In the first case, $N_{RF}$ has no influence on the achievable MSEs of the two estimators, due to the fixed $\alpha_1 = 1$ when varying $N_{RF}$. In the second case, the performances of the two estimators become better for large $N_{RF}$, since it can be inferred that the value of $\alpha_1$ increases for a large $N_{RF}$. Conversely, in the third case, the MSEs of the two estimators become poorer for large $N_{RF}$, due to the decreased value of $\alpha_1$ corresponding to a large $N_{RF}$. Naturally, the proposed MMSE estimator always outperforms the MMSE-DFT benchmark in terms of MSE performance.

Fig. 8 compares the normalized MSEs of the MMSE estimator of Section IV-B.1 and benchmark MMSE-DFT by varying $N_{RF}$, under two antenna setups and given SNR $= 15\,$dB. Observe from Fig. 8 that as $N_{RF}$ increases, the both estimators have worse performance, due to the increased dimension of the composite channel $\boldsymbol{G}_{\mathrm{A}}\boldsymbol{H}\boldsymbol{F}_{\mathrm{A}}$. In particular, the MMSE-DFT estimator achieves a comparable performance to the MMSE estimator at low $N_{RF}$, since in this case the optimal training with the decreased design freedom is asymptotically orthogonal. Like Fig. 6, increasing $N_t$ also leads to better performance.

## B. Mutual Information Performance

Fig. 9 shows the MI performance of the MMSE estimator of Section III-B and benchmark MMSE-DFT versus the SNR, under three different antenna setups and two cases of channel

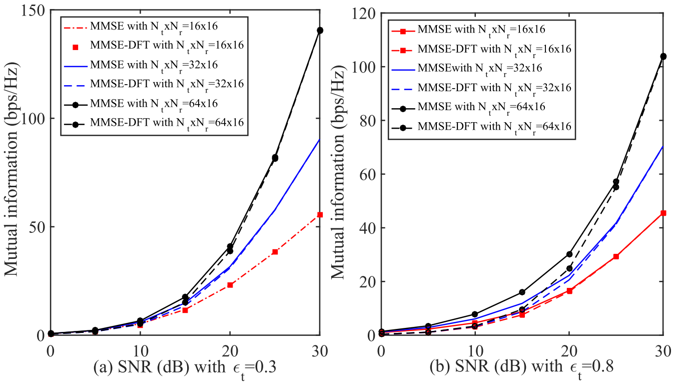

Fig. 9. Achievable MI performance of the MMSE estimator of Section III-B and benchmark MMSE-DFT versus the SNR, under different antenna setups of $N_t\times N_r = 16\times16$, $32\times16$ and $64\times16$ with (a) weak channel correlation of $\epsilon_t = 0.3$, and (b) strong channel correlation of $\epsilon_t = 0.8$.

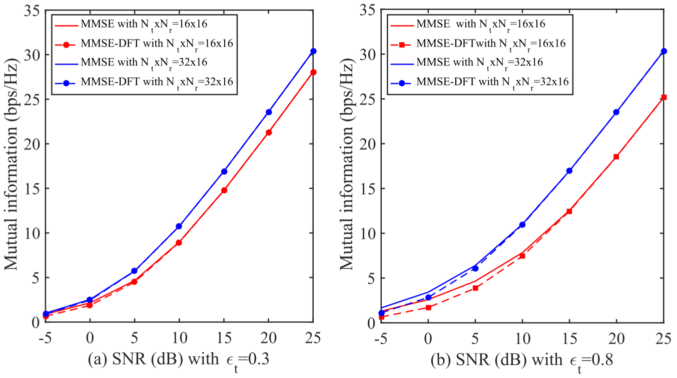

Fig. 10. Achievable MI performance of the MMSE estimator of Section IV-B.1 and benchmark MMSE-DFT versus the SNR, under different antenna setups of $N_t \times N_r = 16 \times 16$ and $32 \times 16$ with (a) weak channel correlation of $\epsilon_t = 0.3$, and (b) strong channel correlation of $\epsilon_t = 0.8$.

correlation. Firstly, we observe that increasing $N_t$ leads to better MI performance for the two estimator. Moreover, the MI performance of the two estimators are almost identical at low $N_t$ value and/or high SNR. Comparing Fig. 9 (a) and Fig. 9 (b), the MI performance gap between the two estimators for the weak channel correlation is clearly narrower than that for the strong counterpart, especially for large $N_t$ value.

