IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received May 8, 2019, accepted June 10, 2019, date of publication June 14, 2019, date of current version July 2, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2922970*

# Robust Superimposed Training Designs for MIMO Relaying Systems Under General Power Constraints

**BEINI RONG, ZHONGSHAN ZHANG, XIN ZHAO, AND XIAOYUN YU**
School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China

Corresponding author: Zhongshan Zhang (zhangzs@bit.edu.cn)

This work was supported in part by the Key Project of the National Natural Science Foundation of China under Grant 61431001, in part by the Beijing Institute of Technology Research Fund Program for Young Scholars, and in part by the Ericsson.

**ABSTRACT** In this paper, we investigate the superimposed training matrix designs for the channel estimation of amplify-and-forward (AF) multiple-input multiple-output (MIMO) relaying systems under general power constraints. Furthermore, the imperfect channel and colored noise statistical information models with the corresponding nominal terms being Kronecker structure and the corresponding statistical errors belonging to the unitarily-invariant uncertainty sets are adopted. Based on the above analysis, the linear minimum mean-squared-error (LMMSE)-based robust training optimization problem is formulated, which is generally nonconvex and intractable. In order to effectively address this problem, we propose an iterative semidefinite programming (SDP) algorithm and two low-complexity upper bound optimization schemes. Particularly, for the proposed two upper bound optimization schemes, the diagonal structured optimal solutions of the relaxed robust training problems can be verified. Besides, the low-complexity iterative bisection search (IBS) can also be applied to derive the diagonal training matrix. Furthermore, we extend our work into the robust mutual information maximization of the AF MIMO relaying channel and demonstrate that all proposed robust training designs are still applicable. Finally, the numerical simulations illustrate the excellent performance of the proposed robust training designs in terms of the channel estimation MSE minimization and mutual information maximization.

**INDEX TERMS** AF MIMO relaying systems, LMMSE channel estimation, robust training designs, general power constraints, unitarily-invariant uncertainty sets.

## I. INTRODUCTION

Shortening the distances between source and destination is one of the most effective way to increase network capacities. Deploying relay node between source and destination can greatly reduce the communication distances to improve the communication quality between source and destination [1]–[10]. In addition, the deployment of multi-antenna arrays at communication terminals is a widely accepted and also a promising way to improve communication quality of cooperative networks [11]–[18]. As a result, multiple-input multiple-output (MIMO) cooperative communications receive lots of attention from academic society [12]. The three most widely used relaying strategies are amplify-and-forward (AF), decode-and-forward (DF) and compress-and-forward (CF) [6], [19]–[22]. Among these protocols, AF MIMO relaying is the most promising technique due to its simplicity, which means that only linear transformation is required at relay to forward the received signals. Therefore, the AF MIMO relaying strategy is widely adopted and has been investigated in depth among a series of existing works e.g., [6], [22]–[28].

Based on different performance metrics, e.g., channel estimation mean square error (MSE) minimization [29]–[36], and mutual information maximization [37]–[39], several channel estimation methods have been proposed for the spatially correlated MIMO with the aid of training sequences [40]. In fact, the criteria of channel estimation MSE minimization and mutual information maximization are distinguished by considering different tradeoffs among the elements of channel estimation MSE matrix [39]. Besides, in most of these works, the kronecker structured channel

The associate editor coordinating the review of this manuscript and approving it for publication was Chenhao Qi.

and colored noise models are assumed due to its mathematical tractability. However, compared with the channel estimation for point-to-point MIMO systems, estimating wireless channels in AF MIMO relaying networks is more difficult because of the compound wireless channels from source to relay and relay to destination, especially when they are correlated [33], [34], [41]. In fact, the literature [33] has investigated the training design for spatially correlated AF MIMO relaying networks without the direct link. Specifically, in [33], the linear minimum mean-squared-error (LMMSE) channel matrix estimation is conducted by only considering the transmit correlation of MIMO channels, which does not fully exploit the complete channel statistical information. In literature [34], both the direct link and complete channel correlations are considered for AF MIMO relaying networks. The formulated training optimization problem under the total power constraint is proved to be convex, which can be efficiently solved with the iterative bisection procedure. Further the literature [41] extended both the minimum MSE and maximum SNR based training designs to AF MIMO multi-relay channel estimation.

Clearly, for most of the existing MIMO training designs, only total power constraint is considered [42]–[44]. However, considering the distributed nature of wireless systems, the power threshold for each communication node may be significantly different [45], [46]. Therefore, in practical communication systems, the general power constraints consisting of a series of linear weighted power constraints are more suitable to promote the highly flexible MIMO configuration. Particularly, the general power constraints contain the total power constraint and individual power constraints as special cases.

Another critical issue is that all aforementioned works assume that the channel and noise spatial correlations are perfect for channel estimation. However, such an assumption is not practical in most wireless systems especially when the highly mobile communication nodes exist [23], [47]. As a result, providing robustness against the channel and noise statistical errors is important for channel estimation of different MIMO systems, so that various communication techniques utilizing the channel information becomes practical [48]–[51]. Besides, as is pointed out in previous work [48], [49], the long-term statistics of channel and noise are generally subject to estimation or quantization errors. Thus, the robust training sequence was proposed for spatially correlated MIMO channels according to different transmission protocols [48], [49]. In particular, the literature [48] considered adding the superimposed training matrix into the precoded transmit signals, which, thus enables the system to release valuable time slots previously occupied by the time-multiplexed pilot sequences. It has been demonstrated that this approach performed well in terms of channel estimation performance. Nevertheless, in literature [49], the conventional time-multiplexed training and signal transmission strategy was adopted for MIMO channel estimation. Moreover, both the arbitrarily correlated MIMO channels and the unitarily-invariant uncertainty set are considered.

To the best of author's knowledge, the robust training designs for AF MIMO relaying systems still remains largely open. Compared with most existing literature, we consider the general power constraints instead of the total power constraints. Statistical errors of both the correlation matrices of channel and noise are taken into account. Moreover, in the considered superimposed training scheme, the training signal is arithmetically added to the transmitted signal, which is effective for improving channel estimation performance. Regarding the importance of this issue, in this paper, we will provide an preliminary and comprehensive work on this topic. The main contributions of our work are briefly summarized in the following.

