A Unified Joint Optimization of Training Sequences and Transceivers Based on Matrix-Monotonic Optimization

Chengwen Xing, *Member, IEEE*, Tao Yu, Jinpeng Song, *Member, IEEE*, Zhong Zheng, *Member, IEEE*, Lian Zhao, *Senior Member, IEEE*, and Lajos Hanzo, *Life Fellow, IEEE*

Abstract—Channel estimation and data transmission constitute the most fundamental functional modules of multiple-input multiple-output (MIMO) communication systems. The underlying key tasks corresponding to these modules are training sequence optimization and transceiver optimization. Hence, we jointly optimize the linear transmit precoder and the training sequence of MIMO systems using the metrics of their effective mutual information (MI), effective mean squared error (MSE), effective weighted MI, effective weighted MSE, as well as their effective generic Schur-convex and Schur-concave functions. Both statistical channel state information (CSI) and estimated CSI are considered at the transmitter in the joint optimization. A unified framework termed as joint matrix-monotonic optimization is proposed. Based on this, the optimal precoder matrix and training matrix structures can be derived for both CSI scenarios. Then, based on the optimal matrix structures, our linear transceivers and their training sequences can be jointly optimized. Compared to state-of-the-art benchmark algorithms, the proposed algorithms visualize the bold explicit relationships between the attainable system performance of our linear transceivers conceived and their training sequences, leading to implementation ready recipes. Finally, several numerical results are provided, which corroborate our theoretical results and demonstrate the compelling benefits of our proposed pilot-aided MIMO solutions.

Index Terms—Channel estimation, data transmission, resource allocation, matrix-monotonic optimization.

I. Introduction

In the evolution of commercial cellular networks from 4G to next generation solutions, multiple-input multiple-output

C. Xing, T. Yu, and Z. Zheng are with School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: xingchengwen@gmail.com, taoyu.bit@gmail.com, zhong.zheng@bit.edu.cn).

J. Song is with School of Cyberspace Science and Technology, Beijing Institute of technology, Beijing 100081, China (e-mail: jinpeng@bit.edu.cn).

L. Zhao is with the Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: l5zhao@ryerson.ca).

L. Hanzo is with the Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK (e-mail: lh@ecs.soton.ac.uk).

(MIMO) technology constitutes a salient milestone [1]–[9], leading to significant spectral and power efficiency improvements. However, they require accurate channel state information (CSI) [10]–[15]. Therefore, pilot-aided channel estimation plays an important role in multi-antenna system design [16]–[21]. Based on the noise-contaminated pilot observations, the MIMO channel matrix can be estimated by relying on performance metrics such as the least square error [22], minimum mean squared error (MSE) [23], [24], mutual information (MI) [26], [27], etc. Then, by leveraging the estimated CSI, MIMO transceivers can be designed for optimizing the overall system performance [23]–[25].

The specific choice of the training sequence has a pivotal impact on the performance of channel estimation for MIMO systems [21], [28], and the key task of channel estimation is to recover the MIMO channel matrix instead of a scalar parameter. Hence, MIMO channel estimation constitutes a multiple-parameter estimation problem, where the training sequence optimization relies either on signal processing oriented [29], [30] or on information theoretic metrics [31]. It has been shown based on matrix-monotonic optimization [10] that in fact a diverse variety of training sequence optimization techniques relying on heterogeneous performance metrics intrinsically aim for maximizing a matrix-valued signal-to-noise ratio (SNR) during the channel estimation procedure.

It is widely exploited that MIMO transceivers are capable of striking a tradeoff between the spatial diversity gain and the spatial multiplexing gain attained by the MIMO channels. When multiple data streams are transmitted simultaneously from the source to destination, the transceiver optimization constitutes a multi-objective optimization problem [32]–[34]. Given the CSI, diverse performance metrics, such as the capacity, or the sum MSE of the estimation may be optimized [11]. In order to accommodate as many optimization objective functions (OFs) as possible, majorization theory is used in [11], [35] to formulate unified OFs [1]. Recently, the framework of matrix-monotonic optimization was proposed for MIMO transceiver optimization relying on diverse OFs [10], [32], [33]. Again, as pointed out in [10], the MIMO precoder optimization task is reminiscent of maximizing a matrix-valued SNR in the data transmission procedure.

---

[1]Explicitly, we will use the adjective "unified" upon referring to unifying diverse OFs, and the adjective "joint" upon referring to the optimization of both the training sequence and precoder.

When channel estimation is more accurate, at first right the illusion might appear that the successive data transmission will have improved performance. However, when either the pilot overload or the pilot power is increased, either the effective throughput or the power - or potentially both - have to be reduced for data transmission for the sake of fair comparison. This tradeoff reveals that the joint optimization of the transmit precoder (TPC) and of the training sequence is of critical importance for MIMO systems. Hence, there is a rich body of literature on joint training and transceiver optimization [23], [25], but in most of the related literature, the joint optimization tends to rely on a single performance metric. For example, joint TPC and training sequence optimization relying on MI maximization was extensively studied in [23]–[25]. However, using a joint training sequence and transceiver weight optimization framework relying on multiple performance metrics - rather than a single one, like the MI in [23]–[25] - has important theoretical and practical benefits. Having said that, there is a distinct paucity of literature on this topic - probably because their joint optimization is much more challenging than that of its separate counterparts. This becomes particularly challenging, when both the time domain as well as the inter-element correlation of the channel's transmit side are taken into account [25]. The main challenge arises from the complex performance analysis of the associated random matrix variables [19], [23], [24]. The distinct contributions of this paper are shown in Table I.

TABLE I
KEY CONTRIBUTIONS OF THIS PAPER CONTRASTED TO THE EXISTING
JOINT DESIGN OF TRAINING SEQUENCES AND MIMO TRANSCEIVERS
WITH THE POWER CONSTRAINED

	[23]	[25]	Our
Transmit-side correlation channel model	$\checkmark$	$\checkmark$	$\checkmark$
Statistical CSI known at transmitter	$\checkmark$	$\checkmark$	$\checkmark$
Estimated CSI known at transmitter			$\checkmark$
OFs	effective MI	different OFs	applicable to OFs in [25] and Schur-concave/convex functions
Maximum number of antennas in simulations	3	2	16

In general MIMO communication systems, multi-dimensional signals have to be processed both in channel estimation and data transmission, where diverse tradeoffs must be struck. The matrix-version SNR is capable of accurately characterizing these tradeoffs. Matrix-monotonic optimization handles matrix variables, where a positive semi-definite matrix is used as the performance metric. This optimization procedure exploits the monotonicity in the positive semi-definite matrix cone for deriving the optimal structures of the Pareto optimal matrix variables [10]. Generally speaking, as for MIMO systems based on matrix-monotonic optimization, both the precoder optimization and training optimization aim for maximizing a matrix-version SNR instead of a scalar SNR.

This general multi-objective optimization procedure offers substantial flexibility for mathematical derivations. Based on the optimal structures derived, the optimization problems of MIMO-aided communications can be significantly simplified, resulting into traditional vector-valued optimization problems.

Moreover, a close scrutiny reveals that the nature of training optimization and of transceiver optimization is quite similar to each other [10]. Hence, we are inspired to investigate the joint optimization of the training sequence design and the TPC design based on the matrix-monotonic optimization framework of [10]. A novel matrix-monotonic optimization technique termed as joint matrix-monotonic optimization is proposed for the joint optimization of the training sequence and of a linear TPC, which involves a pair of matrix variables, namely the training sequence matrix and linear TPC matrix. Our main contributions are as follows:

- A whole suite of performance metrics is considered in the joint optimization of linear transceivers and training sequences, including the effective MI, effective sum MSE, effective weighted MI, effective weighted MSE, as well as the effective general Schur-convex and Schur-concave functions. Furthermore, a range of practical settings is considered. Specifically, for channel estimation, the power of the amplifier is limited, hence the estimation accuracy could be improved by increasing the length of the training sequence instead of increasing the power of the amplifiers.
- For the joint optimization of linear TPCs and training sequences, the spatial correlation of antennas experienced at the transmitter is taken into account. Moreover, both statistical CSI and estimated CSI are discussed. More explicitly for the former scenario, the transmitter only has statistical CSI information, while the receiver relies on the estimated CSI. By contrast, for the latter case, the transmitter and the receiver have the same estimated CSI, i.e. the feedback channel between the transmitter and receiver is assumed to be perfect.
- Our matrix-monotonic optimization framework of [10] derived for separate transceiver optimization or training optimization is extended here to the new joint matrix-monotonic optimization concept, which jointly optimizes our linear TPC as well as the training sequence, and judiciously allocates the system resources to optimize the effective performances of integrated pilot-aided MIMO communication systems. Based on our new joint matrix-monotonic optimization framework, the optimal structures of the linear TPC matrix and of the training sequence matrix are derived. Therefore, the resultant joint optimization problems can be significantly simplified. Although the original joint optimization problems are complex, the proposed solution is of appealingly low complexity, which provides valuable guidelines for practical system designs.

The rest of the paper is organized as follows. In Section II, the signal models of channel estimation and information transmission are provided, respectively. The general joint training sequence and linear transceiver optimization problem

Step 0. Channel model

Build transmit-side correlated channel model.

Step 1. Characteristic of separate optimization

Explore the duality between the optimization of training sequences and transceivers from P. 1 to P.12.

Step 2. Characteristic of joint optimization

Formulate the unified joint optimization of training sequences and transceivers in P.13.

Step 3. Solution with statistical CSI

Derive optimal structure and power allocation based on matrix-monotonic optimization with statistical CSI known at transmitter and discuss the specific examples in Section III-B and C.

Step 4. Solution with estimated CSI

Analyze in the same way as Step 3 with estimated CSI known at transmitter in Section III-D and E.

Fig. 1. Flow of the mathematical analysis.

is formulated and solved in Section III. Specifically, the joint optimization of the training sequence and transceiver is investigated in Section III-B when the transmitter has only statistical CSI. By contrast, the transmitter relying on estimated CSI is investigated in Section III-D. Our simulation results are discussed in Section IV, while our conclusions are offered in Section V. Fig. 1 illustrates the flow of mathematical analysis. **Notation:** In order to clarify the following mathematical derivations, we introduce the notations, symbols and definitions used throughout this paper. Referring to the fundamental matrix operations, the symbols Z^{H} and Z^{T} denote the Hermitian transpose and the transpose of a general matrix Z, respectively. The trace and determinant of a square matrix Z are denoted as $\mathrm{Tr}(Z)$ and $|Z|$, respectively. For a positive semidefinite matrix Z, the matrix $Z^{\frac{1}{2}}$ or $Z^{1/2}$ is the Hermitian square root of Z, which is also positive semidefinite. The mathematical notation z^{+} represents $\max\{0, z\}$. For matrix decompositions in this paper, the notations $\mathbf{\Lambda} \searrow$ and $\mathbf{\Lambda} \nearrow$ represent rectangular or square diagonal matrices with their diagonal elements sorted in decreasing and increasing order, respectively.

II. Signal Models for Channel Estimation and Data Transmission

A typical point-to-point MIMO system is considered. Specifically, an $N_{\mathrm{R}} \times N_{\mathrm{T}}$ MIMO channel H associated with transmit-side spatial antenna correlation is investigated, where the channel H can be formulated as in [23], [25], [32]:

$$H = H_{\mathrm{W}} \mathbf{\Psi}^{1/2}, \qquad (1)$$

where the entries of H_{W} are independent and identically distributed (i.i.d.) Gaussian random variables having zero mean and unit variance. In the signal model (1), the positive definite matrix $\mathbf{\Psi}$ represents the transmit correlation matrix. Both channel estimation and data transmission are considered assuming the channel model (1), which are performed consecutively. In the first phase, training sequences are transmitted. Based on the noise-contaminated training signal observations, the channel matrix is estimated at the receiver. In the second phase, based on the estimated channel matrix, the transmit precoded signals are constructed and transmitted, which are

then recovered at the receiver. In the following, we discuss the signal models of these two phases respectively. The connections and equivalences among different optimization metrics and OFs are illustrated in Fig. 2. Firstly, we discuss different OFs and their unified matrix-monotonic optimizations of separate designs in channel estimation and data transmission respectively, in order to explore the common traits within them [10]. More details on the differences and physical meanings among these OFs may be found in Appendix A. Then, we take the resource allocation into consideration and formulate a joint optimization problem relying on a unified framework in Section III, which is applicable to various OFs. Specifically, two scenarios are taken into consideration, which rely on either statistical or estimated CSI at the transmitter. The unified problem is derived, and the optimal structure is obtained. Then, some specific OFs are discussed and the corresponding solutions are provided.

A. Channel Estimation and Training Optimization

For the training based channel estimation, the training sequence X is transmitted to the receivers, yielding the received signal sequence of

$$Y = HX + N, \qquad (2)$$

where N is the additive noise matrix at the destination. Based on the signal model in (2), the key task is to recover the channel matrix H from the noise-contaminated observation Y as accurately as possible. For a linear channel estimator, the estimated channel equals [25]

$$\widehat{H} = Y G_{\mathrm{E}}, \qquad (3)$$

where the estimated channel matrix $\widehat{H}$ and the true channel matrix H satisfy

$$H = \widehat{H} + \Delta H, \qquad (4)$$

where ΔH is the estimation error. Then the MSE matrix of the corresponding channel estimation can be expressed as

$$\begin{aligned} E_{\mathrm{MSE}} &= \mathbb{E}\{\Delta H^{\mathrm{H}} \Delta H\} \\ &= (I - X G_{\mathrm{E}})^{\mathrm{H}} R_{\mathrm{H}} (I - X G_{\mathrm{E}}) + G_{\mathrm{E}}^{\mathrm{H}} R_{\mathrm{N}} G_{\mathrm{E}}, \end{aligned} \qquad (5)$$

where the channel's correlation matrix R_{H} and the noise covariance matrix R_{N} are defined as

$$R_{\mathrm{H}} = \mathbb{E}\{H^{\mathrm{H}} H\} = N_{\mathrm{R}} \mathbf{\Psi}, \ R_{\mathrm{N}} = \mathbb{E}\{N^{\mathrm{H}} N\}, \qquad (6)$$

where the scalar N_{R} denotes the number of receive antennas. In order to minimize E_{MSE}, the optimal G_{E} can be chosen as the linear minimum mean squared error (LMMSE) channel estimator as in [36], [37]:

$$G_{\mathrm{E,Opt}} = (X^{\mathrm{H}} R_{\mathrm{H}} X + R_{\mathrm{N}})^{-1} X^{\mathrm{H}} R_{\mathrm{H}}. \qquad (7)$$

Therefore, the channel estimation MSE matrix E_{MSE} in (5) can be reformulated as

$$\begin{aligned} E_{\mathrm{MSE}} &= \mathbb{E}\{\Delta H^{\mathrm{H}} \Delta H\} \\ &= (R_{\mathrm{H}}^{-1} + X R_{\mathrm{N}}^{-1} X^{\mathrm{H}})^{-1} \\ &= N_{\mathrm{R}} \left(\mathbf{\Psi}^{-1} + N_{\mathrm{R}} X R_{\mathrm{N}}^{-1} X^{\mathrm{H}}\right)^{-1}. \end{aligned} \qquad (8)$$
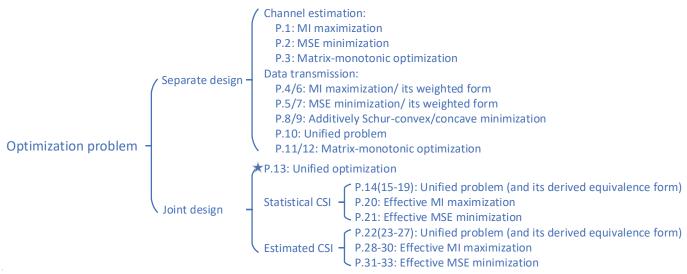
Fig. 2. Overview of the OFs in the paper.