We also extend this simulation to the composite channel estimation and compare the MI performance of the MMSE estimator proposed in Section IV-B.1 with the benchmark MMSE-DFT in Fig. 10. It is readily seen that the conclusions obtained in Fig. 9 are still applicable to Fig. 10. Furthermore, considering the much reduced channel dimensions to be estimated in Section IV, the achievable MI values of the two estimators are evidently lower than those of Fig. 9.

In summary, the proposed LS estimator requiring no channel statistics performs close to the proposed MMSE estimator of Section III-B relying on the second-order statistics of channel and noise for both weak and strong channel correlations. This is particularly so at high SNR. By contrast in the low-SNR regime, it can be concluded that for the case of weak channel correlation, all the proposed MMSE estimators of Sections III-B, IV-B1 and IV-B2 exhibit a similar performance to the corresponding MMSE-DFT estimator. In this case, it would be preferable to apply the low-complexity MMSE-DFT estimator. In the case of strong channel correlation, all the proposed MMSE estimators notably outperform the MMSE-DFT estimator and thus are preferred. Conversely, the proposed MMSE estimator of Section IV-B1 used for estimating the composite channel $\boldsymbol{G}_A\boldsymbol{H}\boldsymbol{F}_A$ is recommended in case of a large number of RF chains. Similarly, all the above conclusions for our proposed MMSE estimators still hold for the mutual information criterion.

## VI. Conclusion

In this paper, the training optimizations for hybrid MIMO communications has been investigated. Different from most of the existing works which rely on the existence of some special structures in the channel matrix to simplify channel estimation, we have considered a more general channel matrix without any special structures. Two channel estimation schemes have been proposed for hybrid MIMO systems. In each scheme, the training sequence and the analog matrices at the receiver and transmitter have been optimized for both sum MSE minimization and MI maximization. In the first scheme, the optimal structures of the analog matrices have been derived rigorously, which can overcome the nonconvex nature of analog matrix optimizations and simplify the joint optimizations of training sequence and analog matrices simultaneously. For the second scheme, the analog matrices have been optimized based on the statistical optimization framework, and the training sequence have been effectively optimized according to different channel statistical information. Simulation results have been used to demonstrate the effectiveness of the proposed designs.

## Appendix

Consider the following statistical optimization

$$\max\ \mathbb{E}\big\{f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}^{\mathrm{H}}\boldsymbol{\Sigma}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)\big\}, \tag{88}$$

where the elements of $\boldsymbol{H}_{\mathrm{W}}$ are i.i.d. Gaussian distributed random variables. The objective function is expressed by

$$\begin{aligned}&\mathbb{E}\big\{f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}^{\mathrm{H}}\boldsymbol{\Sigma}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)\big\}\\&\quad=\int f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}^{\mathrm{H}}\boldsymbol{\Sigma}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)p\big(\boldsymbol{H}_{\mathrm{W}}\big)\mathrm{d}\boldsymbol{H}_{\mathrm{W}},\end{aligned} \tag{89}$$

where $p(\boldsymbol{H}_{\mathrm{W}})$ is the probability density function of $\boldsymbol{H}_{\mathrm{W}}$, and $f(\cdot)$ is a unitary invariant function, i.e.,

$$f\big(\boldsymbol{U}_{\mathrm{L}}\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}^{\mathrm{H}}\boldsymbol{\Sigma}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\boldsymbol{U}_{\mathrm{L}}^{\mathrm{H}}\big)=f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}^{\mathrm{H}}\boldsymbol{\Sigma}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big), \tag{90}$$

for any unitary matrix $\boldsymbol{U}_{\mathrm{L}}$ of appropriate dimension. In addition, $f(\cdot)$ is a matrix-monotone increasing function, i.e., for two positive semidefinite matrices $\boldsymbol{A}\succeq\boldsymbol{B}$, $f(\boldsymbol{A})\geq f(\boldsymbol{B})$.