- Firstly, in our work, we investigate the superimposed training designs for AF MIMO relaying systems, which are characterized by high efficiency and low overhead. During the considered training designs, both the kronecker structured correlated MIMO channels and the colored noises are taken into account. Moreover, the training designs are under more general power constraints including both sum power constraint and per-antenna power constraints as its special cases. In other words, our designs are more general than the traditional designs with only sum power constraint.
- Secondly, instead of assuming statistical information is perfectly known, the statistical information errors of both the correlation matrices of channels and noises are taken into consideration. The robust training optimization problems are formulated following the worst-case robust philosophy. Then, the underlying optimal structures of the optimal solutions are derived, which can greatly simplify the corresponding robust training optimization problems.
- Thirdly, to overcome the non-convexity and intractability of the dimension-reduced robust training optimization problem, we propose an iterative SDP optimization algorithm and two low-complexity upper bound optimization schemes, respectively. Specifically, based on the proposed iterative optimization algorithm, the robust training problem is decomposed into a series of convex sub-problems, and can be solved using linear matrix inequality (LMI) to construct SDP solvers. With the proposed two upper-bound optimization schemes, the diagonal structured optimal solutions are proved, and the iterative bisection search (IBS) can also be applied to obtain the suboptimal diagonal training matrix.
- Finally, we extend our work into the robust mutual information maximization for AF MIMO relaying systems. It is found that all the proposed robust training designs can still work well under new performance metric.

*Notations:* Throughout this paper, without other specifications, vectors and matrices are denoted by bold-faced lowercase and uppercase letters, respectively. Besides, $\boldsymbol{A}^T$, $\boldsymbol{A}^*$, $\boldsymbol{A}^H$ and $\boldsymbol{A}^{-1}$ denote transpose, conjugate, Hermitian transpose

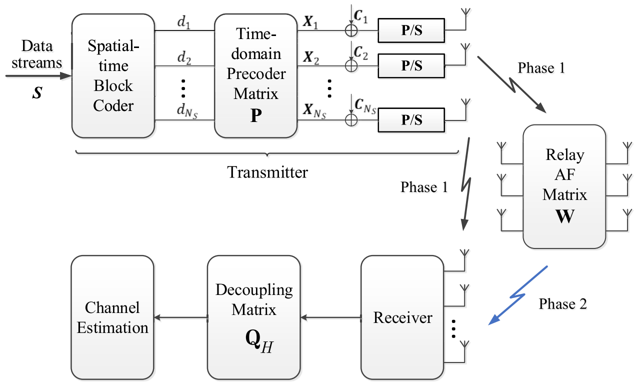


**FIGURE 1. The diagram of the investigated AF MIMO relaying systems.**

and inverse of matrix $\boldsymbol{A}$, respectively. The symbols $\mathrm{Tr}(\boldsymbol{A})$ and $\mathrm{rank}(\boldsymbol{A})$ represent the trace and rank of $\boldsymbol{A}$. In addition, the operation $\mathrm{vec}(\cdot)$ denotes the vector formed by stacking the columns of a matrix. On the other hand, the operation $\mathrm{diag}(\cdot)$ represents a vector consisting of the main diagonal elements of a matrix or a diagonal matrix constructed from a vector. The symbol $\boldsymbol{I}_d$ denotes a $d$-dimensional identity matrix. For a positive semi-definite matrix $\boldsymbol{X}$, its Gaussian square root is denoted as $\boldsymbol{X}^{\frac{1}{2}}$ which is also a Hermitian matrix.

### A. THE AF MIMO RELAYING CHANNEL

In this paper, we consider a AF MIMO relaying system consisting of one $N_S$-antenna source $S$, one $N_R$-antenna relay $R$ and one $N_D$-antenna destination, which is depicted in FIGURE 1. The destination receives the signals transmitted in two consecutive phases. Specifically, in the first phase, both the relay and destination receive the signal transmitted from the source. In the second phase, the relay forwards the received signal from the source to the destination after multiplying an amplifying matrix. The destination recovers its desired signals by jointly processing the two observed signals.

In the considered cooperative network, the channel matrices from source $S$ to destination $D$, source $S$ to relay $R$ and relay $R$ destination $D$ are denoted by $\boldsymbol{H}_{SD} \in \mathbb{C}^{N_D\times N_S}$, $\boldsymbol{H}_{SR} \in \mathbb{C}^{N_R\times N_S}$ and $\boldsymbol{H}_{RD} \in \mathbb{C}^{N_D\times N_R}$, respectively. The transmit signal $\boldsymbol{s}$ is firstly processed by the space-time block coding (STBC) at source, i.e., $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_{N_S}]^T \in \mathbb{C}^{N_S\times K}$, in which $\boldsymbol{d}_i \in \mathbb{C}^K$, $\forall i = 1, 2, \cdots, N_S$ is the $K$-dimensional time-domain symbols of $i$th antenna [52]. Generally, $\mathbb{E}(\boldsymbol{D}\boldsymbol{D}^H) = K\sigma_d^2\boldsymbol{I}_{N_S}$ is assumed, which means the variance of each symbol of $\boldsymbol{D}$ is $\sigma_d^2$. Then the information matrix $\boldsymbol{D}$ is time-domain precoded by the matrix $\boldsymbol{P} \in \mathbb{C}^{K\times K+L}$, where $L \geq N_S$ denotes the length of the introduced redundant vectors to promote the reliable communications [52]. After the time-domain precoding, the new transmit signal $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{N_S}]^T = \boldsymbol{D}\boldsymbol{P} \in \mathbb{C}^{N_S\times(K+L)}$ is obtained, where the precoder $\boldsymbol{P}$ satisfies $\mathrm{Tr}(\boldsymbol{P}\boldsymbol{P}^H) = K + L$ for guaranteing the constant transmit power of the transmit signal $\boldsymbol{D}$.[1] Besides, as seen in most literature, by combining with the superimposed training technique, the precoder $\boldsymbol{P}$ can also be utilized to estimate channel by eliminating the information term at the receiver, thus leaving the intact superimposed training sequence for channel estimation [48], [52], [53]. Therefore, in our work, we also consider the joint transmission of the superimposed training matrix $\boldsymbol{C} \in \mathbb{C}^{N_S\times(K+L)}$ and the transmit signal $\boldsymbol{X}$ to effectively estimate the MIMO AF relaying channel. Based on this technique, the compound signal $\boldsymbol{X} + \boldsymbol{C} = \boldsymbol{D}\boldsymbol{P} + \boldsymbol{C}$ is actually transmitted by the transmitter $S$, and the total transmit power equals

$$\begin{aligned}\mathrm{Tr}((\boldsymbol{X}+\boldsymbol{C})(\boldsymbol{X}+\boldsymbol{C})^H) &= \mathbb{E}[\mathrm{Tr}(\boldsymbol{D}\boldsymbol{P}(\boldsymbol{D}\boldsymbol{P})^H)] + \mathrm{Tr}(\boldsymbol{C}\boldsymbol{C}^H)\\ &= \mathrm{Tr}(\sigma_d^2 N_S \boldsymbol{P}\boldsymbol{P}^H + \boldsymbol{C}\boldsymbol{C}^H). \end{aligned} \quad (1)$$