Meanwhile, based on the LMMSE estimator in (7), the resultant channel estimation error $\Delta \boldsymbol{H}$ can be written in the following form [37]

$$\Delta \boldsymbol{H} = \Delta \boldsymbol{H}_{\mathrm{W}} \boldsymbol{\Phi}^{1/2}, \qquad (9)$$

where $\Delta \boldsymbol{H}_{\mathrm{W}}$ is a random matrix whose elements are i.i.d. Gaussian distributed with zero mean and unit variance, while $\boldsymbol{\Phi}$ is given in (8) as

$$\boldsymbol{\Phi} = \left(\boldsymbol{\Psi}^{-1} + N_{\mathrm{R}} \boldsymbol{X} \boldsymbol{R}_{\mathrm{N}}^{-1} \boldsymbol{X}^{\mathrm{H}} \right)^{-1}. \qquad (10)$$

Moreover, upon applying the LMMSE channel estimator of (7), the correlation matrix of the estimated channel $\mathbb{E}\{\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\}$ equals

$$\begin{aligned} \mathbb{E}\{\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\} &= \boldsymbol{R}_{\mathrm{H}} - (\boldsymbol{R}_{\mathrm{H}}^{-1} + \boldsymbol{X}\boldsymbol{R}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}})^{-1} \\ &= N_{\mathrm{R}}\boldsymbol{\Psi} - N_{\mathrm{R}}\left(\boldsymbol{\Psi}^{-1} + N_{\mathrm{R}}\boldsymbol{X}\boldsymbol{R}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\right)^{-1} \triangleq \boldsymbol{\Pi}. \end{aligned} \qquad (11)$$

When the LMMSE channel estimator is adopted, the channel estimation MSE matrix $\boldsymbol{E}_{\mathrm{MSE}}$ is a function of the training sequence $\boldsymbol{X}$. Therefore, the performance of channel estimation may indeed be improved by optimizing the choice of the training sequence $\boldsymbol{X}$ [10]. The most widely used performance metrics are the MI [31] and MSE [10]. In the following, these two classic training optimization techniques are reviewed in order to reveal the general optimal structure of the training sequence $\boldsymbol{X}$. As for MI maximization, the corresponding training optimization problem is formulated as in [26], [27], [31]:

$$\begin{aligned} \textbf{P. 1:} \quad & \max_{\boldsymbol{X}} \ \log \det\left(\boldsymbol{R}_{\mathrm{H}}^{-1} + \boldsymbol{X}\boldsymbol{R}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\right), \\ & \text{s.t. } \mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}) \leq P_{\mathrm{T}}T_{\mathrm{T}}, \end{aligned} \qquad (12)$$

where P_{T} and T_{T} are the power and the time interval of channel estimation, respectively, and $P_{\mathrm{T}}T_{\mathrm{T}}$ is the energy allocated to the channel estimation. This energy constraint can be viewed as the Frobenius norm of the training sequence matrix. It is worth noting that for matrix-valued training sequences there are multiple solutions that have the same

Frobenius. As a result, the optimization aims for finding the optimal training sequence matrix in the whole feasible solution set. On the other hand, the training optimization aiming for minimizing the sum MSE can be formulated in the following form [30]

$$\begin{aligned} \textbf{P. 2:} \quad & \min_{\boldsymbol{X}} \ \mathrm{Tr}[(\boldsymbol{R}_{\mathrm{H}}^{-1} + \boldsymbol{X}\boldsymbol{R}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}})^{-1}], \\ & \text{s.t. } \mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}) \leq P_{\mathrm{T}}T_{\mathrm{T}}. \end{aligned} \qquad (13)$$

Based on **P. 1** and **P. 2**, it may be inferred that using different performance metrics for MIMO channel estimation results in different optimal sequences, because multiple parameters have to be estimated in MIMO channel estimation. Different performance metrics result in different tradeoffs. Therefore, a general MIMO training optimization results in a multi-objective optimization problem. From a multi-objective optimization perspective, all the optimal solutions of the training optimization associated with different performance metrics are the Pareto optimal[2] solutions of the following matrix-monotonic optimization problem [10]

$$\textbf{P. 3:} \quad \max_{\boldsymbol{X}} \ \boldsymbol{X}\boldsymbol{R}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}, \text{ s.t. } \mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}) \leq P_{\mathrm{T}}T_{\mathrm{T}}. \qquad (14)$$

We would like to highlight that in the above optimization problem, the OF is a positive semi-definite matrix, which may be viewed as a matrix-valued SNR in the channel estimation procedure [10]. Bearing in mind that the constraint in **P. 3** is unitary invariant, the matrix-monotonic optimization **P. 3** is of a vector optimization nature, which is a function of the vector consisting of the eigenvalues of $\boldsymbol{X}\boldsymbol{R}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}$. The different Pareto optimal solutions of **P. 3** achieve different levels of fairness among the different accuracies of the multiple estimated parameters. The channel estimation performance may be improved by optimizing the structure of the training sequence matrix depending on the availability of the channel's statistical information. From a practical implementation

[2]Pareto optimality refers to the fact that, in a multi-objective optimization problem, it is impossible to improve any of the objectives without degrading at least one of the other objectives.

perspective, another design option is to gather more energy at the receiver instead of increasing the transmit power, because a practical power amplifier has a limited linear operating region. In other words, the channel estimation accuracy may also be improved by increasing the number of columns in $\boldsymbol{X}$, instead of increasing the amplitudes of the elements of $\boldsymbol{X}$. Based on the matrix-monotonic optimization framework of [10], we have the following conclusions for training optimization.

Conclusion 1: In MIMO channel estimation relying on an LMMSE estimator, the optimal training sequences satisfy the following structure

$$\boldsymbol{X}_{\text{opt}} = \boldsymbol{U_X}\boldsymbol{\Lambda_X}\boldsymbol{U}_{\boldsymbol{R}_{\text{N}}}^{\text{H}}, \tag{15}$$

where the unitary matrix $\boldsymbol{U}_{\boldsymbol{R}_{\text{N}}}$ is defined based on the eigenvalue decomposition (EVD) as

$$\boldsymbol{R}_{\text{N}}^{-1} = \boldsymbol{U}_{\boldsymbol{R}_{\text{N}}}\boldsymbol{\Lambda}_{\boldsymbol{R}_{\text{N}}}^{-1}\boldsymbol{U}_{\boldsymbol{R}_{\text{N}}}^{\text{H}} \text{ with } \boldsymbol{\Lambda}_{\boldsymbol{R}_{\text{N}}}^{-1} \searrow. \tag{16}$$

The unitary matrix $\boldsymbol{U_X}$ is determined by the specific performance metrics. Finally, $\boldsymbol{\Lambda_X}$ is a rectangular diagonal matrix, which is also determined by the specific performance metric.

Conclusion 2: Referring to $\boldsymbol{U_X}$ for both MI maximization and MSE minimization, the optimal $\boldsymbol{U_X}$ obeys [10]

$$\boldsymbol{U_X} = \boldsymbol{U_\Psi}, \tag{17}$$

where the unitary matrix $\boldsymbol{U_\Psi}$ is defined based on the following EVD:

$$\boldsymbol{\Psi} = \boldsymbol{U_\Psi}\boldsymbol{\Lambda_\Psi}\boldsymbol{U_\Psi}^{\text{H}} \text{ with } \boldsymbol{\Lambda_\Psi} \searrow. \tag{18}$$

Based on **Conclusion 2**, the structure $\boldsymbol{U_X} = \boldsymbol{U_\Psi}$ is applied in the general joint optimization of the training and the transceiver in Section III. Moreover, it is also worth noting that at high SNRs[3] the MI maximization **P. 1** has the optimal solution of $\boldsymbol{\Lambda_X} = \boldsymbol{I}$, i.e., uniformly allocating the powers along spatial directions [10]. By contrast, for MSE minimization, the optimal $\boldsymbol{\Lambda_X}$ is the water-filling solution [10]. Undoubtedly, uniform power allocation is the simplest scheme for training optimization, which can be adopted as a suboptimal power allocation scheme along the spatial directions.

B. Transceiver Optimization

Upon assuming that the estimated CSI is available at the transmitter with the aid of a perfect feedback channel, the transmitted signal is preprocessed by a channel-dependent TPC and the signal $\boldsymbol{y}$ received at the destination is of the following form [11]:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{F}\boldsymbol{s} + \boldsymbol{n}, \tag{19}$$

where $\boldsymbol{F}$ is the TPC matrix, $\boldsymbol{s}$ is the signal vector and $\boldsymbol{n}$ is the additive receiver noise having the covariance matrix $\boldsymbol{R_n} = \sigma_{\text{N}}^2\boldsymbol{I}$. The covariance matrix of $\boldsymbol{s}$ is assumed to be an identity matrix, i.e., $\mathbb{E}\{\boldsymbol{s}\boldsymbol{s}^{\text{H}}\} = \boldsymbol{I}$. In data transmission, based on the channel estimation error model of (4), the signal model of (19) used for data transmission is rewritten as

$$\boldsymbol{y} = \widehat{\boldsymbol{H}}\boldsymbol{F}\boldsymbol{s} + \Delta\boldsymbol{H}\boldsymbol{F}\boldsymbol{s} + \boldsymbol{n} = \widehat{\boldsymbol{H}}\boldsymbol{F}\boldsymbol{s} + \underbrace{\Delta\boldsymbol{H}_{\text{W}}\boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{F}\boldsymbol{s} + \boldsymbol{n}}_{\triangleq\boldsymbol{v}}, \tag{20}$$

---

[3]Channel estimation is performed invariably at high SNR.

where $\boldsymbol{v}$ is the equivalent noise that consists of the additive noise and the channel estimation error. The covariance matrix $\boldsymbol{R_v}$ of the equivalent noise $\boldsymbol{v}$ can be expressed as [25], [36]

$$\boldsymbol{R_v} = [\sigma_{\text{N}}^2 + \text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}\boldsymbol{\Phi})]\boldsymbol{I}. \tag{21}$$

As for the transceiver optimization, the receiver has the information including the estimated channel matrix $\widehat{\boldsymbol{H}}$ and the covariance matrix $\boldsymbol{R_v}$. Again, at the transmitter, we assume that either the estimated CSI $\widehat{\boldsymbol{H}}$ or only the statistical CSI given in (1) is available. In the following, both of these two cases are discussed. Note that similarly to training optimization, for the MIMO TPC designs there are also diverse performance metrics that correspond to distinctly different OFs [10], [11]. In Appendix A, the OFs considered in the following are derived in detail.

First, the TPC is designed by maximizing the capacity or MI. The corresponding optimization problem is given by [11]

$$\textbf{P. 4:} \max_{\boldsymbol{F}} \mathbb{E}\left\{\log\det\left(\boldsymbol{I} + \boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R_v}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F}\right)\right\}$$
$$\text{s.t.}\,\boldsymbol{R_v} = [\sigma_{\text{N}}^2 + \text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}\boldsymbol{\Phi})]\boldsymbol{I}, \text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}) \leq P_{\text{D}}, \tag{22}$$

where P_{D} is the maximum power during the data transmission. In **P. 4**, when the transmitter has the same CSI at the receiver, the expectation $\mathbb{E}\{\cdot\}$ can be removed. Otherwise, when only statistical CSI is available at the transmitter, $\mathbb{E}\{\cdot\}$ is performed over $\widehat{\boldsymbol{H}}$ for the optimization of $\boldsymbol{F}$. On the other hand, the TPC optimization aiming for minimizing the sum MSE may be written as [11]

$$\textbf{P. 5:} \min_{\boldsymbol{F}} \mathbb{E}\left\{\text{Tr}[(\boldsymbol{I} + \boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R_v}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F})^{-1}]\right\}$$
$$\text{s.t.}\,\boldsymbol{R_v} = [\sigma_{\text{N}}^2 + \text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}\boldsymbol{\Phi})]\boldsymbol{I}, \text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}) \leq P_{\text{D}}. \tag{23}$$

In order to extend the optimization **P. 4** to a general case, a weighted MI is formulated as [32]

$$\textbf{P. 6:} \max_{\boldsymbol{F}} \mathbb{E}\left\{\log\det\left(\boldsymbol{I} + \boldsymbol{A}^{\text{H}}\boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R_v}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F}\boldsymbol{A}\right)\right\}$$
$$\text{s.t.}\,\boldsymbol{R_v} = [\sigma_{\text{N}}^2 + \text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}\boldsymbol{\Phi})]\boldsymbol{I}, \text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}) \leq P_{\text{D}}, \tag{24}$$

where $\boldsymbol{A}$ is a complex matrix of appropriate dimensions. In **P. 6**, it can be observed that the weighting matrix $\boldsymbol{A}$ is applied to the matrix-valued SNR. Similarly, the optimization problem of weighted MSE minimization can be written as [10]

$$\textbf{P. 7:} \min_{\boldsymbol{F}} \mathbb{E}\left\{\text{Tr}[\boldsymbol{W}(\boldsymbol{I} + \boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R_v}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F})^{-1}]\right\}$$
$$\text{s.t.}\,\boldsymbol{R_v} = [\sigma_{\text{N}}^2 + \text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}\boldsymbol{\Phi})]\boldsymbol{I}, \text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}) \leq P_{\text{D}}, \tag{25}$$

where the positive semidefinite matrix $\boldsymbol{W}$ is the weighting matrix.