*Conclusion 5: The optimization problem (88) in nature aims to maximize* $\boldsymbol{\lambda}\big(\boldsymbol{\Sigma}^{\mathrm{H}}\boldsymbol{\Sigma}\big)$ *and* $\boldsymbol{\lambda}\big(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{H}}\big)$.

*Proof:* For $\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\succeq\boldsymbol{\Sigma}_2^{\mathrm{H}}\boldsymbol{\Sigma}_2$, we have

$$\begin{aligned}&f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)p\big(\boldsymbol{H}_{\mathrm{W}}\big)\mathrm{d}\boldsymbol{H}_{\mathrm{W}}\\&\qquad\geq f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}_2^{\mathrm{H}}\boldsymbol{\Sigma}_2\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)p\big(\boldsymbol{H}_{\mathrm{W}}\big)\mathrm{d}\boldsymbol{H}_{\mathrm{W}}\end{aligned} \tag{91}$$

based on which the following inequality holds

$$\mathbb{E}\big\{f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)\big\} \geq \mathbb{E}\big\{f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)\big\}. \quad (92)$$

Moreover, for $\boldsymbol{\lambda}\big(\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\big) \succeq \boldsymbol{\lambda}\big(\boldsymbol{\Sigma}_2^{\mathrm{H}}\boldsymbol{\Sigma}_2\big)$, there always exists a unitary matrix $\boldsymbol{U}$ that makes $\boldsymbol{U}\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{U}^{\mathrm{H}} \succeq \boldsymbol{\Sigma}_2^{\mathrm{H}}\boldsymbol{\Sigma}_2$. Therefore,

$$\begin{aligned}
&\int f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)p\big(\boldsymbol{H}_{\mathrm{W}}\big)\mathrm{d}\boldsymbol{H}_{\mathrm{W}}\\
&= \int f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{U}^{\mathrm{H}}\boldsymbol{U}\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{U}^{\mathrm{H}}\boldsymbol{U}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)p\big(\boldsymbol{H}_{\mathrm{W}}\big)\mathrm{d}\boldsymbol{H}_{\mathrm{W}}\\
&= \int f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{U}^{\mathrm{H}}\boldsymbol{U}\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{U}^{\mathrm{H}}\boldsymbol{U}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)p\big(\boldsymbol{U}\boldsymbol{H}_{\mathrm{W}}\big)\mathrm{d}\boldsymbol{U}\boldsymbol{H}_{\mathrm{W}}\\
&= \int f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{U}\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{U}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)p\big(\boldsymbol{H}_{\mathrm{W}}\big)\mathrm{d}\boldsymbol{H}_{\mathrm{W}}\\
&\geq \int f\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}_2^{\mathrm{H}}\boldsymbol{\Sigma}_2\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big)p\big(\boldsymbol{H}_{\mathrm{W}}\big)\mathrm{d}\boldsymbol{H}_{\mathrm{W}}, \quad (93)
\end{aligned}$$

where the second and the third equalities are due to the facts that the elements of $\boldsymbol{H}_{\mathrm{W}}$ are i.i.d. Gaussian distributed and $\boldsymbol{H}_{\mathrm{W}}$ and $\boldsymbol{U}\boldsymbol{H}_{\mathrm{W}}$ have the same distribution. The final inequality is derived based on (91).

Note that $\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}^{\mathrm{H}}\boldsymbol{\Sigma}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}^{\mathrm{H}}$ have the same nonzero eigenvalues. As a result, the following two optimization problems are equivalent

$$\max \boldsymbol{\lambda}\big(\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}^{\mathrm{H}}\boldsymbol{\Sigma}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\big) \Leftrightarrow \max \boldsymbol{\lambda}\big(\boldsymbol{\Sigma}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}^{\mathrm{H}}\big). \quad (94)$$