As discussed above, in our work the direct link between the source and the destination is also taken into account. Thus the received signals at the receiver and relay in the first phase can be written in the following form

$$\begin{aligned}\boldsymbol{Y}_D^{[1]} &= \boldsymbol{H}_{SD}(\boldsymbol{D}\boldsymbol{P} + \boldsymbol{C}) + \boldsymbol{N}_D^{[1]},\\ \boldsymbol{Y}_R &= \boldsymbol{H}_{SR}(\boldsymbol{D}\boldsymbol{P} + \boldsymbol{C}) + \boldsymbol{N}_R. \end{aligned} \quad (2)$$

Without loss of generality, it is assumed that $\boldsymbol{N}_D^{[1]} \in \mathbb{C}^{N_D\times(K+L)}$ and $\boldsymbol{N}_R \in \mathbb{C}^{N_R\times(K+L)}$ are temporally uncorrelated but spatially correlated colored noise. Furthermore, in the second phase the relay forwards the received signal $\boldsymbol{Y}_R$ by pre-multiplying a matrix $\boldsymbol{F}_R \in \mathbb{C}^{N_R\times N_R}$. Particularly, as indicated in [34], considering that fact that in the first phase relay is the receiver, the channel statistical information (CSI) of the first hop, i.e., $\mathbb{E}[\boldsymbol{H}_{SR}\boldsymbol{H}_{SR}^H] = \boldsymbol{\Sigma}_{SR}$, can be perfectly estimated at the relay. Then a diagonal relay forwarding matrix is designed subject to the individual relay antenna power constraints. For example, in [34], $\boldsymbol{F}_R$ is defined as

$$\begin{aligned}\boldsymbol{F}_R = \mathrm{diag}\Bigg[&\sqrt{\frac{P_{R_1}}{P_S\boldsymbol{\Sigma}_{SR}[1,1]+\sigma_n^2}}, \sqrt{\frac{P_{R_2}}{P_S\boldsymbol{\Sigma}_{SR}[2,2]+\sigma_n^2}},\\ &\cdots, \sqrt{\frac{P_{R_{N_R}}}{P_S\boldsymbol{\Sigma}_{SR}[N_R,N_R]+\sigma_n^2}}\Bigg], \end{aligned} \quad (3)$$

where $P_{R_i}$ $\forall i = 1, 2, \cdots, N_R$ is the maximum transmit power of $i$th relay antenna and $\boldsymbol{\Sigma}_{SR}[i,i]$ denotes the $i$th diagonal element of $\boldsymbol{\Sigma}_{SR}$. Further, the relay forwards the signal $\boldsymbol{F}_R\boldsymbol{Y}_R$ to the receiver via channel $\boldsymbol{H}_{RD}$ and then the received signal at the receiver is given by

$$\begin{aligned}\boldsymbol{Y}_D^{[2]} &= \boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{Y}_R + \boldsymbol{N}_D^{[2]}\\ &= \boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{H}_{SR}(\boldsymbol{D}\boldsymbol{P}+\boldsymbol{C}) + \boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{N}_R + \boldsymbol{N}_D^{[2]}, \end{aligned} \quad (4)$$

It is noteworthy that the colored noise $\boldsymbol{N}_D^{[2]} \in \mathbb{C}^{N_D\times(K+L)}$ is also temporally uncorrelated and spatially correlated. Further by combining (2) and (4), the received signals at the

[1] Specifically, by assuming the variance of each symbol of $\boldsymbol{X}$ as $\sigma_x^2$, we have $\sigma_x^2 = \frac{\mathrm{Tr}\left(\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^H)\right)}{N_S(K+L)} = \frac{\sigma_d^2 N_S \mathrm{Tr}(\boldsymbol{P}\boldsymbol{P}^H)}{N_S(K+L)} = \sigma_d^2$, which indicates that transmit signal power of transmitter $S$ keeps invariant after time-domain precoding.

$$\underbrace{\begin{bmatrix} \boldsymbol{Y}_D^{[1]} \\ \boldsymbol{Y}_D^{[2]} \end{bmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{bmatrix} \boldsymbol{H}_{SD} \\ \boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{H}_{SR} \end{bmatrix}}_{\boldsymbol{H}} (\boldsymbol{D}\boldsymbol{P}+\boldsymbol{C}) + \underbrace{\begin{bmatrix} \boldsymbol{I}_{N_D} & \boldsymbol{0}_{N_D\times N_R} & \boldsymbol{0}_{N_D\times N_D} \\ \boldsymbol{0}_{N_D\times N_D} & \boldsymbol{H}_{RD}\boldsymbol{F}_R & \boldsymbol{I}_{N_D} \end{bmatrix} \begin{bmatrix} \boldsymbol{N}_D^{[1]} \\ \boldsymbol{N}_R \\ \boldsymbol{N}_D^{[2]} \end{bmatrix}}_{\boldsymbol{N}} \tag{5}$$

$$\underbrace{\begin{bmatrix} \mathrm{vec}(\boldsymbol{Y}_D^{[1]}\boldsymbol{Q}_H) \\ \mathrm{vec}(\boldsymbol{Y}_D^{[2]}\boldsymbol{Q}_H) \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{(\boldsymbol{I}_2 \otimes (\boldsymbol{C}\boldsymbol{Q}_H)^T \otimes \boldsymbol{I}_{N_D})}_{\bar{\boldsymbol{C}}} \underbrace{\begin{bmatrix} \mathrm{vec}(\boldsymbol{H}_{SD}) \\ \mathrm{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{H}_{SR}) \end{bmatrix}}_{\boldsymbol{h}} + \underbrace{(\boldsymbol{I}_2 \otimes \boldsymbol{Q}_H^T \otimes \boldsymbol{I}_{N_D})}_{\bar{\boldsymbol{Q}}_H} \underbrace{\begin{bmatrix} \mathrm{vec}(\boldsymbol{N}_D^{[1]}) \\ \mathrm{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{N}_R + \boldsymbol{N}_D^{[2]}) \end{bmatrix}}_{\boldsymbol{n}} \tag{7}$$

$$\widetilde{\boldsymbol{R}}_h = \begin{bmatrix} \sigma_h^2(\widehat{\boldsymbol{\Sigma}}_S^T + \boldsymbol{E}_S^T) \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D) & \boldsymbol{0} \\ \boldsymbol{0} & (\hat{\sigma}_h^2 \widehat{\boldsymbol{\Sigma}}_S^T + \widetilde{\boldsymbol{E}}_S^T) \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D) \end{bmatrix} \tag{10}$$