Generally speaking, for linear transceiver optimization having additively Schur-convex objectives, the problem can be represented as [10], [11], [32]

$$\textbf{P. 8:} \min_{\boldsymbol{F}} \mathbb{E}\left\{f_{\text{Convex}}^{\text{A-Schur}}\left(\mathbf{d}[(\boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R_v}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F} + \boldsymbol{I})^{-1}]\right)\right\}$$
$$\text{s.t.}\,\boldsymbol{R_v} = [\sigma_{\text{N}}^2 + \text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}\boldsymbol{\Phi})]\boldsymbol{I}, \text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}) \leq P_{\text{D}}, \tag{26}$$

where $f_{\text{Convex}}^{\text{A-Schur}}(\cdot)$ is an additively Schur-convex function of the vector consisting of the diagonal elements of $(\boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R_v}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F} + \boldsymbol{I})^{-1}$. The symbol $\mathbf{d}(\boldsymbol{Z})$ represents the

vector consisting of the diagonal elements of $\boldsymbol{Z}$. On the other hand, the linear transceiver optimization having additively Schur-concave objective is given by [10], [11], [32]

$$\textbf{P. 9:}\min_{\boldsymbol{F}}\mathbb{E}\left\{f_{\text{Concave}}^{\text{A--Schur}}\left(\mathbf{d}\!\left[(\boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R}_{\boldsymbol{v}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F}+\boldsymbol{I})^{-1}\right]\right)\right\}$$
$$\text{s.t.}\,\boldsymbol{R}_{\boldsymbol{v}}=[\sigma_{\text{N}}^{2}+\text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}\boldsymbol{\Phi})]\boldsymbol{I},\text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}})\leq P_{\text{D}},\quad (27)$$

where $f_{\text{Concave}}^{\text{A--Schur}}(\cdot)$ is an additively Schur-concave function of the vector consisting of the diagonal elements of $(\boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R}_{\boldsymbol{v}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F}+\boldsymbol{I})^{-1}$.

In order to accommodate the OFs in **P. 4** to **P. 9**, the unified general optimization problem is formulated as follows

$$\textbf{P. 10:}\max_{\boldsymbol{F}}f_{\text{unified}}\left(\boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R}_{\boldsymbol{v}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F}\right)$$
$$\text{s.t.}\,\boldsymbol{R}_{\boldsymbol{v}}=[\sigma_{\text{N}}^{2}+\text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}\boldsymbol{\Phi})]\boldsymbol{I},\text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}})\leq P_{\text{D}}.\quad (28)$$

It is worth noting that **P. 10** aims for maximizing a performance metric, while some optimization problems from the set of **P. 4** to **P. 9** aim for minimizing some performance metrics. In order to make the mathematical formulas consistent, when the optimization problem considered aims for minimizing a nonnegative function, its OF is replaced by its inverse function. When the optimization problem considered aims for minimizing a negative function, its OF is replaced by its negative counterpart. It should be highlighted that the expectation operation is contained in **P. 10**, and has either different mathematical meanings when the transmitter has either estimated CSI or statistical CSI. Specifically, when the transmitter relies on estimated CSI, the TPC $\boldsymbol{F}$ is optimized based on a specific realization of $\widehat{\boldsymbol{H}}$ and then the expectation operation is applied over the resultant OF. On the other hand, when the transmitter has only statistical information, the TPC $\boldsymbol{F}$ is optimized over the whole distribution of $\widehat{\boldsymbol{H}}$, instead of a realization of $\widehat{\boldsymbol{H}}$.

As for the TPC optimizations from the set **P. 4** to **P. 9**, when the estimated CSI is available at the transmitter, the optimal solutions are the Pareto optimal solutions of the following matrix-monotonic optimization [10]

$$\textbf{P. 11:}\max_{\boldsymbol{F}}\boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R}_{\boldsymbol{v}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F}$$
$$\text{s.t.}\,\boldsymbol{R}_{\boldsymbol{v}}=[\sigma_{\text{N}}^{2}+\text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}\boldsymbol{\Phi})]\boldsymbol{I},\text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}})\leq P_{\text{D}}.\quad (29)$$

On the other hand, when the transmitter only has statistical information, the optimal solutions of **P. 4** to **P. 9** aim for maximizing the distribution of $\boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R}_{\boldsymbol{v}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F}$. When $\widehat{\boldsymbol{H}}$ has only column correlations, maximizing the distribution of a random positive semidefinite matrix is equivalent to maximizing its expectation, i.e., $\boldsymbol{F}^{\text{H}}\mathbb{E}\{\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R}_{\boldsymbol{v}}^{-1}\widehat{\boldsymbol{H}}\}\boldsymbol{F}$ [38]. This is because when the distribution is maximized, for any given realization in the original distribution, it is always possible to find a realization from the optimized distribution, which is larger than the original realization and has the same probability density function (pdf). The detailed proof can be found in [38]. As a result, when $\widehat{\boldsymbol{H}}$ has only column correlations and the transmitter has only statistical CSI, the optimal solutions of

P. 4 to **P. 9** are the Pareto optimal solutions of the following matrix-monotonic optimization [32]

$$\textbf{P. 12:}\max_{\boldsymbol{F}}\boldsymbol{F}^{\text{H}}\mathbb{E}\{\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R}_{\boldsymbol{v}}^{-1}\widehat{\boldsymbol{H}}\}\boldsymbol{F}$$
$$\text{s.t.}\,\boldsymbol{R}_{\boldsymbol{v}}=[\sigma_{\text{N}}^{2}+\text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}}\boldsymbol{\Phi})]\boldsymbol{I},\text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}})\leq P_{\text{D}}.\quad (30)$$

P. 12 can be viewed as the optimization of the distribution of $\boldsymbol{F}^{\text{H}}\widehat{\boldsymbol{H}}^{\text{H}}\boldsymbol{R}_{\boldsymbol{v}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F}$. When multiple data streams are transmitted, the MIMO TPC optimization is a multi-objective optimization problem. Generally speaking, there is no solution that is optimal for all OFs. It is plausible that the channel estimation procedure has a grave impact on the TPC optimization. Note that in **P. 10** and **P. 11**, the positive definite matrix $\boldsymbol{\Phi}$ is a function of the training sequence $\boldsymbol{X}$. Moreover, it is also worth highlighting that the communication resources are limited, i.e., $(T - T_{\text{T}})P_{\text{D}} + P_{\text{T}}T_{\text{T}} \leq E_{\text{total}}$, where T is the channel's coherence time. The more resources are allocated to channel estimation, the more accurately the channel matrix can be estimated, but leaving less resources for information transmissions. As a result, there exist tradeoffs between the optimizations of **P. 3** and **P. 10**. The focus of our work is hence to jointly optimize these two matrix-monotonic optimization problems.

III. Unified Optimization of the Training Sequence and the TPC

Based on the discussions in Section II, when having perfect CSI at the transmitter, the matrix-monotonic optimization problem **P. 10** can be simplified to the following matrix-monotonic optimization [10]

$$\max_{\boldsymbol{F}}\,\boldsymbol{F}^{\text{H}}\boldsymbol{H}^{\text{H}}\boldsymbol{R}_{n}^{-1}\boldsymbol{H}\boldsymbol{F},\,\text{s.t. }\text{Tr}(\boldsymbol{F}\boldsymbol{F}^{\text{H}})\leq P_{\text{D}},\quad (31)$$

which is of the same form as **P. 3**. They try to recover the received signals from the noise-contaminated observations of the known training sequence for channel estimation and for data transmission by relying on the known channel matrix, respectively. This fact reveals that there exists a conceptual duality between the TPC optimization and training sequence optimization [10], albeit the constraints in **P. 3** and **P. 10** have slightly different physical meanings from a physical perspective. Specifically, the constraint in **P. 3** guarantees having limited total energy for channel estimation. On the other hand, the constraint in **P. 10** assures that the maximum power is lower than a specific threshold.

A. Unified Optimization Problems

For the joint optimization of the training sequence $\boldsymbol{X}$ and the TPC $\boldsymbol{F}$, a joint performance metric is a function of both $\boldsymbol{X}$ and $\boldsymbol{F}$. Again, here the training optimization is carried out by only using the statistical CSI. By contrast, the TPC optimization is carried out by using either the statistical CSI or the estimated CSI. As a result, for the joint optimization considered, the performance metric should be an average performance that is independent of the instantaneous CSI.

The corresponding unified joint optimization of the training sequence and the TPC is written in the following form

P. 13: $\displaystyle\max_{\boldsymbol{F},\boldsymbol{X},T_{\mathrm{T}}} \quad \frac{T-T_{\mathrm{T}}}{T} f_{\mathrm{unified}}\left(\frac{\boldsymbol{F}^{\mathrm{H}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\right)$

$$\text{s.t. } \boldsymbol{\Phi} = \left(\boldsymbol{\Psi}^{-1} + N_{\mathrm{R}}\boldsymbol{X}\boldsymbol{R}_{\boldsymbol{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\right)^{-1}$$
$$\mathrm{Tr}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}) \leq P_{\mathrm{D}}, \ \mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}) \leq P_{\mathrm{T}}T_{\mathrm{T}},$$
$$(T-T_{\mathrm{T}})P_{\mathrm{D}} + P_{\mathrm{T}}T_{\mathrm{T}} \leq E_{\mathrm{total}}. \tag{32}$$

In contrast to **P. 10**, there are three optimization variables in **P. 13**, namely $\boldsymbol{F}$, $\boldsymbol{X}$, and T_{T}. The discrete time interval T_{T} is the training length, which must satisfy $1 \leq T_{\mathrm{T}} \leq T-1$. The scalar T_{T} determines the amount of resources allocated to channel estimation and hence also the resources left for data transmission. After introducing T_{T} in the OF, **P. 13** aims for striking a compelling tradeoff between the channel estimation accuracy and data transmission efficiency, which is distinctly different from **P. 10**. In order to boldly differentiate the traditional designs given by **P. 10**, the performance metrics in **P. 13** given by **P. 4** to **P. 9** are termed as the effective metrics. For example, when considering the MI maximization problem **P. 4**, **P. 13** aims for maximizing the *effective MI* (termed as achievable rate in [25]) instead of the original MI. The corresponding OF is

$$\frac{T-T_{\mathrm{T}}}{T} f_{\mathrm{unified}}\left(\frac{\boldsymbol{F}^{\mathrm{H}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\right)$$
$$= \frac{T-T_{\mathrm{T}}}{T}\mathbb{E}\left\{\log\det\left(\boldsymbol{I} + \frac{\boldsymbol{F}^{\mathrm{H}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\right)\right\}. \tag{33}$$

The OF in (33) simultaneously maximizes the MI and data transmission time interval $T - T_{\mathrm{T}}$.

As discussed above, for the optimization problems **P. 4** to **P. 9**, there are several ones aiming for minimizing a specific performance metric. When these kinds of optimizations are unified in **P. 13**, the OFs are replaced by their corresponding inverse functions. For example, for the weighted MSE minimization of **P. 7**, the corresponding OF of **P. 13** is formulated as the reciprocal of the weighted MSE, i.e.,

$$\frac{T-T_{\mathrm{T}}}{T} f_{\mathrm{unified}}\left(\frac{\boldsymbol{F}^{\mathrm{H}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\right)$$
$$= \frac{T-T_{\mathrm{T}}}{T}\frac{1}{\mathbb{E}\left\{\mathrm{Tr}\left[\boldsymbol{W}\left(\boldsymbol{I} + \frac{\boldsymbol{F}^{\mathrm{H}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\right)^{-1}\right]\right\}}. \tag{34}$$

Here, similar to the concept of effective MI, the term

$$\frac{T}{T-T_{\mathrm{T}}}\mathbb{E}\left\{\mathrm{Tr}\left[\boldsymbol{W}\left(\boldsymbol{I} + \frac{\boldsymbol{F}^{\mathrm{H}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\right)^{-1}\right]\right\} \tag{35}$$

may be viewed as the effective weighted MSE. Minimizing the effective MSE aims for simultaneously minimizing the traditional MSE and maximizing the data transmission time interval.

B. Joint Optimization for Statistical CSI at the Transmitter

In this section, we investigate the case in which the transmitter only has statistical CSI but the destination has more accurate estimated CSI. Here we would like to highlight that in contrast to most of the existing studies on transceiver optimization [32], in the joint optimization of a linear TPC and training sequence, only the average performances are considered instead of their counterparts relying on the instantaneous CSI. Strictly speaking, only the upper bounds of the system performance are investigated. Following the rationale in [25], the expectation operations are moved into the OFs of **P. 4** to **P. 9**. The resultant joint optimization problem becomes

P. 14: $\displaystyle\max_{\boldsymbol{F},\boldsymbol{X},T_{\mathrm{T}}} \quad \frac{T-T_{\mathrm{T}}}{T} f_{\mathrm{unified}}\left(\frac{\boldsymbol{F}^{\mathrm{H}}\mathbb{E}\{\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\right)$

$$\text{s.t. } \boldsymbol{\Phi} = \left(\boldsymbol{\Psi}^{-1} + N_{\mathrm{R}}\boldsymbol{X}\boldsymbol{R}_{\boldsymbol{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\right)^{-1}$$
$$\mathrm{Tr}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}) \leq P_{\mathrm{D}}, \ \mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}) \leq P_{\mathrm{T}}T_{\mathrm{T}},$$
$$(T-T_{\mathrm{T}})P_{\mathrm{D}} + P_{\mathrm{T}}T_{\mathrm{T}} \leq E_{\mathrm{total}}. \tag{36}$$

In the OF of **P. 14**, the average matrix-valued SNR becomes:

$$\frac{\boldsymbol{F}^{\mathrm{H}}\mathbb{E}\{\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})} = \frac{\boldsymbol{F}^{\mathrm{H}}\boldsymbol{\Pi}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}, \tag{37}$$

with $\boldsymbol{\Pi} = N_{\mathrm{R}}\boldsymbol{\Psi} - N_{\mathrm{R}}\boldsymbol{\Phi}$.