Based on the unitary invariant property in (90) and (93), it can be proved that when $\boldsymbol{\lambda}\big(\boldsymbol{\Psi}_1\boldsymbol{\Psi}_1^{\mathrm{H}}\big) \succeq \boldsymbol{\lambda}\big(\boldsymbol{\Psi}_2\boldsymbol{\Psi}_2^{\mathrm{H}}\big)$, the following inequality holds

$$\int f\big(\boldsymbol{\Psi}_1^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}_1\big)p\big(\boldsymbol{H}_{\mathrm{W}}\big)\mathrm{d}\boldsymbol{H}_{\mathrm{W}} \geq \int f\big(\boldsymbol{\Psi}_2^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{\Sigma}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Psi}_2\big)p\big(\boldsymbol{H}_{\mathrm{W}}\big)\mathrm{d}\boldsymbol{H}_{\mathrm{W}}. \quad (95)$$

This completes the proof. ■

## References

[1] P. Zhang, S. Chen, and L. Hanzo, "Two-tier channel estimation aided near-capacity MIMO transceivers relying on norm-based joint transmit and receive antenna selection," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 122–137, Jan. 2015.

[2] Z. Wang, W. Liu, C. Qian, S. Chen, and L. Hanzo, "Two-dimensional precoding for 3-D massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5485–5490, Jun. 2017.

[3] J. Zhang, S. Chen, R. G. Maunder, R. Zhang, and L. Hanzo, "Adaptive coding and modulation for large-scale antenna array-based aeronautical communications in the presence of co-channel interference," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1343–1357, Feb. 2018.

[4] W. Liu, Z. Wang, J. Cao, S. Chen, and L. Hanzo, "Partially-activated conjugate beamforming for LoS massive MIMO communications," *IEEE Access*, vol. 6, pp. 56504–56513, 2018.

[5] A. Pastore, M. Joham, and J. R. Fonollosa, "A framework for joint design of pilot sequence and linear precoder," *IEEE Trans. Inf. Theory*, vol. 62, no. 9, pp. 5059–5079, Sep. 2016.

[6] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[7] A. Soysal and S. Ulukus, "Joint channel estimation and resource allocation for MIMO systems-part I: Single-user analysis," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 624–631, Feb. 2010.

[8] A. Soysal and S. Ulukus, "Joint channel estimation and resource allocation for MIMO systems-part II: Multi-user and numerical analysis," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 632–640, Feb. 2010.

[9] M. Coldrey and P. Bohlin, "Training-based MIMO systems—Part I: Performance comparison," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5464–5476, Nov. 2007.

[10] M. Coldrey and P. Bohlin, "Training-based MIMO systems: Part II—Improvements using detected symbol information," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 296–303, Jan. 2008.

[11] Z. Gao, L. Dai, Z. Wang, and S. Chen, "Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169–6183, Dec. 2015.

[12] T. F. Wong and B. Park, "Training sequence optimization in MIMO systems with colored interference," *IEEE Trans. Commun.*, vol. 52, no. 11, pp. 1939–1947, Nov. 2004.

[13] Y. Liu, T. F. Wong, and W. W. Hager, "Training signal design for estimation of correlated MIMO channels with colored interference," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1486–1497, Apr. 2007.

[14] M. Biguesh and A. B. Gershman, "Training-based MIMO channel estimation: A study of estimator tradeoffs and optimal training signals," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 884–893, Mar. 2006.

[15] F. Gao, T. Cui, and A. Nallanathan, "Optimal training design for channel estimation in decode-and-forward relay networks with individual and total power constraints," *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 5937–5949, Dec. 2008.

[16] X. Guo, S. Chen, J. Zhang, X. Mu, and L. Hanzo, "Optimal pilot design for pilot contamination elimination/reduction in large-scale multiple-antenna aided OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7229–7243, Nov. 2016.

[17] H. Xie, F. Gao, S. Jin, J. Fang, and Y.-C. Liang, "Channel estimation for TDD/FDD massive MIMO systems with channel covariance computing," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4206–4218, Jun. 2018.

[18] X. Cheng, J. Sun, and S. Li, "Channel estimation for FDD multi-user massive MIMO: A variational Bayesian inference-based approach," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7590–7602, Nov. 2017.

[19] C.-Y. Wu, W.-J. Huang, and W.-H. Chung, "Low-complexity semiblind channel estimation in massive MU-MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6279–6290, Sep. 2017.

[20] X. Xiong, X. Wang, X. Gao, and X. You, "Beam-domain channel estimation for FDD massive MIMO systems with optimal thresholds," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4669–4682, Jul. 2017.