$$\widetilde{\boldsymbol{R}}_n = \begin{bmatrix} \sigma_n^2 \boldsymbol{I}_{K+L} \otimes (\widehat{\boldsymbol{\Psi}}_D + \boldsymbol{E}_{ND}) & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_n^2 \boldsymbol{I}_{K+L} \otimes (\hat{\sigma}_r^2 \widehat{\boldsymbol{\Sigma}}_D + \widehat{\boldsymbol{\Psi}}_D + \widetilde{\boldsymbol{E}}_{ND}) \end{bmatrix} \tag{12}$$

receiver through two phases are ultimately expressed as the formulation (5), as shown at the top of this page. Particularly, we simplify the formulation (5) as $\boldsymbol{Y} = \boldsymbol{H}(\boldsymbol{D}\boldsymbol{P}+\boldsymbol{C}) + \boldsymbol{N}$. Then at the receiver, in order to extract the superimposed training $\boldsymbol{C}$ from the received signal $\boldsymbol{Y}$ to perform channel estimation, we post-multiply $\boldsymbol{Y}$ by a proper decoupling matrix $\boldsymbol{Q}_H$ satisfying $\boldsymbol{P}\boldsymbol{Q}_H = \boldsymbol{0}$. As a result, the received signal model is ultimately simplified as

$$\boldsymbol{Y}\boldsymbol{Q}_H = \boldsymbol{H}\boldsymbol{C}\boldsymbol{Q}_H + \boldsymbol{N}\boldsymbol{Q}_H. \tag{6}$$

*Remark 1:* From the formulation (6), it can be found that in the context of $\boldsymbol{P}\boldsymbol{Q}_H = \boldsymbol{0}$, if $\boldsymbol{Q}_H = \boldsymbol{I}_{K+L}$, then we have $\boldsymbol{P} = \boldsymbol{0}_{K\times(K+L)}$. In this case, our work can be reduced to the conventional training designs for MIMO wireless channels [30]–[34]. However, for the case of $\boldsymbol{Q}_H \neq \boldsymbol{I}_{K+L}$, we can define $\boldsymbol{Q}_H \in \mathbb{C}^{(K+L)\times L}$ as an arbitrary orthogonal subspace of $\boldsymbol{P}$ to guarantee $\boldsymbol{P}\boldsymbol{Q}_H = \boldsymbol{0}$. Then we naturally have $(\boldsymbol{Q}_H)^H\boldsymbol{Q}_H = \boldsymbol{I}_{N_S}$.

### B. THE ROBUST TRAINING OPTIMIZATION PROBLEM

In order to simplify the following derivations, some vectorization operations are adopted to reformulate (6). Specifically, based on the identity $\mathrm{vec}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = (\boldsymbol{C}^T \otimes \boldsymbol{A})\mathrm{vec}(\boldsymbol{B})$, we can vectorize (6) to obtain the formulation (7), as shown at the top of this page, i.e., $\boldsymbol{y} = \bar{\boldsymbol{C}}\boldsymbol{h} + \bar{\boldsymbol{Q}}_H\boldsymbol{n}$, at the top of this page, in which the equivalent channel vector $\boldsymbol{h} \in \mathbb{C}^{2N_SN_D\times 1}$ is to be estimated. Based on (7), the LMMSE based channel estimation can be expressed as $\hat{\boldsymbol{h}} = \boldsymbol{G}_h\boldsymbol{y}$, with $\boldsymbol{G}_h \in \mathbb{C}^{2N_SN_D\times 2N_DL}$ being the LMMSE channel estimator. The estimated channel $\hat{\boldsymbol{h}}$ equals [54]

$$\begin{aligned} \hat{\boldsymbol{h}} &= \boldsymbol{G}_h\boldsymbol{y} \\ &= (\widetilde{\boldsymbol{R}}_h^{-1} + \bar{\boldsymbol{C}}^H\bar{\boldsymbol{Q}}_H\widetilde{\boldsymbol{R}}_n^{-1}\bar{\boldsymbol{Q}}_H^H\bar{\boldsymbol{C}})^{-1}\bar{\boldsymbol{C}}^H\bar{\boldsymbol{Q}}_H\widetilde{\boldsymbol{R}}_n^{-1}\bar{\boldsymbol{Q}}_H^H\boldsymbol{y}, \end{aligned} \tag{8}$$

where $\widetilde{\boldsymbol{R}}_h = \mathbb{E}[\boldsymbol{h}\boldsymbol{h}^H]$ and $\widetilde{\boldsymbol{R}}_n = \mathbb{E}[\boldsymbol{n}\boldsymbol{n}^H]$ are the covariance matrices of the equivalent channel $\boldsymbol{h}$ and colored noise $\boldsymbol{n}$, respectively.

In this work, the widely used Kronecker structured channel models are used i.e.,[2]

$$\begin{aligned} \boldsymbol{H}_{SD} &= \boldsymbol{\Sigma}_D^{\frac{1}{2}}\bar{\boldsymbol{H}}_{SD}\boldsymbol{\Sigma}_S^{\frac{1}{2}}, \\ \boldsymbol{H}_{SR} &= \boldsymbol{\Sigma}_R^{\frac{1}{2}}\bar{\boldsymbol{H}}_{SR}\boldsymbol{\Sigma}_S^{\frac{1}{2}}, \\ \boldsymbol{H}_{RD} &= \boldsymbol{\Sigma}_D^{\frac{1}{2}}\bar{\boldsymbol{H}}_{RD}\boldsymbol{\Sigma}_R^{\frac{1}{2}}, \end{aligned} \tag{9}$$

where the elements of $\{\boldsymbol{\Sigma}_S, \boldsymbol{\Sigma}_R, \boldsymbol{\Sigma}_D\}$ can be determined by the exponential model or the one-ring model [54]. The inner matrices $\bar{\boldsymbol{H}}_{SD}$, $\bar{\boldsymbol{H}}_{SR}$ and $\bar{\boldsymbol{H}}_{RD}$ have i.i.d. Gaussian distributed entries with zero mean and variance $\sigma_h^2$.

#### 1) THE MODELING OF $\widetilde{\boldsymbol{R}}_h$

In our work, the more realistic situation is considered, in which all the spatial correlation matrices $\{\boldsymbol{\Sigma}_S, \boldsymbol{\Sigma}_R, \boldsymbol{\Sigma}_D\}$ are imperfect due to inevitable estimation errors and quantization errors. As derived in Appendix A, the imperfect channel covariance matrix $\widetilde{\boldsymbol{R}}_h$ in (8) can be modeled as the formulation (10), as shown at the top of this page.