It can be concluded that **P. 14** maximizes this matrix-valued SNR. As a result, our joint matrix-monotonic optimization is formulated as

P. 15: $\displaystyle\max_{\boldsymbol{F},\boldsymbol{X}} \quad \frac{\boldsymbol{F}^{\mathrm{H}}\boldsymbol{\Pi}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}$

$$\text{s.t.} \boldsymbol{\Pi} = N_{\mathrm{R}}\boldsymbol{\Psi} - N_{\mathrm{R}}\boldsymbol{\Phi}, \boldsymbol{\Phi} = \left(\boldsymbol{\Psi}^{-1} + N_{\mathrm{R}}\boldsymbol{X}\boldsymbol{R}_{\boldsymbol{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\right)^{-1},$$
$$\mathrm{Tr}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}) \leq P_{\mathrm{D}}, \ \mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}) \leq P_{\mathrm{T}}T_{\mathrm{T}},$$
$$(T-T_{\mathrm{T}})P_{\mathrm{D}} + P_{\mathrm{T}}T_{\mathrm{T}} \leq E_{\mathrm{total}}. \tag{38}$$

Explicitly this joint matrix-monotonic optimization is intrinsically different from traditional single matrix-monotonic optimization applied separately for transceiver design and training design, since a pair of optimization variables, namely the linear TPC $\boldsymbol{F}$ and the training sequence $\boldsymbol{X}$, are jointly optimized. The optimal solutions of **P. 14** belong to the Pareto-optimal solution set of **P. 15**. We would like to point out that **P. 13** optimizes the distribution of $\boldsymbol{F}^{\mathrm{H}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{R}_v^{-1}\widehat{\boldsymbol{H}}\boldsymbol{F}$ by relying on the statistical CSI available at transmitter. Provided that $\widehat{\boldsymbol{H}}$ has only column correlations, there exists an equivalence between optimizing the distribution of a random positive semidefinite matrix and maximizing its expectation, i.e., $\boldsymbol{F}^{\mathrm{H}}\mathbb{E}\{\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{R}_v^{-1}\widehat{\boldsymbol{H}}\}\boldsymbol{F}$ [38]. As a result, in **P. 14**, there is an approximation, since the expectation operation is moved into the OF. However, based on our discussion concerning **P. 12**, this approximation does not change the fact that the optimal solutions of **P. 13** belong to the Pareto-optimal solution set of **P. 15**. We can conjecture that the approximation in **P. 14** might change the correspondence between the optimal solutions of **P. 13** and the Pareto-optimal solutions of **P. 15**.

It is worth noting that the constraint $\mathrm{Tr}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}) \leq P_{\mathrm{D}}$ is equivalent to the following constraint

$$\sigma_{\mathrm{N}}^2\mathrm{Tr}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}) \leq \sigma_{\mathrm{N}}^2 P_{\mathrm{D}}, \tag{39}$$

which is also equivalent to the following one

$$\sigma_{\mathrm{N}}^2 \mathrm{Tr}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}) + P_{\mathrm{D}}\mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}) \le \sigma_{\mathrm{N}}^2 P_{\mathrm{D}} + P_{\mathrm{D}}\mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}). \tag{40}$$

The power constraint (40) is equivalent furthermore to the following constraint

$$\frac{\mathrm{Tr}[(\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}]}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})} \le P_{\mathrm{D}}. \tag{41}$$

Here, it is worth noting that the equivalence implies that, (41) can be proved and vice versa from (39). Therefore, based on the equivalence between (39) and (41), the joint matrix-monotonic optimization problem **P. 15** is equivalent to the following one

P. 16: $\displaystyle\max_{\boldsymbol{F},\boldsymbol{X}} \frac{\boldsymbol{F}^{\mathrm{H}}\boldsymbol{\Pi}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}$

$\text{s.t.} \boldsymbol{\Pi} = N_{\mathrm{R}}\boldsymbol{\Psi} - N_{\mathrm{R}}\boldsymbol{\Phi}, \ \boldsymbol{\Phi} = (\boldsymbol{\Psi}^{-1} + N_{\mathrm{R}}\boldsymbol{X}\boldsymbol{R}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}})^{-1},$

$\mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}) \le P_{\mathrm{T}}T_{\mathrm{T}}, \ \dfrac{\mathrm{Tr}[(\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}]}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\mathbf{FF}^{\mathrm{H}})} \le P_{\mathrm{D}},$

$(T - T_{\mathrm{T}})P_{\mathrm{D}} + P_{\mathrm{T}}T_{\mathrm{T}} \le E_{\mathrm{total}}. \tag{42}$

Upon defining the following new matrix variable

$$\widetilde{\boldsymbol{F}} = (\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})^{1/2}[\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})]^{-1/2}\boldsymbol{F}, \tag{43}$$

the optimization problem **P. 16** can be simplified into the following one

P. 17: $\displaystyle\max_{\widetilde{\boldsymbol{F}},\boldsymbol{X}} \ \widetilde{\boldsymbol{F}}^{\mathrm{H}}(\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})^{-1/2}(N_{\mathrm{R}}\boldsymbol{\Psi} - N_{\mathrm{R}}\boldsymbol{\Phi})$

$\times (\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})^{-1/2}\widetilde{\boldsymbol{F}}$

$\text{s.t.} \ \boldsymbol{\Phi} = (\boldsymbol{\Psi}^{-1} + N_{\mathrm{R}}\boldsymbol{X}\boldsymbol{R}_{\mathrm{N}}^{-1}\boldsymbol{X}^{\mathrm{H}})^{-1}$

$\mathrm{Tr}(\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}^{\mathrm{H}}) \le P_{\mathrm{D}}, \ \mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}) \le P_{\mathrm{T}}T_{\mathrm{T}},$

$(T - T_{\mathrm{T}})P_{\mathrm{D}} + P_{\mathrm{T}}T_{\mathrm{T}} \le E_{\mathrm{total}}. \tag{44}$

Based on the fundamental results of matrix-monotonic optimization [10], [32], we have the following theorem concerning the optimal $\widetilde{\boldsymbol{F}}$.

Theorem 1: The Pareto-optimal solution of $\widetilde{\boldsymbol{F}}$ for **P. 17** satisfies the following structure

$$\widetilde{\boldsymbol{F}}_{\mathrm{opt}} = \boldsymbol{V}_{\mathcal{H}}\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{F}}}\boldsymbol{U}_{\widetilde{\boldsymbol{F}}}^{\mathrm{H}}, \tag{45}$$

where $\boldsymbol{V}_{\mathcal{H}}$ and $\boldsymbol{U}_{\widetilde{\boldsymbol{F}}}$ are unitary matrices and $\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{F}}}$ is a diagonal matrix. The unitary matrix $\boldsymbol{V}_{\mathcal{H}}$ is defined based on the following SVD

$$(N\boldsymbol{\Psi} - N\boldsymbol{\Phi})^{\frac{1}{2}}(\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})^{-\frac{1}{2}} = \boldsymbol{U}_{\mathcal{H}}\boldsymbol{\Lambda}_{\mathcal{H}}\boldsymbol{V}_{\mathcal{H}}^{\mathrm{H}} \text{ with } \boldsymbol{\Lambda}_{\mathcal{H}} \searrow. \tag{46}$$

On the other hand, the unitary matrix $\boldsymbol{U}_{\widetilde{\boldsymbol{F}}}$ is determined by the specific OFs [10]. In the following, the optimal solutions of $\boldsymbol{U}_{\widetilde{\boldsymbol{F}}}$ are enumerated briefly and the detailed derivations are provided in [10].

Conclusion 3: For the optimization problems **P. 4** to **P. 9**, the corresponding optimal values of $\boldsymbol{U}_{\widetilde{\boldsymbol{F}}}$ are listed as follows:

P. 4: $\boldsymbol{U}_{\widetilde{\boldsymbol{F}},\mathrm{opt}} = \boldsymbol{U}_{\mathrm{Arb}}$; **P. 5:** $\boldsymbol{U}_{\widetilde{\boldsymbol{F}},\mathrm{opt}} = \boldsymbol{I}$; **P. 6:** $\boldsymbol{U}_{\widetilde{\boldsymbol{F}},\mathrm{opt}} = \boldsymbol{U}_{\boldsymbol{A}}$;
P. 7: $\boldsymbol{U}_{\widetilde{\boldsymbol{F}},\mathrm{opt}} = \boldsymbol{U}_{\boldsymbol{W}}$; **P. 8:** $\boldsymbol{U}_{\widetilde{\boldsymbol{F}},\mathrm{opt}} = \boldsymbol{U}_{\mathrm{DFT}}$; **P. 9:** $\boldsymbol{U}_{\widetilde{\boldsymbol{F}},\mathrm{opt}} = \boldsymbol{I}$, $\tag{47}$

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

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{H}} = \boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{U}_{\boldsymbol{A}}^{\mathrm{H}} \text{ with } \boldsymbol{\Lambda}_{\boldsymbol{A}} \searrow,$$

$$\boldsymbol{W} = \boldsymbol{U}_{\boldsymbol{W}}\boldsymbol{\Lambda}_{\boldsymbol{W}}\boldsymbol{U}_{\boldsymbol{W}}^{\mathrm{H}} \text{ with } \boldsymbol{\Lambda}_{\boldsymbol{W}} \searrow. \tag{48}$$

Moreover, $\boldsymbol{U}_{\mathrm{Arb}}$ denotes an arbitrary unitary matrix of appropriate dimensions and $\boldsymbol{U}_{\mathrm{DFT}}$ represents a DFT unitary matrix of suitable dimensions, respectively.

According to the definition of $\widetilde{\boldsymbol{F}}$ in (43), the following equality always holds

$$(\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})^{-1/2}\widetilde{\boldsymbol{F}} = [\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})]^{-1/2}\boldsymbol{F}, \tag{49}$$

based on which the following equality can be derived

$$\frac{1}{\sigma_{\mathrm{N}}^2 + \mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})} = \frac{\mathrm{Tr}[(\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})^{-\frac{1}{2}}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}^{\mathrm{H}}(\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})^{-\frac{1}{2}}]}{\mathrm{Tr}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}$$

$$= \frac{\mathrm{Tr}[(\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})^{-1}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}^{\mathrm{H}}]}{P_{\mathrm{D}}}, \tag{50}$$

where the second equality is due to the fact that for the optimal $\boldsymbol{F}$, we have $\mathrm{Tr}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}) = P_{\mathrm{D}}$. Finally, when the optimal solution $\widetilde{\boldsymbol{F}}$ has been computed, the optimal $\boldsymbol{F}$ equals

$$\boldsymbol{F}_{\mathrm{opt}} = \sqrt{\frac{P_{\mathrm{D}}}{\mathrm{Tr}[(\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})^{-1}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}^{\mathrm{H}}]}}(\sigma_{\mathrm{N}}^2\boldsymbol{I} + P_{\mathrm{D}}\boldsymbol{\Phi})^{-1/2}\widetilde{\boldsymbol{F}}_{\mathrm{opt}}. \tag{51}$$

Based on **Conclusion 1**, **Conclusion 2** and **Theorem 1**, we define the following variables and parameters

$$[\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{F}}}^{\mathrm{T}}\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{F}}}]_{i,i} = f_i^2, \ [\boldsymbol{\Lambda}_{\boldsymbol{\Psi}}]_{i,i} = \psi_i,$$

$$[\boldsymbol{\Lambda}_{\boldsymbol{X}}\boldsymbol{\Lambda}_{\boldsymbol{X}}^{\mathrm{T}}]_{i,i} = x_i^2, \ [\boldsymbol{\Lambda}_{\boldsymbol{R}_{\mathrm{N}}}^{-1}]_{i,i} = 1/\sigma_i^2. \tag{52}$$

The joint matrix-monotonic optimization problem **P. 17** is equivalent to the following one:

P. 18: $\displaystyle\max_{\{f_i^2\},\{x_i^2\}} \ \boldsymbol{U}_{\widetilde{\boldsymbol{F}}}\boldsymbol{\Lambda}_{\mathrm{IN}}\boldsymbol{U}_{\widetilde{\boldsymbol{F}}}^{\mathrm{H}}$

$\text{s.t.} \ [\boldsymbol{\Lambda}_{\mathrm{IN}}]_{i,i} = \dfrac{N_{\mathrm{R}}^2 f_i^2 x_i^2 \psi_i/\sigma_i^2}{\sigma_{\mathrm{N}}^2\psi_i^{-1} + N_{\mathrm{R}}\sigma_{\mathrm{N}}^2 x_i^2/\sigma_i^2 + P_{\mathrm{D}}}$

$\displaystyle\sum_{i=1}^{N_{\mathrm{Data}}} f_i^2 \le P_{\mathrm{D}}, \ \sum_i^{N_{\mathrm{Data}}} x_i^2 \le P_{\mathrm{T}}T_{\mathrm{T}},$

$(T - T_{\mathrm{T}})P_{\mathrm{D}} + P_{\mathrm{T}}T_{\mathrm{T}} \le E_{\mathrm{total}}. \tag{53}$

Based on this, the joint optimization problem **P. 14** can be further rewritten as

P. 19: $\displaystyle\max_{\{f_i^2\},\{x_i^2\},T_{\mathrm{T}}} f_{\mathrm{unified}}\left(\left\{\dfrac{N_{\mathrm{R}}^2 f_i^2 x_i^2 \psi_i/\sigma_i^2}{\sigma_{\mathrm{N}}^2\psi_i^{-1} + N_{\mathrm{R}}\sigma_{\mathrm{N}}^2 x_i^2/\sigma_i^2 + P_{\mathrm{D}}}\right\}_{i=1}^{N_{\mathrm{Data}}}\right)$

$\times \dfrac{T - T_{\mathrm{T}}}{T}$

$\text{s.t.} \ \displaystyle\sum_{i=1}^{N_{\mathrm{Data}}} f_i^2 \le P_{\mathrm{D}}, \ \sum_i^{N_{\mathrm{Data}}} x_i^2 \le P_{\mathrm{T}}T_{\mathrm{T}},$

$(T - T_{\mathrm{T}})P_{\mathrm{D}} + P_{\mathrm{T}}T_{\mathrm{T}} \le E_{\mathrm{total}}. \tag{54}$

The specific optimal solutions of **P. 19** are determined by the particular mathematical formulas of the OF in **P. 19**. In the following, two specific examples are given to show how to solve **P. 19**. Their methodology may also be applied to other OFs.

C. Specific Examples for Statistical CSI

In this subsection, some specific joint optimization problems are investigated one-by-one to show how to jointly optimize the training sequence and the transceiver.