[21] J. Fang, X. Li, H. Li, and F. Gao, "Low-rank covariance-assisted downlink training and channel estimation for FDD massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1935–1947, Mar. 2017.

[22] D. Fan, F. Gao, G. Wang, Z. Zhong, and A. Nallanathan, "Angle domain signal processing-aided channel estimation for indoor 60-GHz TDD/FDD massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1948–1961, Sep. 2017.

[23] J. Zhao, F. Gao, W. Jia, S. Zhang, S. Jin, and H. Lin, "Angle domain hybrid precoding and channel tracking for millimeter wave massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6868–6880, Oct. 2017.

[24] W. Xu, W. Xiang, Y. Jia, Y. Li, and Y. Yang, "Downlink performance of massive-MIMO systems using EVD-based channel estimation," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3045–3058, Apr. 2017.

[25] D. Kong, D. Qu, K. Luo, and T. Jiang, "Channel estimation under staggered frame structure for massive MIMO system," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1469–1479, Feb. 2016.

[26] C. Xing, X. Zhao, W. Xu, X. Dong, and G. Y. Li, "A framework on hybrid MIMO transceiver design based on matrix-monotonic optimization," *IEEE Trans. Signal Process.*, vol. 67, no. 13, pp. 3531–3546, Jul. 2019.

[27] L. Pan, L. Liang, W. Xu, and X. Dong, "Framework of channel estimation for hybrid analog-and-digital processing enabled massive MIMO communications," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3902–3915, Sep. 2018.

[28] W. Ni and X. Dong, "Hybrid block diagonalization for massive multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 201–211, Jan. 2016.

[29] W. Ni, X. Dong, and W.-S. Lu, "Near-optimal hybrid processing for massive MIMO systems via matrix decomposition," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 3922–3933, Aug. 2017.

[30] S. Wang, S. Ma, C. Xing, S. Gong, J. An, and H. V. Poor, "Optimal training design for MIMO systems with general power constraints," *IEEE Trans. Signal Process.*, vol. 66, no. 14, pp. 3649–3664, Jul. 2018.

[31] C. Xing, S. Ma, and Y. Zhou, "Matrix-monotonic optimization for MIMO systems," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 334–348, Jan. 2015.

[32] J. Jin, C. Xiao, W. Chen, and Y. Wu, "Channel-statistics-based hybrid precoding for millimeter-wave MIMO systems with dynamic subarrays," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 3991–4003, Jun. 2019.

[33] D. Zhu, B. Li, and P. Liang, "A novel hybrid beamforming algorithm with unified analog beamforming by subspace construction based on partial CSI for massive MIMO-OFDM systems," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 594–607, Feb. 2017.

[34] A. Liu and V. K. N. Lau, "Impact of CSI knowledge on the codebook-based hybrid beamforming in massive MIMO," *IEEE Trans. Signal Process.*, vol. 64, no. 24, pp. 6545–6556, Dec. 2016.

[35] D. Katselis, E. Kofidis, and S. Theodoridis, "On training optimization for estimation of correlated MIMO channels in the presence of multiuser interference," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4892–4904, Oct. 2008.

[36] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[37] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*. New York, NY, USA: Academic, 1979.

[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[39] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications* (MPS-SIAM Series on Optimization). Philadelphia, PA, USA: SIAM, 2001.

[40] M. C. Grant and S. P. Boyd. (2015). The CVX Users' Guide (Release 2.1). CVX Research. [Online]. Available: http://web.cvxr.com/cvx/beta/doc/CVX.pdf

**Chengwen Xing** (Member, IEEE) received the B.Eng. degree from Xidian University, Xi'an, China, in 2005, and the Ph.D. degree from The University of Hong Kong, Hong Kong, in 2010. Since September 2010, he has been with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, where he is currently a Full Professor. From September 2012 to December 2012, he was a Visiting Scholar with the University of Macau, Macau, China. His current research interests include statistical signal processing, convex optimization, multivariate statistics, combinatorial optimization, massive MIMO systems, and high frequency band communication systems. He was an Associate Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2013 to 2019, and currently serves as an Associate Editor of *KSII Transactions on Internet and Information Systems, Transactions on Emerging Telecommunications Technologies*, and *China Communications*.