In (10), $\hat{\sigma}_h^2 = \sigma_h^4\mathrm{Tr}(\widehat{\boldsymbol{\Sigma}}_R\boldsymbol{F}_R\widehat{\boldsymbol{\Sigma}}_R\boldsymbol{F}_R^H)$, and $\{\widehat{\boldsymbol{\Sigma}}_S, \widehat{\boldsymbol{\Sigma}}_R\ \widehat{\boldsymbol{\Sigma}}_D\}$ are the estimated channel correlation matrices at transmitter, relay and destination, respectively. Besides, the Hermitian matrices $\boldsymbol{E}_S$, $\widetilde{\boldsymbol{E}}_S$ and $\boldsymbol{E}_D$ are the statistical errors of $\boldsymbol{\Sigma}_S$, $\bar{\sigma}_h^2\boldsymbol{\Sigma}_S$ and $\boldsymbol{\Sigma}_D$, respectively.

[2] This channel model is adopted because of the two reasons: 1) This model completely shows the correlation introduced by limited antenna spacings at both communication ends. 2) It is a tractable mathematical model to develop the convenient and meaningful communication techniques for spatially correlated MIMO systems [55].

### 2) THE MODELING OF $\widetilde{R}_n$

Similarly to model $\widetilde{\boldsymbol{R}}_h$, we firstly define the temporally uncorrelated and spatially correlated colored noise as

$$\boldsymbol{N}_D^{[1]} = \boldsymbol{\Psi}_D^{\frac{1}{2}}\bar{\boldsymbol{N}}_D^{[1]}, \quad \boldsymbol{N}_R = \boldsymbol{\Psi}_R^{\frac{1}{2}}\bar{\boldsymbol{N}}_R, \ \boldsymbol{N}_D^{[2]} = \boldsymbol{\Psi}_D^{\frac{1}{2}}\bar{\boldsymbol{N}}_D^{[2]}. \quad (11)$$

The Gaussian elements of $\{\bar{\boldsymbol{N}}_D^{[1]}, \bar{\boldsymbol{N}}_R, \bar{\boldsymbol{N}}_D^{[2]}\}$ are also i.i.d. with zero mean and variance $\sigma_n^2$. Then according to Appendix A, the imperfect noise covariance matrix $\widetilde{\boldsymbol{R}}_n$ is modeled as formulation (12), as shown at the top of the previous page, where $\hat{\sigma}_r^2 = \sigma_h^2 \mathrm{Tr}(\widehat{\boldsymbol{\Psi}}_R^T \boldsymbol{F}_R^T \widehat{\boldsymbol{\Sigma}}_R^T \boldsymbol{F}_R^*)$. $\{\widehat{\boldsymbol{\Psi}}_R, \widehat{\boldsymbol{\Psi}}_D\}$ are the estimated noise spatial correlation matrices at relay and destination, respectively. The Hermitian matrices $\boldsymbol{E}_{ND}$ and $\widetilde{\boldsymbol{E}}_{ND}$ denote the statistical errors of $\boldsymbol{\Psi}_D$ and $\bar{\sigma}_r^2\boldsymbol{\Sigma}_D + \boldsymbol{\Psi}_D$, respectively.

### 3) THE ROBUST TRAINING OPTIMIZATION PROBLEM

In our work, the channel uncertainty is unitarily-invariant $\{\boldsymbol{E}_S, \widetilde{\boldsymbol{E}}_S, \boldsymbol{E}_D, \boldsymbol{E}_{ND}, \widetilde{\boldsymbol{E}}_{ND}\}$, which means that for the arbitrary compact and convex uncertainty set $\xi$, if $\boldsymbol{E} \in \xi$, then $\boldsymbol{U}\boldsymbol{E}\boldsymbol{U}^H \in \xi$ also holds for any unitary matrix $\boldsymbol{U}$ [49]. Generally, this model includes the most popular three types of channel uncertainty sets as its special cases [50]. Such as, the Frobenius norm constrained channel uncertainty [48]–[50], the Spectral norm constrained channel uncertainty [49], [56], and the Nuclear norm constrained channel uncertainty [49]. After substituting (8) into $\mathbb{E}[(\boldsymbol{h} - \boldsymbol{G}_h\boldsymbol{y})(\boldsymbol{h} - \boldsymbol{G}_h\boldsymbol{y})^H]$, the channel estimation MSE matrix can be written as follows

$$\boldsymbol{\Psi}_M = (\widetilde{\boldsymbol{R}}_h^{-1} + \bar{\boldsymbol{C}}^H \bar{\boldsymbol{Q}}_H \widetilde{\boldsymbol{R}}_n^{-1} \bar{\boldsymbol{Q}}_H^H \bar{\boldsymbol{C}})^{-1}. \quad (13)$$

based on which the LMMSE based robust training optimization problem under the general power constraints is formulated as

$$\begin{aligned} &\min_{\boldsymbol{C}} \max_{E\in\xi} \ \mathrm{Tr}\left\{(\widetilde{\boldsymbol{R}}_h^{-1} + \bar{\boldsymbol{C}}^H \bar{\boldsymbol{Q}}_H \widetilde{\boldsymbol{R}}_n^{-1} \bar{\boldsymbol{Q}}_H^H \bar{\boldsymbol{C}})^{-1}\right\}, \\ &s.t. \ \mathrm{Tr}\left(\boldsymbol{W}_l(\boldsymbol{X}+\boldsymbol{C})(\boldsymbol{X}+\boldsymbol{C})^H\right) \le P_{S_l}, \quad \forall\, l = 1, \cdots, L_p, \end{aligned} \quad (14)$$

where the set of statistical errors is defined as

$$E = \{\boldsymbol{E}_S, \widetilde{\boldsymbol{E}}_S, \boldsymbol{E}_D, \boldsymbol{E}_{ND}, \widetilde{\boldsymbol{E}}_{ND}\}. \quad (15)$$

Besides, multiple positive semi-definite weighting matrices $\boldsymbol{W}_l$, $l = 1, \cdots, L_p$ are defined and $P_{S_l}$ denotes the corresponding maximum transmit power. Clearly, the general power constraints cover the total power constraint and the individual power constraints as special cases. Particularly, by setting $\boldsymbol{W}_l = \boldsymbol{I}_{N_S}, \forall\, l = 1, \cdots, L_p$, this model is equivalent to the total power constraint. However, by defining $\boldsymbol{W}_l$ as a diagonal matrix with the $l$-th diagonal element being 1, this model is reduced to the individual power constraints. In fact, the general power constraints are more suitable to the distributed MIMO systems, where multiple single-antenna sources are independently powered by its own battery. Traditionally, the widely adopted method for worst case robustness optimization is S-procedure [57]. However, it is not applied to the problem (14) due to the $\mathrm{Tr}(\boldsymbol{A}^{-1})$ objective function. Moreover, the problem (14) is NP-hard difficult because of the coupled training matrix and error matrices. The underlying optimal structures of the optimal solution of the problem (14) are firstly exploited to reduce the dimensionality. Then alternating algorithms are proposed for the resulting simplified optimization problem in the following section.