Effective MI Maximization:

Based on the results derived above, the joint optimization problem **P. 14** is finally simplified into the following form for the optimization problem of effective MI maximization

$$\textbf{P. 20:} \max_{\{f_n^2\},\{x_n^2\},T_{\mathrm{T}}} \sum_{i=1}^{N_{\mathrm{Data}}} \log\left(1+\frac{N_{\mathrm{R}}^2 f_i^2 x_i^2 \psi_i/\sigma_i^2}{\sigma_{\mathrm{N}}^2\psi_i^{-1}+N_{\mathrm{R}}\sigma_{\mathrm{N}}^2 x_i^2/\sigma_i^2+P_{\mathrm{D}}}\right)$$
$$\times \frac{T-T_{\mathrm{T}}}{T}$$
$$\text{s.t.} \sum_{i=1}^{N_{\mathrm{Data}}} f_i^2 \le P_{\mathrm{D}}, \ \sum_{i=1}^{N_{\mathrm{Data}}} x_i^2 \le P_{\mathrm{T}}T_{\mathrm{T}},$$
$$(T-T_{\mathrm{T}})P_{\mathrm{D}}+P_{\mathrm{T}}T_{\mathrm{T}} \le E_{\mathrm{total}}. \tag{55}$$

P. 20 can be optimized in an alternating manner with respect to $\{f_n^2\}$, $\{x_n^2\}$, and T_{T}. When T_{T} is given, for fixed x_i^2 values, the optimal solutions of f_i^2 are the following water-filling solutions [39]

$$f_i^2 = \left(\frac{1}{\mu_f} - \frac{1}{h_i}\right)^+, \tag{56}$$

where μ_f is the Lagrange multiplier corresponding to the constraint $\sum_{i=1}^{N_{\mathrm{Data}}} f_i^2 \le P_{\mathrm{D}}$. The parameter h_i is defined as

$$h_i = \frac{N_{\mathrm{R}}^2 x_i^2 \psi_i/\sigma_i^2}{\sigma_{\mathrm{N}}^2\psi_i^{-1}+N_{\mathrm{R}}\sigma_{\mathrm{N}}^2 x_i^2/\sigma_i^2+P_{\mathrm{D}}}. \tag{57}$$

Then, for fixed f_i^2 values, the optimal solutions of x_i^2 are the following water-filling solutions:

$$x_i^2 = \left(\frac{-c_i(a_i+2b_i)+\sqrt{c_i^2(a_i+2b_i)^2-4(a_i+b_i)b_i(c_i^2-a_ic_i/\mu_x)}}{2(a_i+b_i)b_i}\right)^+, \tag{58}$$

where the parameters a_i, b_i and c_i are defined as follows

$$a_i = N_{\mathrm{R}}^2 f_i^2 \psi_i/\sigma_i^2, \ b_i = N_{\mathrm{R}}\sigma_{\mathrm{N}}^2/\sigma_i^2, \ c_i = \sigma_{\mathrm{N}}^2\psi_i^{-1}+P_{\mathrm{D}}. \tag{59}$$

Finally, the optimization of T_{T} can be carried out using a one-dimensional search [25] [23], given its discrete nature. The solution is illustrated in Algorithm 1. This procedure is quite similar to the effective MSE minimization. The outer loop represents the exhaustive search of T_{T}, while the inner loop is used for iteratively updating $\boldsymbol{X}_{\mathrm{P}}$, $\boldsymbol{F}$ until the convergence of the OF. The results are averaged over a number of random realizations to find the robust optimal training length and the optimal structure of both the training sequences and of the precoder.

Effective Weighted MSE Minimization

According to the previous discussions, the joint optimization problem **P. 14** aims for maximizing a performance metric. For the effective weighted MSE minimization, the OF of the joint

---

Algorithm 1 Joint Effective MI Maximization of Training Sequences and Transceivers Based on Matrix-Monotonic Optimization for Statistical CSI

1. Initialize the parameters and begin the exhaustive search from T_{T}=1.
repeat
  repeat
    2. Update f_i^2 with fixed x_i^2 in (56).
    3. Update x_i^2 with fixed f_i^2 in (58).
  until the convergence of OF in (55)
  4. $T_{\mathrm{T}} \leftarrow T_{\mathrm{T}} + 1$.
until $T_{\mathrm{T}} = T - 1$
5. Find the optimal training length T_{T}, training sequence as in Conclusion 1 and 2, and precoder as in Theorem 1 and (51).

---

optimization problem **P. 14** is formulated as the reciprocal of the effective weighted MSE, i.e.,

$$\frac{T-T_{\mathrm{T}}}{T} f_{\mathrm{unified}}\left(\frac{\boldsymbol{F}^{\mathrm{H}}\mathbb{E}\{\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2+\mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\right)$$
$$=\frac{T-T_{\mathrm{T}}}{T} \frac{1}{\mathrm{Tr}\left[\boldsymbol{W}\left(\boldsymbol{I}+\frac{\boldsymbol{F}^{\mathrm{H}}\mathbb{E}\{\widehat{\boldsymbol{H}}^{\mathrm{H}}\widehat{\boldsymbol{H}}\}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2+\mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\right)^{-1}\right]}, \tag{60}$$

based on which the joint optimization problem **P. 14** with respect to the effective weighted MSE performance finally becomes equivalent to

$$\textbf{P. 21:} \min_{\{f_n^2\},\{x_n^2\},T_{\mathrm{T}}} \frac{T}{T-T_{\mathrm{T}}} \sum_{i=1}^{N_{\mathrm{Data}}} \frac{w_i}{1+\frac{N_{\mathrm{R}}^2 f_i^2 x_i^2 \psi_i/\sigma_i^2}{\sigma_{\mathrm{N}}^2\psi_i^{-1}+N_{\mathrm{R}}\sigma_{\mathrm{N}}^2 x_i^2/\sigma_i^2+P_{\mathrm{D}}}}$$
$$\text{s.t.} \sum_{i=1}^{N_{\mathrm{Data}}} f_i^2 \le P_{\mathrm{D}}, \ \sum_{i=1}^{N_{\mathrm{Data}}} x_i^2 \le P_{\mathrm{T}}T_{\mathrm{T}},$$
$$(T-T_{\mathrm{T}})P_{\mathrm{D}}+P_{\mathrm{T}}T_{\mathrm{T}} \le E_{\mathrm{total}}. \tag{61}$$

P. 21 can also be optimized in an alternating manner with respect to $\{f_n^2\}$, $\{x_n^2\}$, and T_{T}. When T_{T} is given, for fixed x_i^2 values, the optimal solutions of f_i^2 are the following water-filling solutions:

$$f_i^2 = \left(\sqrt{\frac{w_i}{\mu_f h_i}} - \frac{1}{h_i}\right)^+, \tag{62}$$

where μ_f is the Lagrange multiplier corresponding to the constraint $\sum_{i=1}^{N_{\mathrm{Data}}} f_i^2 \le P_{\mathrm{D}}$. On the other hand, for fixed f_i^2 values, the optimal solutions of x_i^2 are the following water-filling solutions:

$$x_i^2 = \left(\frac{1}{a_i+b_i}\sqrt{\frac{w_i a_i c_i}{\mu_x}} - \frac{c_i}{a_i+b_i}\right)^+, \tag{63}$$

where a_i, b_i and c_i are defined in (59). Similar to effective MI maximziation, the optimization of T_{T} can be also performed using a one-dimensional search, given its discrete nature.

D. Joint Optimization using the Receiver's Estimated CSI at the Transmitter

In this section, we investigate the scenario, in which both the transmitter and the receiver have the same estimated CSI,

assuming that a feedback channel is available between them. The training optimization relies on the statistical information available, while the precoder is optimized based on the channel matrix estimated at the destination and the estimation error model. The final performance metric is averaged over all channel realizations. In this case, based on the above mathematical formulation along with our TPC optimization, the joint optimization problem can be further rewritten as follows:

$$
\textbf{P. 22:} \max_{\boldsymbol{F},\boldsymbol{X},T_{\mathrm{T}}} f_{\mathrm{unified}}\left(\frac{\boldsymbol{F}^{\mathrm{H}}(\boldsymbol{\Psi}-\boldsymbol{\Phi})^{\frac{1}{2}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}(\boldsymbol{\Psi}-\boldsymbol{\Phi})^{\frac{1}{2}}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2+\mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\right)
$$
$$
\times \frac{T-T_{\mathrm{T}}}{T}
$$
$$
\mathrm{s.t.}\,\boldsymbol{\Phi}=\left(\boldsymbol{\Psi}^{-1}+N_{\mathrm{R}}\boldsymbol{X}\boldsymbol{R}_{\boldsymbol{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\right)^{-1},
$$
$$
\mathrm{Tr}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})\le P_{\mathrm{D}},\,\mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}})\le P_{\mathrm{T}}T_{\mathrm{T}},
$$
$$
(T-T_{\mathrm{T}})P_{\mathrm{D}}+P_{\mathrm{T}}T_{\mathrm{T}}\le E_{\mathrm{total}}. \tag{64}
$$

This is equivalent to the following by replacing the sum power constraint $\mathrm{Tr}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})\le P_{\mathrm{D}}$ by (41)

$$
\textbf{P. 23:} \max_{\boldsymbol{F},\boldsymbol{X},T_{\mathrm{T}}} f_{\mathrm{unified}}\left(\frac{\boldsymbol{F}^{\mathrm{H}}(\boldsymbol{\Psi}-\boldsymbol{\Phi})^{\frac{1}{2}}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}(\boldsymbol{\Psi}-\boldsymbol{\Phi})^{\frac{1}{2}}\boldsymbol{F}}{\sigma_{\mathrm{N}}^2+\mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\right)
$$
$$
\times \frac{T-T_{\mathrm{T}}}{T}
$$
$$
\mathrm{s.t.}\,\boldsymbol{\Phi}=\left(\boldsymbol{\Psi}^{-1}+N_{\mathrm{R}}\boldsymbol{X}\boldsymbol{R}_{\boldsymbol{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\right)^{-1},
$$
$$
\frac{\mathrm{Tr}[(\sigma_{\mathrm{N}}^2\boldsymbol{I}+P_{\mathrm{D}}\boldsymbol{\Phi})\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}]}{\sigma_{\mathrm{N}}^2+\mathrm{Tr}(\boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}})}\le P_{\mathrm{D}},\,\mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}})\le P_{\mathrm{T}}T_{\mathrm{T}},
$$
$$
(T-T_{\mathrm{T}})P_{\mathrm{D}}+P_{\mathrm{T}}T_{\mathrm{T}}\le E_{\mathrm{total}}. \tag{65}
$$

Similar to **P. 17**, based on the definition of $\widetilde{\boldsymbol{F}}$ in (43), **P. 23** can further be reformulated as

$$
\textbf{P. 24:} \max_{\widetilde{\boldsymbol{F}},\boldsymbol{X},T_{\mathrm{T}}} \frac{T-T_{\mathrm{T}}}{T} f_{\mathrm{unified}}\big(\widetilde{\boldsymbol{F}}^{\mathrm{H}}(\sigma_{\mathrm{N}}^2\boldsymbol{I}+P_{\mathrm{D}}\boldsymbol{\Phi})^{-\frac{1}{2}}(\boldsymbol{\Psi}-\boldsymbol{\Phi})^{\frac{1}{2}}
$$
$$
\times \boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}(\boldsymbol{\Psi}-\boldsymbol{\Phi})^{\frac{1}{2}}(\sigma_{\mathrm{N}}^2\boldsymbol{I}+P_{\mathrm{D}}\boldsymbol{\Phi})^{-\frac{1}{2}}\widetilde{\boldsymbol{F}}\big)
$$
$$
\mathrm{s.t.}\,\boldsymbol{\Phi}=\left(\boldsymbol{\Psi}^{-1}+N_{\mathrm{R}}\boldsymbol{X}\boldsymbol{R}_{\boldsymbol{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\right)^{-1}
$$
$$
\mathrm{Tr}(\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}^{\mathrm{H}})\le P_{\mathrm{D}},\,\mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}})\le P_{\mathrm{T}}T_{\mathrm{T}},
$$
$$
(T-T_{\mathrm{T}})P_{\mathrm{D}}+P_{\mathrm{T}}T_{\mathrm{T}}\le E_{\mathrm{total}}. \tag{66}
$$

By exploiting that all elements of $\boldsymbol{H}_{\mathrm{W}}$ are i.i.d. random variables, **P. 24** is equivalent to

$$
\textbf{P. 25:} \max_{\widetilde{\boldsymbol{F}},\boldsymbol{X},T_{\mathrm{T}}} \frac{T-T_{\mathrm{T}}}{T} f_{\mathrm{unified}}\big(\widetilde{\boldsymbol{F}}^{\mathrm{H}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{F}}\big)
$$
$$
\mathrm{s.t.}\,\boldsymbol{\Sigma}=(\sigma_{\mathrm{N}}^2\boldsymbol{I}+P_{\mathrm{D}}\boldsymbol{\Phi})^{-\frac{1}{2}}(\boldsymbol{\Psi}-\boldsymbol{\Phi})(\sigma_{\mathrm{N}}^2\boldsymbol{I}+P_{\mathrm{D}}\boldsymbol{\Phi})^{-\frac{1}{2}},
$$
$$
\boldsymbol{\Phi}=\left(\boldsymbol{\Psi}^{-1}+N_{\mathrm{R}}\boldsymbol{X}\boldsymbol{R}_{\boldsymbol{N}}^{-1}\boldsymbol{X}^{\mathrm{H}}\right)^{-1}
$$
$$
\mathrm{Tr}(\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}^{\mathrm{H}})\le P_{\mathrm{D}},\,\mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}})\le P_{\mathrm{T}}T_{\mathrm{T}},
$$
$$
(T-T_{\mathrm{T}})P_{\mathrm{D}}+P_{\mathrm{T}}T_{\mathrm{T}}\le E_{\mathrm{total}}. \tag{67}
$$

Based on our matrix-monotonic optimization framework, the OF of **P. 25** equals

$$
f_{\mathrm{unified}}\big(\widetilde{\boldsymbol{F}}^{\mathrm{H}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{F}}\big)
$$
$$
=f_{\mathrm{unified}}\left(\left\{f_i^2\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})\right\}_{i=1}^{N_{\mathrm{Data}}}\right), \tag{68}
$$

where the elements of the diagonal matrix $\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}$ are defined as

$$
[\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}]_{i,i}=\frac{N_{\mathrm{R}}x_i^2\psi_i/\sigma_i^2}{\sigma_{\mathrm{N}}^2\psi_i^{-1}+N_{\mathrm{R}}\sigma_{\mathrm{N}}^2x_i^2/\sigma_i^2+P_{\mathrm{D}}}. \tag{69}
$$