**Dekang Liu** received the B.S. degree in electronic engineering from the Beijing Institute of Technology, Beijing, China, in 2013, where he is currently pursuing the Ph.D. degree with the School of Electronic and Information. His research interests are in the area of signal processing, massive MIMO systems, high frequency band communication systems, and convex optimization.

**Shiqi Gong** (Student Member, IEEE) received the B.S. degree in electronic engineering from the Beijing Institute of Technology, Beijing, China, in 2014, where she is currently pursuing the Ph.D. degree with the School of Electronic and Information. She also served as a Research Assistant of the Faculty of Science and Technology, University of Macau. Her research interests are in the area of signal processing, physical-layer security, resource allocation, and convex optimization.

**Wei Xu** (Senior Member, IEEE) received the B.Sc. degree in electrical engineering and the M.S. and Ph.D. degrees in communication and information engineering from Southeast University, Nanjing, China, in 2003, 2006, and 2009, respectively. From 2009 to 2010, he was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, University of Victoria, Canada. He is currently a Professor with the National Mobile Communications Research Laboratory, Southeast University. He is also an Adjunct Professor with the University of Victoria, Canada, and a Distinguished Visiting Fellow of the Royal Academy of Engineering, U.K. He has coauthored over 100 refereed journal articles in addition to 36 domestic patents and four U.S. patents granted. His research interests include cooperative communications, information theory, signal processing, and machine learning for wireless communications. He received the Best Paper Awards from IEEE MAPE 2013, IEEE/CIC ICCC 2014, IEEE Globecom 2014, IEEE ICUWB 2016, WCSP 2017, and ISWCS 2018. He was a co-recipient of the First Prize of the Science and Technology Award in Jiangsu, China, in 2014. He received the Youth Science and Technology Award of the China Institute of Communications in 2018. He was an Editor of the IEEE COMMUNICATIONS LETTERS from 2012 to 2017. He is currently an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE ACCESS.

**Sheng Chen** (Fellow, IEEE) received the B.Eng. degree in control engineering from the East China Petroleum Institute, Dongying, China, in 1982, the Ph.D. degree in control engineering from the City, University of London, in 1986, and the D.Sc. degree from the University of Southampton, Southampton, U.K., in 2005. From 1986 to 1999, he held research and academic appointments at the Universities of Sheffield, Edinburgh and Portsmouth, all in U.K. Since 1999, he has been with the School of Electronics and Computer Science, University of Southampton, U.K., where he holds the position of Professor in intelligent systems and signal processing. He has published over 650 research papers. He has more than 13,000 Web of Science citations and more than 27,000 Google Scholar citations. His research interests include adaptive signal processing, wireless communications, modeling and identification of nonlinear systems, neural networks, and machine learning, intelligent control system design, evolutionary computation methods, and optimization. He is a fellow of the United Kingdom Royal Academy of Engineering, a fellow of IET, a Distinguished Adjunct Professor at King Abdulaziz University, Jeddah, Saudi Arabia, and was recognized as an original ISI highly cited researcher in engineering in March 2004.

**Lajos Hanzo** (Fellow, IEEE) received the master's degree in electronics and the Ph.D. degree from the Technical University (TU) of Budapest in 1976 and 1983, respectively, the D.Sc. degree from the University of Southampton, in 2004, and the Ph.D. degree (Hons.) from The University of Edinburgh in 2015. He has published more than 1900 contributions on IEEE Xplore, 19 Wiley-IEEE Press books, and has helped fast-track the careers of 119 Ph.D. students. Over 40 of these Ph.D. candidates are professors at various stages of their careers in academia and many are leading scientists in the wireless industry. He is a fellow of the Royal Academy of Engineering, Institution of Engineering and Technology, and EURASIP. He was awarded Honorary Doctorates by the TU of Budapest in 2009 and by The University of Edinburgh in 2015. He is a Foreign Member of the Hungarian Academy of Sciences and a former Editor-in-Chief of the IEEE Press. He has served as a Governor of both IEEE ComSoc and VTS.