## II. THE PROPOSED ALGORITHMS FOR THE ROBUST SUPERIMPOSED TRAINING DESIGN

### A. THE DIMENSION-REDUCED ROBUST TRAINING DESIGN

Substituting (10) and (12) into the problem (14) and after some tedious but straightforward derivations, we have the following formula for robust training optimization under general power constraints

$$\begin{aligned} \min_{\boldsymbol{C}} \max_{E\in\xi} \ & \mathrm{Tr}\Big\{\Big(\sigma_h^{-2}(\widehat{\boldsymbol{\Sigma}}_S^T + \boldsymbol{E}_S^T)^{-1} \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D)^{-1} \\ & + \sigma_n^{-2}(\boldsymbol{C}^*\boldsymbol{Q}_H^*\boldsymbol{Q}_H^T\boldsymbol{C}^T) \otimes (\widehat{\boldsymbol{\Psi}}_D + \boldsymbol{E}_{ND})^{-1}\Big)^{-1}\Big\} \\ & + \mathrm{Tr}\Big\{\Big((\hat{\sigma}_h^2\widehat{\boldsymbol{\Sigma}}_S^T + \widetilde{\boldsymbol{E}}_S^T)^{-1} \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D)^{-1} \\ & + \sigma_n^{-2}(\boldsymbol{C}^*\boldsymbol{Q}_H^*\boldsymbol{Q}_H^T\boldsymbol{C}^T) \otimes \big(\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D + \widehat{\boldsymbol{\Psi}}_D + \widetilde{\boldsymbol{E}}_{ND}\big)^{-1}\Big)^{-1}\Big\} \\ s.t. \ \ & \mathrm{Tr}(\boldsymbol{W}_l\boldsymbol{C}\boldsymbol{C}^H) \le P'_{S_l}, \ , \ l = 1, \cdots, L_p, \end{aligned} \quad (16)$$

where

$$P'_{S_l} = P_{S_l} - \sigma_d^2 N_S \mathrm{Tr}(\boldsymbol{W}_l\boldsymbol{P}\boldsymbol{P}^H). \quad (17)$$

Then a dimension-reduced auxiliary variable $\widetilde{\boldsymbol{C}} = \boldsymbol{C}\boldsymbol{Q}_H \in \mathbb{C}^{N_S\times L}$ is defined and optimized instead of the original variable $\boldsymbol{C} \in \mathbb{C}^{N_S\times(K+L)}$. According to *Remark 1* in Section II, we have $\boldsymbol{P}\boldsymbol{Q}_H = \boldsymbol{0}_{K\times L}$ and $\boldsymbol{Q}_H^H\boldsymbol{Q}_H = \boldsymbol{I}_L$, thus a reasonable value of $\boldsymbol{C} = \widetilde{\boldsymbol{C}}\boldsymbol{Q}_H^H$ is readily observed. Moreover, we also find that this choice of $\boldsymbol{C}$ is optimal in terms of minimizing the transmit power. As a result, the robust training optimization problem (16) can be equivalently reduced to

$$\begin{aligned} \min_{\widetilde{\boldsymbol{C}}} \max_{E\in\xi} \ & \mathrm{Tr}\Big\{\Big(\sigma_h^{-2}(\widehat{\boldsymbol{\Sigma}}_S^T + \boldsymbol{E}_S^T)^{-1} \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D)^{-1} \\ & + \sigma_n^{-2}(\widetilde{\boldsymbol{C}}^*\widetilde{\boldsymbol{C}}^T) \otimes (\widehat{\boldsymbol{\Psi}}_D + \boldsymbol{E}_{ND})^{-1}\Big)^{-1}\Big\} \\ & + \mathrm{Tr}\Big\{\Big((\hat{\sigma}_h^2\widehat{\boldsymbol{\Sigma}}_S^T + \widetilde{\boldsymbol{E}}_S^T)^{-1} \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D)^{-1} \\ & + \sigma_n^{-2}(\widetilde{\boldsymbol{C}}^*\widetilde{\boldsymbol{C}}^T) \otimes \big(\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D + \widehat{\boldsymbol{\Psi}}_D + \widetilde{\boldsymbol{E}}_{ND}\big)^{-1}\Big)^{-1}\Big\} \\ s.t. \ \ & \mathrm{Tr}(\boldsymbol{W}_l\widetilde{\boldsymbol{C}}\widetilde{\boldsymbol{C}}^H) \le P'_{S_l}, \ , \quad l = 1, \cdots, L_p. \end{aligned} \quad (18)$$

It is worth highlighting that the problem (18) is still not jointly convex w.r.t. $\{\widetilde{\boldsymbol{C}}, E\}$. Then in the following, two alternating algorithms are proposed: a) the numerical iterative SDP algorithm; b) the low-complexity upper bound optimization schemes. Particularly, for the low-complexity upper-bound optimization schemes, the semi-closed-form training solutions are available based on the equivalence transformation proposed in [58].

### B. ITERATIVE SDP OPTIMIZATION

After defining $\boldsymbol{R}_{\widetilde{C}} = \widetilde{\boldsymbol{C}}^{*}\widetilde{\boldsymbol{C}}^{T} \in \mathbb{C}^{N_S \times N_S}$, the problem (18) can be further transferred into the following optimization problem

$$\begin{aligned}\min_{\boldsymbol{R}_{\widetilde{C}}} \max_{E\in\xi} \ & \mathrm{Tr}\Big\{\Big(\sigma_h^{-2}(\widehat{\boldsymbol{\Sigma}}_S^T+\boldsymbol{E}_S^T)^{-1}\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1} \\ & +\sigma_n^{-2}\boldsymbol{R}_{\widetilde{C}}\otimes(\widehat{\boldsymbol{\Psi}}_D+\boldsymbol{E}_{ND})^{-1}\Big)^{-1}\Big\} \\ & +\mathrm{Tr}\Big\{\Big((\hat{\sigma}_h^2\widehat{\boldsymbol{\Sigma}}_S^T+\widetilde{\boldsymbol{E}}_S^T)^{-1}\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1} \\ & +\sigma_n^{-2}\boldsymbol{R}_{\widetilde{C}}\otimes\big(\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D+\widehat{\boldsymbol{\Psi}}_D+\widetilde{\boldsymbol{E}}_{ND}\big)^{-1}\Big)^{-1}\Big\} \\ s.t. \ & \mathrm{Tr}(\boldsymbol{W}_l^*\boldsymbol{R}_{\widetilde{C}})\le P'_{S_l},\ \boldsymbol{R}_{\widetilde{C}}\succeq\boldsymbol{0},\quad l=1,2,\cdots,L_p. \end{aligned} \tag{19}$$

It is obvious that the optimization problem (19) is convex w.r.t. $\boldsymbol{R}_{\widetilde{C}}$. On the other hand, it is concave w.r.t. the arbitrary covariance matrix error in the set of $E$. Based on this fact an alternating optimization algorithm is proposed to solve the optimization problem (19). The details of the proposed algorithm is given in the following.