The equality in (68) exploits the fact that for any unitary matrix $\boldsymbol{U}$, $\boldsymbol{H}_{\mathrm{W}}$ and $\boldsymbol{H}_{\mathrm{W}}\boldsymbol{U}$ have the same distribution. Therefore, based on (68) the joint optimization of our linear TPC and training sequence is formulated as

$$
\textbf{P. 26:} \max_{\{f_i^2\},\{x_i^2\},T_{\mathrm{T}}} f_{\mathrm{unified}}\left(\left\{f_i^2\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})\right\}_{i=1}^{N_{\mathrm{Data}}}\right)
$$
$$
\times \frac{T-T_{\mathrm{T}}}{T}
$$
$$
\mathrm{s.t.}\,[\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}]_{i,i}=\frac{N_{\mathrm{R}}x_i^2\psi_i/\sigma_i^2}{\sigma_{\mathrm{N}}^2\psi_i^{-1}+N_{\mathrm{R}}\sigma_{\mathrm{N}}^2x_i^2/\sigma_i^2+P_{\mathrm{D}}},
$$
$$
\sum_{i=1}^{N_{\mathrm{Data}}}f_i^2\le P_{\mathrm{D}},\,\sum_{i=1}^{N_{\mathrm{Data}}}x_i^2\le P_{\mathrm{T}}T_{\mathrm{T}},
$$
$$
(T-T_{\mathrm{T}})P_{\mathrm{D}}+P_{\mathrm{T}}T_{\mathrm{T}}\le E_{\mathrm{total}}. \tag{70}
$$

In the case when the transmitter has estimated CSI, for a given transmit power P_{D}, the optimal solution of f_i^2 can be expressed as a function of P_{D} and $\left\{\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})\right\}$, i.e.,

$$
f_i^2=p_i\left(P_{\mathrm{D}},\left\{\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})\right\}_{i=1}^{N_{\mathrm{Data}}}\right). \tag{71}
$$

As a result, the optimization problem **P. 26** may also be shown to be equivalent to

$$
\textbf{P. 27:} \max_{\{x_i^2\},T_{\mathrm{T}}} f_{\mathrm{unified}}\left(\left\{f_i^2\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})\right\}_{i=1}^{N_{\mathrm{Data}}}\right)
$$
$$
\times \frac{T-T_{\mathrm{T}}}{T}
$$
$$
\mathrm{s.t.}\,[\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}]_{i,i}=\frac{N_{\mathrm{R}}x_i^2\psi_i/\sigma_i^2}{\sigma_{\mathrm{N}}^2\psi_i^{-1}+N_{\mathrm{R}}\sigma_{\mathrm{N}}^2x_i^2/\sigma_i^2+P_{\mathrm{D}}}
$$
$$
f_i^2=p_i\left(P_{\mathrm{D}},\left\{\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})\right\}_{i=1}^{N_{\mathrm{Data}}}\right)
$$
$$
\sum_{i=1}^{N_{\mathrm{Data}}}x_i^2\le P_{\mathrm{T}}T_{\mathrm{T}},
$$
$$
(T-T_{\mathrm{T}})P_{\mathrm{D}}+P_{\mathrm{T}}T_{\mathrm{T}}\le E_{\mathrm{total}}. \tag{72}
$$

It is worth highlighting that the statistical expectation operation involved in **P. 27** makes the optimization problem difficult to solve. To overcome this difficulty, a pair of algorithms are proposed. For the first one, the power allocation in the training optimization adopts the suboptimal solutions given by **P. 1** or **P. 2**. As a result, in **P. 27** there is only a single real scalar optimization variable T_{T} that can be optimized by using a simple one-dimensional search. The expectation operation in **P. 27** can be carried out numerically. On the other hand, the second one is based on the approximations that replace the eigenvalues $\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})$ in **P. 27** by $\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\mathbb{E}\{\boldsymbol{H}_{\mathrm{W}}^{\mathrm{H}}\boldsymbol{H}_{\mathrm{W}}\}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})$.

E. Specific Examples for Estimated CSI

Effective MI Maximization:

In this subsection, we investigate the effective MI maximization in detail, where the MI term in the OF of **P. 27** equals

$$f_{\text{unified}}\left(\left\{f_i^2\lambda_i\left(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\right)\right\}_{i=1}^{N_{\text{Data}}}\right)$$
$$=\mathbb{E}\left\{\sum_{i=1}^{N_{\text{Data}}}\log\left(1+f_i^2\lambda_i\left(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\right)\right)\right\}. \quad (73)$$

It is widely exploited that for the MI maximization the optimal f_i^2 obeys [10], [11]

$$f_i^2 = \left(\frac{1}{\mu_f} - \frac{1}{\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})}\right)^+. \quad (74)$$

The Lagrange multiplier μ_f is equivalent to

$$\frac{1}{\mu_f} = \frac{P_{\text{D}}+\sum_{i=1}^{\widetilde{N}_{\text{Data}}}\frac{1}{\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})}}{\widetilde{N}_{\text{Data}}}, \quad (75)$$

where $\widetilde{N}_{\text{Data}}$ is the number of eigen-channels allocated non-zero powers during the data transmission phase. Therefore, substituting the value of μ_f into the OF, the following result holds

$$f_{\text{unified}}\left(\left\{f_i^2\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})\right\}_{i=1}^{N_{\text{Data}}}\right)$$
$$=\mathbb{E}\left\{\sum_{i=1}^{\widetilde{N}_{\text{Data}}}\log\left(\frac{P_{\text{D}}+\sum_{j=1}^{\widetilde{N}_{\text{Data}}}\frac{1}{\lambda_j(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{\frac{1}{2}})}}{\widetilde{N}_{\text{Data}}}\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{\frac{1}{2}})\right)\right\} \quad (76)$$

Based on (76), the effective MI maximization problem is written in the following form

P. 28: $\max\limits_{\{x_i^2\},T_{\text{T}}} \mathbb{E}\left\{\sum_{i=1}^{\widetilde{N}_{\text{Data}}}\log\left(\frac{P_{\text{D}}+\sum_{j=1}^{\widetilde{N}_{\text{Data}}}\frac{1}{\lambda_j(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})}}{\widetilde{N}_{\text{Data}}}\right.\right.$

$$\left.\left.\times\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})\right)\right\}\frac{T-T_{\text{T}}}{T}$$
$$\text{s.t.}\sum_{i=1}^{N_{\text{Data}}}x_i^2 \leq P_{\text{T}}T_{\text{T}},$$
$$(T-T_{\text{T}})P_{\text{D}}+P_{\text{T}}T_{\text{T}} \leq E_{\text{total}},$$
$$\left[\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}\right]_{i,i} = \frac{N_{\text{R}}x_i^2\psi_i/\sigma_i^2}{\sigma_{\text{N}}^2\psi_i^{-1}+N_{\text{R}}\sigma_{\text{N}}^2x_i^2/\sigma_i^2+P_{\text{D}}}. \quad (77)$$

Because of the statistical expectation operation in **P. 28**, it is challenging to derive the optimal solutions of **P. 28** in closed forms. In the following two kinds of algorithms are proposed.

In order to avoid complex optimizations involving statistical expectations, some suboptimal solutions are used for $\{x_i^2\}$. Specifically, since channel estimation is usually performed at high SNRs, for MI maximization, the resources are assumed to be allocated uniformly among $\{x_i^2\}$, i.e., $\{x_i^2 =$

$P_{\text{T}}T_{\text{T}}/N_{\text{Data}}\}$. Therefore, **P. 28** is simplified to the following problem:

P. 29: $\max\limits_{T_{\text{T}}}\mathbb{E}\left\{\sum_{i=1}^{\widetilde{N}_{\text{Data}}}\log\left(\frac{P_{\text{D}}+\sum_{j=1}^{\widetilde{N}_{\text{Data}}}\frac{1}{\lambda_j(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})}}{\widetilde{N}_{\text{Data}}}\right.\right.$

$$\left.\left.\times\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})\right)\right\}\frac{T-T_{\text{T}}}{T}$$
$$\text{s.t.}\,x_i^2 = P_{\text{T}}T_{\text{T}}/N_{\text{Data}},$$
$$(T-T_{\text{T}})P_{\text{D}}+P_{\text{T}}T_{\text{T}} \leq E_{\text{total}},$$
$$\left[\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}\right]_{i,i} = \frac{N_{\text{R}}x_i^2\psi_i/\sigma_i^2}{\sigma_{\text{N}}^2\psi_i^{-1}+N_{\text{R}}\sigma_{\text{N}}^2x_i^2/\sigma_i^2+P_{\text{D}}}. \quad (78)$$

The second rationale is to replace the eigenvalues $\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})$ by their expectations, i.e., by $\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\mathbb{E}\{\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})$. As a result, **P. 28** can be approximated as follows:

P. 30: $\max\limits_{\{x_i^2\},T_{\text{T}}}\sum_{i=1}^{\widetilde{N}_{\text{Data}}}\log\left(\frac{P_{\text{D}}+\sum_{j=1}^{\widetilde{N}_{\text{Data}}}\frac{1}{\lambda_j(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\mathbb{E}\{\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})}}{\widetilde{N}_{\text{Data}}}\right.$

$$\left.\times\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\mathbb{E}\{\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})\right)\frac{T-T_{\text{T}}}{T}$$
$$\text{s.t.}\sum_{i=1}^{N_{\text{Data}}}x_i^2 \leq P_{\text{T}}T_{\text{T}},$$
$$(T-T_{\text{T}})P_{\text{D}}+P_{\text{T}}T_{\text{T}} \leq E_{\text{total}},$$
$$\left[\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}\right]_{i,i} = \frac{N_{\text{R}}x_i^2\psi_i/\sigma_i^2}{\sigma_{\text{N}}^2\psi_i^{-1}+N_{\text{R}}\sigma_{\text{N}}^2x_i^2/\sigma_i^2+P_{\text{D}}}. \quad (79)$$

It is worth noting that this approximation is accurate at high SNRs as shown in Section IV, but the proof is omitted due to space limitation. The optimization problem **P. 30** can be efficiently solved via alternating optimization, where the solution of $\{x_i^2\}$ may be found via the MATLAB function "*fmincon*".

Effective Weighted MSE Minimization:

In this subsection, the effective weighted MSE minimization is investigated. For the effective weighted MSE minimization, the MSE term of the OF of **P. 27** may be expressed as

$$f_{\text{unified}}\left(\left\{f_i^2\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})\right\}_{i=1}^{N_{\text{Data}}}\right)$$
$$=\left[\mathbb{E}\left\{\frac{\left(\sum_{i=1}^{\widetilde{N}_{\text{Data}}}\frac{\sqrt{w_i}}{\sqrt{\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})}}\right)^2}{P_{\text{D}}+\sum_{i=1}^{\widetilde{N}_{\text{Data}}}\frac{1}{\lambda_i(\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{H}_{\text{W}}^{\text{H}}\boldsymbol{H}_{\text{W}}\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}^{1/2})}}\right\}+\sum_{i>\widetilde{N}_{\text{Data}}}\omega_i\right]^{-1}, \quad (80)$$

based on which **P. 27** is further reformulated as follows:

$$\textbf{P. 31:} \max_{\{x_i^2\}, T_T} \left[\mathbb{E}\left\{ \frac{\left(\sum_{i=1}^{\widetilde{N}_{\text{Data}}} \frac{\sqrt{w_i}}{\sqrt{\lambda_i(\mathbf{\Lambda_\Sigma}^{1/2} \mathbf{H}_W^H \mathbf{H}_W \mathbf{\Lambda_\Sigma}^{1/2})}}\right)^2}{P_D + \sum_{i=1}^{\widetilde{N}_{\text{Data}}} \frac{1}{\lambda_i(\mathbf{\Lambda_\Sigma}^{1/2} \mathbf{H}_W^H \mathbf{H}_W \mathbf{\Lambda_\Sigma}^{1/2})}} \right\} + \sum_{i>\widetilde{N}_{\text{Data}}} \omega_i \right]^{-1}$$

$$\times \frac{T - T_T}{T}$$

$$\text{s.t. } \sum_{i=1}^{N_{\text{Data}}} x_i^2 \leq P_T T_T, (T-T_T)P_D + P_T T_T \leq E_{\text{total}},$$

$$\left[\mathbf{\Lambda_\Sigma}\right]_{i,i} = \frac{N_R x_i^2 \psi_i/\sigma_i^2}{\sigma_N^2 \psi_i^{-1} + N_R \sigma_N^2 x_i^2/\sigma_i^2 + P_D}. \quad (81)$$

It is worth noting that solving the optimization problem **P. 31** is still very challenging due to the statistical expectation operation. In the following, a pair of approximations are used.

Firstly, similar to **P. 29**, in order to avoid complex mathematical operations, suboptimal power allocation schemes may be applied to $\{x_i^2\}$. At high SNR, the resources are supposed to be allocated proportionally among $\{x_i^2\}$ for sum MSE minimization. Since we are considering the average performance, the suboptimal uniform power allocation scheme is adopted for $\{x_i^2\}$ in the following for simplicity, i.e., $\{x_i^2 = P_T T_T/N_{\text{Data}}\}$. Thus, **P. 31** is simplified to

$$\textbf{P. 32:} \min_{T_T} \frac{T}{T - T_T} \mathbb{E}\left\{ \frac{\left(\sum_{i=1}^{\widetilde{N}_{\text{Data}}} \frac{\sqrt{w_i}}{\sqrt{\lambda_i(\mathbf{\Lambda_\Sigma}^{1/2} \mathbf{H}_W^H \mathbf{H}_W \mathbf{\Lambda_\Sigma}^{1/2})}}\right)^2}{P_D + \sum_{i=1}^{\widetilde{N}_{\text{Data}}} \frac{1}{\lambda_i(\mathbf{\Lambda_\Sigma}^{1/2} \mathbf{H}_W^H \mathbf{H}_W \mathbf{\Lambda_\Sigma}^{1/2})}} \right\}$$

$$+ \frac{T}{T - T_T} \sum_{i>\widetilde{N}_{\text{Data}}} \omega_i$$

$$\text{s.t.} \left[\mathbf{\Lambda_\Sigma}\right]_{i,i} = \frac{N_R x_i^2 \psi_i/\sigma_i^2}{\sigma_N^2 \psi_i^{-1} + N_R \sigma_N^2 x_i^2/\sigma_i^2 + P_D},$$

$$x_i^2 = \frac{P_T T_T}{N_{\text{Data}}}, (T-T_T)P_D + P_T T_T \leq E_{\text{total}}. \quad (82)$$

In **P. 32**, the single optimization variable T_T can be found by a one-dimensional search.