#### 1) THE OPTIMIZATION OF $\boldsymbol{R}_{\widetilde{C}}$

Firstly, for any given channel covariance error, the problem (19) can be reformulated as a standard SDP problem w.r.t. $\boldsymbol{R}_{\widetilde{C}}$ by introducing two auxiliary variables $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, which is

$$\begin{aligned}\min_{\boldsymbol{R}_{\widetilde{C}},\boldsymbol{P}_1,\boldsymbol{P}_2} \ & \mathrm{Tr}(\boldsymbol{P}_1)+\mathrm{Tr}(\boldsymbol{P}_2) \\ s.t. \ & \Big(\sigma_h^{-2}(\widehat{\boldsymbol{\Sigma}}_S^T+\boldsymbol{E}_S^T)^{-1}\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1} \\ & +\sigma_n^{-2}\boldsymbol{R}_{\widetilde{C}}\otimes(\widehat{\boldsymbol{\Psi}}_D+\boldsymbol{E}_{ND})^{-1}\Big)^{-1}\preceq\boldsymbol{P}_1 \\ & \times\Big((\hat{\sigma}_h^2\widehat{\boldsymbol{\Sigma}}_S^T+\widetilde{\boldsymbol{E}}_S^T)^{-1}\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1} \\ & +\sigma_n^{-2}\boldsymbol{R}_{\widetilde{C}}\otimes\big(\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D+\widehat{\boldsymbol{\Psi}}_D+\widetilde{\boldsymbol{E}}_{ND}\big)^{-1}\Big)^{-1}\preceq\boldsymbol{P}_2, \\ & \mathrm{Tr}(\boldsymbol{W}_l^*\boldsymbol{R}_{\widetilde{C}})\le P'_{S_l},\ \boldsymbol{R}_{\widetilde{C}}\succeq\boldsymbol{0},\quad l=1,2,\cdots,L_p. \end{aligned} \tag{20}$$

Based on the Woodbury matrix identity and S-Procedure [56], the optimization problem (20) can be further transferred into the following SDP problem

$$\begin{aligned}\min_{\boldsymbol{R}_{\widetilde{C}},\boldsymbol{P}_1,\boldsymbol{P}_2} \ & \mathrm{Tr}(\boldsymbol{P}_1)+\mathrm{Tr}(\boldsymbol{P}_2) \\ s.t. \ & \begin{bmatrix}\boldsymbol{P}_1 & \boldsymbol{\Sigma}_{SD} \\ \boldsymbol{\Sigma}_{SD} & \boldsymbol{\Sigma}_{SD}+\boldsymbol{\Sigma}_{SD}(\boldsymbol{R}_{\widetilde{C}}\otimes\widetilde{\boldsymbol{\Psi}}_D)\boldsymbol{\Sigma}_{SD}\end{bmatrix}\succeq\boldsymbol{0}, \\ & \begin{bmatrix}\boldsymbol{P}_2 & \boldsymbol{\Sigma}_{SRD} \\ \boldsymbol{\Sigma}_{SRD} & \boldsymbol{\Sigma}_{SRD}+\boldsymbol{\Sigma}_{SRD}\left(\boldsymbol{R}_{\widetilde{C}}\otimes\boldsymbol{\Psi}_{ND}\right)\boldsymbol{\Sigma}_{SRD}\end{bmatrix}\succeq\boldsymbol{0}, \\ & \mathrm{Tr}(\boldsymbol{W}_l^*\boldsymbol{R}_{\widetilde{C}})\le P'_{S_l}, \\ & \boldsymbol{R}_{\widetilde{C}}\succeq\boldsymbol{0},\quad \forall\, l=1,2,\cdots,L_p, \end{aligned} \tag{21}$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_{SD} &= \sigma_h^2(\widehat{\boldsymbol{\Sigma}}_S^T+\boldsymbol{E}_S^T)\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D), \\ \boldsymbol{\Sigma}_{SRD} &= (\hat{\sigma}_h^2\widehat{\boldsymbol{\Sigma}}_S^T+\widetilde{\boldsymbol{E}}_S^T)\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D), \\ \widetilde{\boldsymbol{\Psi}}_D &= \sigma_n^{-2}(\widehat{\boldsymbol{\Psi}}_D+\boldsymbol{E}_{ND})^{-1}, \\ \boldsymbol{\Psi}_{ND} &= \sigma_n^{-2}\big(\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D+\widehat{\boldsymbol{\Psi}}_D+\widetilde{\boldsymbol{E}}_{ND}\big)^{-1}. \end{aligned} \tag{22}$$

*Remark 2:* Particularly, once the optimal $\boldsymbol{R}_{\widetilde{C}}$ is derived from problem (21) given the worst case $E$, we can utilize the eigenvalue decomposition to obtain $\widetilde{\boldsymbol{C}}$ as $\widetilde{\boldsymbol{C}} = [\widetilde{\boldsymbol{C}}_1, \boldsymbol{0}_{N_S\times(L-N_S)}]$, where $\widetilde{\boldsymbol{C}}_1 \in \mathbb{C}^{N_S\times N_S}$ and $\boldsymbol{R}_{\widetilde{C}} = \widetilde{\boldsymbol{C}}_1^*\widetilde{\boldsymbol{C}}_1^T$. Further according to $\boldsymbol{C} = \widetilde{\boldsymbol{C}}\boldsymbol{Q}_H^H$, the optimal training $\boldsymbol{C}$ is expressed as $\boldsymbol{C} = \widetilde{\boldsymbol{C}}_1(\boldsymbol{Q}_H^{[1:N_S]})^H$, where $\boldsymbol{Q}_H^{[1:N_S]}$ denotes the 1-th column to the $N_S$-th column of matrix $\boldsymbol{Q}_H$. More importantly, we observe $\boldsymbol{C}\boldsymbol{P}^H = \widetilde{\boldsymbol{C}}_1(\boldsymbol{Q}_H^{[1:N_S]})^H\boldsymbol{P}^H = \boldsymbol{0}_{N_S\times K}$, which means that the optimal robust superimposed training $\boldsymbol{C}$ of the problem (16) is orthogonal to the time-precoder $\boldsymbol{P}$.