In the second solution, the eigenvalues $\lambda_i(\mathbf{\Lambda_\Sigma}^{1/2} \mathbf{H}_W^H \mathbf{H}_W \mathbf{\Lambda_\Sigma}^{1/2})$ are replaced by their expectations, i.e., by $\lambda_i(\mathbf{\Lambda_\Sigma}^{1/2} \mathbb{E}\{\mathbf{H}_W^H \mathbf{H}_W\} \mathbf{\Lambda_\Sigma}^{1/2})$. Thus, **P. 31** is relaxed to

$$\textbf{P. 33:} \min_{\{x_i^2\}, T_T} \frac{T}{T - T_T} \frac{\left(\sum_{i=1}^{\widetilde{N}_{\text{Data}}} \frac{\sqrt{w_i}}{\sqrt{\lambda_i(\mathbf{\Lambda_\Sigma}^{1/2} \mathbb{E}\{\mathbf{H}_W^H \mathbf{H}_W\} \mathbf{\Lambda_\Sigma}^{1/2})}}\right)^2}{P_D + \sum_{i=1}^{\widetilde{N}_{\text{Data}}} \frac{1}{\lambda_i(\mathbf{\Lambda_\Sigma}^{1/2} \mathbb{E}\{\mathbf{H}_W^H \mathbf{H}_W\} \mathbf{\Lambda_\Sigma}^{1/2})}}$$

$$+ \frac{T}{T - T_T} \sum_{i>\widetilde{N}_{\text{Data}}} \omega_i$$

$$\text{s.t. } \sum_{i=1}^{N_{\text{Data}}} x_i^2 \leq P_T T_T, (T-T_T)P_D + P_T T_T \leq E_{\text{total}}$$

$$\left[\mathbf{\Lambda_\Sigma}\right]_{i,i} = \frac{N_R x_i^2 \psi_i/\sigma_i^2}{\sigma_N^2 \psi_i^{-1} + N_R \sigma_N^2 x_i^2/\sigma_i^2 + P_D}. \quad (83)$$

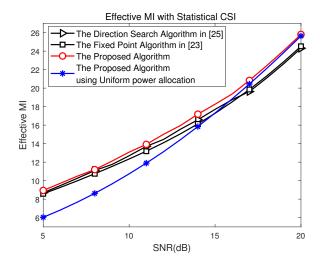

Fig. 3. Joint optimization for effective MI maximization using statistical CSI at the transmitter, where the antenna setting is 8×8.

The optimization problem **P. 33** can be solved efficiently by alternating between $\{x_i^2\}$ and T_T, where the solution of $\{x_i^2\}$ may be found by the MATLAB function "*fmincon*".

F. Complexity Analysis

For the proposed algorithms, the computational complexity mainly arises from the matrix decompositions, water-filling algorithms, numbers of iterations, one-dimensional search over the training interval T_T and the numerical computations of the statistical expectation. For an $N \times N$ matrix, the complexity of its matrix decomposition is on the order of $\mathcal{O}(N^3)$. Moreover, for an N-dimensional separate optimization problem, the complexity of the water-filling algorithm is as low as $\mathcal{O}(N)$. We will demonstrate in Section IV that typically only one or two iterations may be required in Fig. 8 and Fig. 9. The one-dimensional search will increase the computational complexity at an order of $\mathcal{O}(T_T)$. For the following complexity analysis, L_S denotes the number of independent trials used for the numerical computations of the statistical expectation. Upon using statistical CSI at the transmitter, the total computational complexities of the proposed algorithms are $\mathcal{O}(N^3) + \mathcal{O}(NT_T + L_S T_T)$. On the other hand, in the case of estimated CSI at transmitter, the total computational complexities of the proposed algorithms using the eigenvalue approximations are also $\mathcal{O}(N^3) + \mathcal{O}(NT_T + L_S T_T)$. However, in this case the total computational complexities of the proposed algorithms using uniform power allocation for training are a little bit different and the total complexity is $\mathcal{O}(N^3) + \mathcal{O}(L_S T_T)$.

IV. SIMULATION RESULTS AND DISCUSSIONS

In this section, several numerical results are provided for the proposed joint optimization of our linear TPC and training sequence. Firstly, the widely used exponential correlation model is used for $\mathbf{\Psi}$, i.e., $[\mathbf{\Psi}]_{i,j} = \theta^{|i-j|}$ [10], [36]. In the following simulations, $\theta = 0.9$ is chosen, and the block length is 256. For the cases in which the receiver relies on the estimated CSI and the transmitter has only statistical CSI,
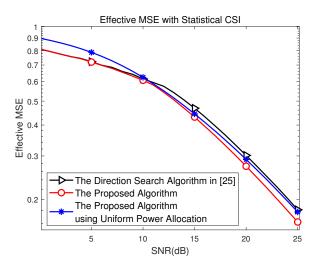
Fig. 4. Joint optimization for effective MSE minimization using statistical CSI at the transmitter, where the antenna setting is 8×8.
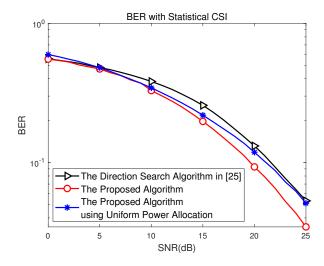

Fig. 5. The BER performance of joint optimization for effective MSE minimization using statistical CSI at the transmitter, where the antenna setting is 8×8.

the joint optimization used for effective MI maximization is investigated first. In Fig. 3, the effective MI represents the values of the OF of **P. 20**, with antenna settings of $N_T = N_R = N = 8$ and $T = 256$ vs the SNR. In order to characterize the performance of the proposed algorithm, the direction search algorithm of [25] and the fixed point algorithm of [23] are also used in Fig. 3. At low SNRs, the effective MI of the proposed algorithm only has modest advantages over the direction search algorithm of [25] and the fixed point algorithm of [23]. However, there is a significant performance gap between the proposed algorithm and the algorithm using uniform power allocation, which shows the superiority of the proposed optimization algorithm. With the SNR increasing, the performance gap between the proposed algorithm and the algorithm employing uniform power allocation is reducing, which is due to the fact that the optimal power allocation required for effective MI maximization tends to be uniform at high SNRs. Furthermore, the effective MI of the proposed algorithm is much higher than that of the direction search algorithm of [25] and that of the fixed point algorithm of [23] at high SNRs.

In contrast to the MI, the MSE exhibits higher signal recovery accuracy at the receiver. If we directly use MSE as our performance metric, the trivial conclusion emerges that all the resources should be allocated to channel estimation. To overcome this deficiency, the OF of the optimization problem **P. 21** is the effective MSE instead of the traditional MSE used for transceiver optimization. The effective MSE minimization aims for simultaneously minimizing the data estimation MSE and maximizing the data transmission time interval. In Fig. 4, we compare the effective MSEs of the direction search algorithm [25], of the proposed algorithm, and of the proposed algorithm using uniform power allocation, but only statistical CSI at the transmitter. At high SNRs, the effective MSE of the proposed algorithm is better than that of the direction search algorithm of [25]. Additionally, the effective MSE of the proposed algorithm is always lower than that of the proposed algorithm employing uniform power allocation.
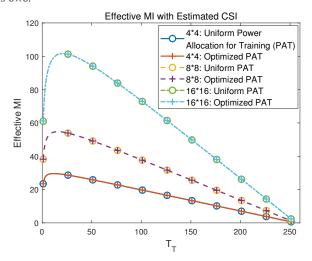

Fig. 6. Joint optimization for effective MI maximization using estimated CSI at transmitter with different antenna number settings when SNR is 30 dB.

In practical communication systems the bit-error rate (BER) is one of the salient performance criteria. However, its analytical expression is generally quite challenging to derive. Since the mathematically tractable MSE metric is closely related to the BER, Fig. 5 illustrates the BER achieved by the MSE solution for the system considered. Observe that at sufficiently high SNRs our solution attains beneficial BER gains over both the direction search algorithm of [25] and over that using uniform power allocation.

In the simulations, the other case is also taken into account, in which the transmitter and receiver have the same estimated CSI. In this case, the optimization problems **P. 29** and **P. 30** are used for the effective MI maximization. In Fig. 6, the effective MIs are demonstrated for $T = 256$ at SNR= 30dB and for different antenna settings of $N = 4, 8, 16$. Each point of the curves in Fig. 6 is an average of 10^4 independent realizations used for calculating the statistical expectations. It can be concluded from the numerical results that there always exist a best operating point for resource allocation between channel estimation and data transmission. Moreover, it can
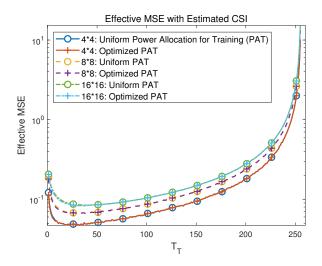
Fig. 7. Joint Optimization for effective MSE minimization using estimated CSI at transmitter with different antenna number settings when SNR is 30 dB.

Fig. 8. The convergence behavior for effective MI maximization using statistical CSI at transmitter, where the antenna setting is 8×8 under different SNR situations and training lengths.

Fig. 9. The convergence behavior for effective MSE minimization using statistical CSI at transmitter, where the antenna setting is 8×8 under different SNR situations and training lengths.

be seen that for the optimal operating point, the resources allocated to channel estimation are much lower than those allocated for data transmission. Furthermore, Fig. 6 shows that the suboptimal solution in **P. 29** using uniform power allocation for training has almost the same performance as the optimized power allocation employed for training in **P. 30**. Similar conclusions can be drawn for the effective MSE minimization characterized in Fig. 7. For the effective MSE minimization, the optimization problems **P. 32** and **P. 33** are considered. It can also be seen that the suboptimal solution in **P. 32** associated with uniform power allocation for training has almost the same performance as the optimized power allocation delivered for training in **P. 33**.

Last but not the least, the convergence behaviors of the proposed algorithms are investigated. It is worth highlighting that for the joint optimization using estimated CSI at the transmitter, the TPC is derived to be a function of the training sequence. In turn, this function is substituted into the original OF. As a result, there is no convergence issue that should be investigated. For the case of statistical CSI at the transmitter, the convergence behaviors of the proposed algorithm are shown in Fig. 8 and Fig. 9. It is observed that for both effective MI maximization and effective MSE minimization, the proposed algorithms exhibit excellent pretty good convergence properties for all simulation settings.

V. CONCLUSIONS

The joint optimization of linear transceivers and training sequences was investigated for various performance metrics, including the effective MI, effective MSE, effective weighted MI, effective weighted MSE, effective additively Schur-convex and effective additively Schur-concave functions. The new joint matrix-monotonic optimization framework was proposed, which has two matrix variables, i.e., the linear TPC matrix and the training matrix. Based on our new framework of joint matrix-monotonic optimization, the optimal structures of linear transceivers and training sequences were derived for the cases in which the transmitter relies either on statistical CSI or
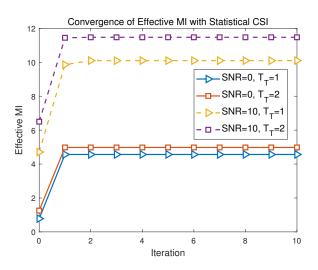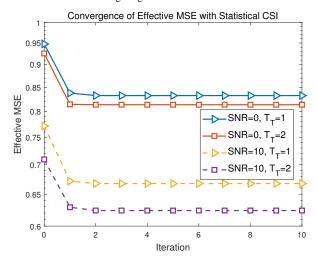
estimated CSI in the face of transmit-side spatial correlation. Based on the optimal structures, the explicit optimization algorithms were proposed for the joint optimization considered. Finally, several numerical simulations were presented for corroborating our theoretical results.

APPENDIX A
THE OBJECTIVE FUNCTIONS FOR PRECODER OPTIMIZATIONS

In this Appendix, based on the channel model in (20), the signal estimation MSE matrix at the receiver is formulated as:

$$\begin{aligned}
\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{G}, \boldsymbol{F}) &= \mathbb{E}\left\{(\widehat{\boldsymbol{s}} - \boldsymbol{s})(\widehat{\boldsymbol{s}} - \boldsymbol{s})^{\mathrm{H}}\right\} \\
&= \mathbb{E}\left\{(\boldsymbol{G}\boldsymbol{y} - \boldsymbol{s})(\boldsymbol{G}\boldsymbol{y} - \boldsymbol{s})^{\mathrm{H}}\right\} \\
&\succeq \boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{G}_{\mathrm{LMMSE}}, \boldsymbol{F}),
\end{aligned} \tag{84}$$

where $\boldsymbol{G}_{\mathrm{LMMSE}}$ is the LMMSE equalizer [11]. For the MSE matrix $\boldsymbol{\Phi}_{\mathrm{MSE}}(\boldsymbol{G}_{\mathrm{LMMSE}}, \boldsymbol{F})$, the nth diagonal element reflects

the MSE of the nth data stream. As discussed in our previous work [10], for MIMO precoder/transmitter beamforming optimization, the OFs are specific matrix functions of the corresponding MSE matrix $\mathbf{\Phi}_{\text{MSE}}(\boldsymbol{G}_{\text{LMMSE}}, \boldsymbol{F})$, which is derived in [10]:

$$\mathbf{\Phi}_{\text{MSE}}(\boldsymbol{G}_{\text{LMMSE}}, \boldsymbol{F}) = (\boldsymbol{I} + \boldsymbol{F}^{\text{H}} \widehat{\boldsymbol{H}}^{\text{H}} \boldsymbol{R}_{\boldsymbol{v}}^{-1} \widehat{\boldsymbol{H}} \boldsymbol{F})^{-1}. \quad (85)$$

The matrix $\boldsymbol{R}_{\boldsymbol{v}}$ is the effective noise covariance matrix including both the channel estimation error and the additive noise. The maximum MI between the transmitter and receiver satisfies $I_{\max}(\boldsymbol{y}; \boldsymbol{s}) \geq -\log \det[\mathbf{\Phi}_{\text{MSE}}(\boldsymbol{G}_{\text{LMMSE}}, \boldsymbol{F})]$ [25]. The corresponding MI maximization OF is

$$\textbf{Obj. 1: } \log \det(\boldsymbol{I} + \boldsymbol{F}^{\text{H}} \widehat{\boldsymbol{H}}^{\text{H}} \boldsymbol{R}_{\boldsymbol{v}}^{-1} \widehat{\boldsymbol{H}} \boldsymbol{F}). \quad (86)$$

The sum data MSE is given by the sum of the diagonal elements of $\mathbf{\Phi}_{\text{MSE}}(\boldsymbol{G}_{\text{LMMSE}}, \boldsymbol{F})$ and the corresponding OF of MSE minimization is [10]

$$\textbf{Obj. 2: } \text{Tr}[(\boldsymbol{I} + \boldsymbol{F}^{\text{H}} \widehat{\boldsymbol{H}}^{\text{H}} \boldsymbol{R}_{\boldsymbol{v}}^{-1} \widehat{\boldsymbol{H}} \boldsymbol{F})^{-1}]. \quad (87)$$

In contrast to the MI, the MSE reflects the recovery accuracy in tangible terms and it is directly related to BER, provided that the error is Gaussian distributed. Moreover, in order to reflect different levels of fairness, a weighted sum performance metric is widely used following the multi-objective optimization framework of [39]. In contrast to traditional vector valued multi-objective optimizations, the specific performance metric is a matrix-valued function. For a positive semidefinite matrix $\mathbf{\Phi}_{\text{V}}$, a general weighting operation may be defined as $\boldsymbol{A}_{\text{W}}^{\text{H}} \mathbf{\Phi}_{\text{V}} \boldsymbol{A}_{\text{W}} + \mathbf{\Pi}_{\text{W}}$, with the weighting parameters $\boldsymbol{A}_{\text{W}}$ and $\mathbf{\Pi}_{\text{W}}$ being an adjustable complex matrix and positive semidefinite matrix, respectively [40]. This kind of weighting operation is much more powerful for matrix-valued multi-objective optimization than traditional element-wise weighting operation. By appropriately varying $\boldsymbol{A}_{\text{W}}$ and $\mathbf{\Pi}_{\text{W}}$, the whole Pareto optimal set can be explored. Thus, for **Obj. 1**, taking $\mathbf{\Phi}_{\text{V}} = \boldsymbol{I} + \boldsymbol{F}^{\text{H}} \widehat{\boldsymbol{H}}^{\text{H}} \boldsymbol{R}_{\boldsymbol{v}}^{-1} \widehat{\boldsymbol{H}} \boldsymbol{F}$ and defining $\boldsymbol{A} = \boldsymbol{A}_{\text{W}} (\boldsymbol{A}_{\text{W}} \boldsymbol{A}_{\text{W}}^{\text{H}} + \mathbf{\Pi}_{\text{W}})^{\frac{1}{2}}$, the OF for the weighted MI is written in the following form

$$\textbf{Obj. 3: } \log \det(\boldsymbol{I} + \boldsymbol{A}^{\text{H}} \boldsymbol{F}^{\text{H}} \widehat{\boldsymbol{H}}^{\text{H}} \boldsymbol{R}_{\boldsymbol{v}}^{-1} \widehat{\boldsymbol{H}} \boldsymbol{F} \boldsymbol{A}). \quad (88)$$

Similarly, for **Obj. 2**, taking $\mathbf{\Phi}_{\text{V}} = (\boldsymbol{I} + \boldsymbol{F}^{\text{H}} \widehat{\boldsymbol{H}}^{\text{H}} \boldsymbol{R}_{\boldsymbol{v}}^{-1} \widehat{\boldsymbol{H}} \boldsymbol{F})^{-1}$ and defining $\boldsymbol{W} = \boldsymbol{A}_{\text{W}} \boldsymbol{A}_{\text{W}}^{\text{H}}$, the OFs of the weighted MSE is formulated as

$$\textbf{Obj. 4: } \text{Tr}[\boldsymbol{W} (\boldsymbol{I} + \boldsymbol{F}^{\text{H}} \widehat{\boldsymbol{H}}^{\text{H}} \boldsymbol{R}_{\boldsymbol{v}}^{-1} \widehat{\boldsymbol{H}} \boldsymbol{F})^{-1}], \quad (89)$$

where $\boldsymbol{W}$ is a positive semidefinite weighting matrix. In the existing literature, a widely used logic for general MIMO transceiver designs is to exploit majorization theory [11], [33]. Based on majorization theory, the OFs are classified into two kinds of functions of the diagonal elements of $\mathbf{\Phi}_{\text{MSE}}(\boldsymbol{G}_{\text{LMMSE}}, \boldsymbol{F})$ (denoted as $\text{d}[\mathbf{\Phi}_{\text{MSE}}(\boldsymbol{G}_{\text{LMMSE}}, \boldsymbol{F})]$), i.e., Schur-convex and Schur-concave functions. The corresponding OFs are expressed as follows

$$\textbf{Obj. 5: } f_{\text{Concave}}^{\text{A-Schur}}\left(\text{d}[(\boldsymbol{F}^{\text{H}} \widehat{\boldsymbol{H}}^{\text{H}} \boldsymbol{R}_{\boldsymbol{v}}^{-1} \widehat{\boldsymbol{H}} \boldsymbol{F} + \boldsymbol{I})^{-1}]\right),$$

$$\textbf{Obj. 6: } f_{\text{Convex}}^{\text{A-Schur}}\left(\text{d}[(\boldsymbol{F}^{\text{H}} \widehat{\boldsymbol{H}}^{\text{H}} \boldsymbol{R}_{\boldsymbol{v}}^{-1} \widehat{\boldsymbol{H}} \boldsymbol{F} + \boldsymbol{I})^{-1}]\right). \quad (90)$$

For Schur-convex functions, the performances for different data streams tend to be the same. On the other hand, for Schur-concave functions, the performances of different data streams tend to depend on the eigenvalues of $\mathbf{\Phi}_{\text{MSE}}(\boldsymbol{G}_{\text{LMMSE}}, \boldsymbol{F})$ [11], [33].

REFERENCES

[1] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 4, pp. 1941–1988, fourthquarter 2015.

[2] H. Xie, F. Gao, S. Zhang and S. Jin, "A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3170-3184, Apr. 2017.

[3] 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, March 2017.

[4] X. Ma, X. Lei, W. Xu and P. T. Mathiopoulos, "Performance analysis of TDD multicell massive MIMO systems with non-orthogonal pilots and hardware imperfections in rician fading channels," *IEEE Trans. Veh. Technol.*, vol. 70, no. 2, pp. 1347-1364, Feb. 2021.

[5] K. Zheng, L. Zhao, J. Mei, B. Shao, W. Xiang and L. Hanzo, "Survey of large-scale MIMO systems," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 3, pp. 1738-1760, thirdquarter 2015.

[6] Y. Wang, C. Li, Y. Huang, D. Wang, T. Ban and L. Yang, "Energy-efficient optimization for downlink massive MIMO FDD systems with transmit-side channel correlation," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7228-7243, Sept. 2016.

[7] B. Wang, M. Jian, F. Gao, G. Y. Li and H. Lin, "Beam squint and channel estimation for wideband mmWave massive MIMO-OFDM systems," *IEEE Trans. Signal Process.*, vol. 67, no. 23, pp. 5893-5908, Dec. 2019.

[8] J. Wang, C. Jiang, H. Zhang, Y. Ren, K. -C. Chen and L. Hanzo, "Thirty years of machine learning: The road to pareto-optimal wireless networks," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 3, pp. 1472-1514, thirdquarter 2020.

[9] Z. Wan, Z. Gao, F. Gao, M. D. Renzo and M. -S. Alouini, "Terahertz massive MIMO with holographic reconfigurable intelligent surfaces," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4732-4750, July 2021.

[10] 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.

[11] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization," *IEEE Trans. Signal Process.*, vol. 51, pp. 2381–2401, Sept. 2003.

[12] H. Xie, F. Gao and S. Jin, "An overview of low-rank channel estimation for massive MIMO systems," *IEEE Access*, vol. 4, pp. 7313-7321, 2016.

[13] B. Jiang, F. Gao, X. Gao and A. Nallanathan, "Channel estimation and training design for two-way relay networks with power allocation," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 2022-2032, June 2010.

[14] K. Xu, S. Gong, M. Cui, G. Zhang and S. Ma, "Statistically robust transceiver design for multi-RIS assisted multi-user MIMO systems," *IEEE Commun. Lett.*, vol. 26, no. 6, pp. 1428-1432, June 2022.

[15] S. Wu, H. Yao, C. Jiang, X. Chen, L. Kuang and L. Hanzo, "Downlink channel estimation for massive MIMO systems relying on vector approximate message passing," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 5145-5148, May 2019.

[16] S. Jin, X. Wang, Z. Li, K. -K. Wong, Y. Huang and X. Tang, "On massive MIMO zero-forcing transceiver using time-shifted pilots," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 59-74, Jan. 2016.

[17] L. You, X. Gao, X. -G. Xia, N. Ma and Y. Peng, "Pilot reuse for massive MIMO transmission over spatially correlated rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3352-3366, June 2015.

[18] C. Qi, G. Yue, L. Wu, Y. Huang and A. Nallanathan, "Pilot design schemes for sparse channel estimation in OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 64, no. 4, pp. 1493-1505, Apr. 2015.

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

[20] Y. Wu, S. Ma and Y. Gu, "A unified framework of non-orthogonal pilot design for multi-cell massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7623-7633, Dec. 2020.

[21] X. Yuan, C. Fan and Y. J. Zhang, "Fundamental limits of training-based uplink multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7544–7558, Nov. 2018.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] X. Ma, F. Yang, S. Liu, J. Song and Z. Han, "Design and optimization on training sequence for mmWave communications: A new approach for sparse channel estimation in massive MIMO," *IEEE J. Select. Areas Commun.*, vol. 35, no. 7, pp. 1486–1497, July 2017.

[29] N. Shariati, J. Wang and M. Bengtsson, "Robust training sequence design for correlated MIMO channel estimation," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 107–120, Jan. 2014.

[30] 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.

[31] Y. Gu and Y. D. Zhang, "Information-theoretic pilot design for downlink channel estimation in FDD massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2334–2346, May 2019.

[32] C. Xing, S. Wang, S. Chen, S. Ma, H. V. Poor and L. Hanzo, "Matrix-monotonic optimization - Part I: Single-variable optimization," *IEEE Trans. Signal Process.*, vol. 69, pp. 738-754, 2021.

[33] C. Xing, S. Wang, S. Chen, S. Ma, H. V. Poor and L. Hanzo, "Matrix-monotonic optimization - Part II: Multi-variable optimization," *IEEE Trans. Signal Process.*, vol. 69, pp. 179-194, 2021.

[34] Z. Fei, B. Li, S. Yang, C. Xing, H. Chen and L. Hanzo, "A survey of multi-objective optimization in wireless sensor networks: Metrics, algorithms, and open problems," *IEEE Commun. Surv. Tutor.*, vol. 19, no. 1, pp. 550-586, Firstquarter 2017.

[35] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. New York: Academic, 1979.

[36] M. Ding and S. D. Blostein, "MIMO minimum total MSE transceiver design with imperfect CSI at both ends," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 1141–1150, Mar. 2009.

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

[38] C. Xing, D. Liu, S. Gong, W. Xu, S. Chen, and L. Hanzo, "Training optimization for hybrid MIMO communication systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5473–5487, May 2020.

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

[40] C. Xing, Y. Jing, and Y. Zhou, "On weighted MSE model for MIMO transceiver optimization," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7072–7085, Feb. 2017.

Tao Yu received the B.S. degree in 2020 from the Beijing Institute of Technology, Beijing, China. She is currently working towards the Ph.D. degree with the School of Information and Electronics, Beijing Institute of Technology. Her main research area includes statistical signal processing, and convex optimization.

Jinpeng Song (Member, IEEE) received the B.S. and Ph.D. degrees from the University of Electronic Science and Technology of China, in 2013 and 2019, respectively. From 2016 to 2018, he was a Visiting Researcher with the University of Wisconsin-Madison, USA. From 2019 to 2021, he was a Postdoctoral Research Fellow with the Beijing Institute of Technology, China, where he is currently an Associate Professor. His research interests include the theory, design, and implementation of digital, analog, and mixed signal processing systems with special emphasis on satellite communication and digital enhancement techniques for analog signal processing systems.

Zhong Zheng (Member, IEEE) received the B.Eng. degree from the Beijing University of Technology, Beijing, China, in 2007, the M.Sc. degree from the Helsinki University of Technology, Espoo, Finland, in 2010, and the D.Sc. degree from Aalto University, Espoo, Finland, in 2015. From 2015 to 2018, he held visiting positions at the University of Texas at Dallas and National Institute of Standards and Technology. In 2019, he joined School of Information and Electronics at Beijing Institute of Technology, Beijing, China, as an Associate Professor. His research interests include massive MIMO, secure communications, millimeter wave communications, random matrix theory, and free probability theory.

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, China, 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 at the University of Macau, Macau, China. His current research interests include machine learning, statistical signal processing, convex optimization, multivariate statistics, array signal processing, and high-frequency band communication systems.

Lian Zhao (S'99-M'03-SM'06) received the Ph.D. degree from the Department of Electrical and Computer Engineering (ELCE), University of Waterloo, Canada, in 2002. She joined the Department of Electrical and Computer Engineering at Toronto Metropolitan University (formerly Ryerson University), Canada, in 2003.

She has been an IEEE Communication Society (ComSoc) and IEEE Vehicular Technology (VTS) Distinguished Lecturer (DL); received the Best Land Transportation Paper Award from IEEE Vehicular Technology Society in 2016, Top 15 Editor Award in 2016 for IEEE Transaction on Vehicular Technology, Best Paper Award from the 2013 International Conference on Wireless Communications and Signal Processing (WCSP), and the Canada Foundation for Innovation (CFI) New Opportunity Research Award in 2005.

She has been serving as an Editor for IEEE Transactions on Wireless Communications, IEEE Internet of Things Journal, and IEEE Transactions on Vehicular Technology (2013-2021). She has been an elected Board of Governor (BoG) committee member since 2023. She has severed as a panel expert in various federal, provincial, and international evaluation committees. She is a licensed Professional Engineer in the Province of Ontario.

Lajos Hanzo (FIEEE'04) received Honorary Doctorates from the Technical University of Budapest and Edinburgh University. He is a Foreign Member of the Hungarian Science-Academy, Fellow of the Royal Academy of Engineering (FREng), of the IET, of EURASIP and holds the IEEE Eric Sumner Technical Field Award. (http://www-mobile.ecs.soton.ac.uk/,https://en.wikipedia.org/wiki/Lajos_Hanzo)