#### 2) THE OPTIMIZATION OF $E$

Similar to $\boldsymbol{R}_{\widetilde{C}}$, the robust optimization problem (19) can also be formulated as a jointly concave one w.r.t. $\{\boldsymbol{E}_S, \widetilde{\boldsymbol{E}}_S\}$ when $\{\boldsymbol{R}_{\widetilde{C}}, \boldsymbol{E}_{ND}, \widetilde{\boldsymbol{E}}_{ND}, \boldsymbol{E}_D\}$ is fixed, $\boldsymbol{E}_D$ when $\{\boldsymbol{R}_{\widetilde{C}}, \boldsymbol{E}_S, \widetilde{\boldsymbol{E}}_S, \boldsymbol{E}_{ND}, \widetilde{\boldsymbol{E}}_{ND}\}$ is fixed or $\{\boldsymbol{E}_{ND}, \widetilde{\boldsymbol{E}}_{ND}\}$ when $\{\boldsymbol{R}_{\widetilde{C}}, \boldsymbol{E}_S, \widetilde{\boldsymbol{E}}_S, \boldsymbol{E}_D\}$ is fixed. Here, we take the optimization of $\{\boldsymbol{E}_{ND}, \widetilde{\boldsymbol{E}}_{ND}\}$ as an example and formulate the corresponding SDP problem as the formulation

$$\begin{aligned}\max_{\{\boldsymbol{E}_{ND},\widetilde{\boldsymbol{E}}_{ND}\}\in\xi} \ & \mathrm{Tr}(\boldsymbol{P}_1)+\mathrm{Tr}(\boldsymbol{P}_2) \\ s.t. \ & \begin{bmatrix}\boldsymbol{I}+\boldsymbol{\Sigma}_{SD}^{-\frac{1}{2}}\widehat{\boldsymbol{R}}_{D,1}(\boldsymbol{E}_{ND})\boldsymbol{\Sigma}_{SD}^{-\frac{1}{2}} & \boldsymbol{\Sigma}_{SD}^{-\frac{1}{2}}\widehat{\boldsymbol{R}}_{D,1}(\boldsymbol{E}_{ND}) \\ \widehat{\boldsymbol{R}}_{D,1}(\boldsymbol{E}_{ND})\boldsymbol{\Sigma}_{SD}^{-\frac{1}{2}} & \widehat{\boldsymbol{R}}_{D,1}(\boldsymbol{E}_{ND})-\boldsymbol{P}_1\end{bmatrix}\succeq\boldsymbol{0} \\ & \begin{bmatrix}\boldsymbol{I}+\boldsymbol{\Sigma}_{SRD}^{-\frac{1}{2}}\widehat{\boldsymbol{R}}_{D,2}(\widetilde{\boldsymbol{E}}_{ND})\boldsymbol{\Sigma}_{SRD}^{-\frac{1}{2}} & \boldsymbol{\Sigma}_{SRD}^{-\frac{1}{2}}\widehat{\boldsymbol{R}}_{D,2}(\widetilde{\boldsymbol{E}}_{ND}) \\ \widehat{\boldsymbol{R}}_{D,2}(\widetilde{\boldsymbol{E}}_{ND})\boldsymbol{\Sigma}_{SRD}^{-\frac{1}{2}} & \widehat{\boldsymbol{R}}_{D,2}(\widetilde{\boldsymbol{E}}_{ND})-\boldsymbol{P}_2\end{bmatrix}\succeq\boldsymbol{0} \end{aligned} \tag{23}$$

where $\widehat{\boldsymbol{R}}_{D,1}(\boldsymbol{E}_{ND}) = \sigma_n^{-2}\boldsymbol{R}_{\widetilde{C}}^{-1}\otimes(\widehat{\boldsymbol{\Psi}}_D+\boldsymbol{E}_{ND})$ and $\widehat{\boldsymbol{R}}_{D,2}(\widetilde{\boldsymbol{E}}_{ND}) = \sigma_n^{-2}\boldsymbol{R}_{\widetilde{C}}^{-1}\otimes\big(\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D+\widehat{\boldsymbol{\Psi}}_D+\widetilde{\boldsymbol{E}}_{ND}\big)$. As discussed above, the optimization of $\boldsymbol{E}_D$, $\{\boldsymbol{E}_S, \widetilde{\boldsymbol{E}}_S\}$ can also be formulated as standard SDP optimization problems and their optimal solutions can be found using some numerical algorithms, such as interior-point method.

Thus, by iteratively computing the derived SDP optimizations given in (21), (23) and the corresponding SDP problems w.r.t. $\boldsymbol{E}_S, \widetilde{\boldsymbol{E}}_S$ and $\boldsymbol{E}_D$, the dimension-reduced problem (18) can be solved efficiently.

As presented in [59], the computational complexity for a standard SDP problem is $\mathcal{O}(M_{sdp}N_{sdp}^{3.5} + M_{sdp}^2N_{sdp}^{2.5} + M_{sdp}^3\,N_{sdp}^{0.5})\log(1/\epsilon)$, where $N_{sdp}$ and $M_{sdp}$ denote the dimension of semidefinite cone and the number of semidefinite cone constraints, respectively, and $\epsilon$ indicates the solution accuracy. In our proposed iterative SDP algorithm, each iteration mainly involves three SDP subproblems in terms of the variable blocks $\boldsymbol{R}_{\widetilde{C}}$, $\{\boldsymbol{E}_N, \widetilde{\boldsymbol{E}}_N\}$ and $\{\boldsymbol{E}_H, \widetilde{\boldsymbol{E}}_H\}$. Therefore, the order of total complexity of our proposed iterative SDP algorithm is $I_{tol}\mathcal{O}\big((N_SN_D)^{3.5}\big)\log(1/\epsilon)$, where $I_{tol}$ denotes the total number of iterations. However, in terms of our work,

**Algorithm 1** The Subgradient Algorithm for Problem (25)

1: Initialize: iteration index $n = 0$; maximum iteration number $I_{max}$; auxiliary parameters $\mu_l^{(0)}, \forall\, l = 1, \cdots, L$;
2: **repeat**
3: Solve the problem (25) to obtain $\boldsymbol{X}^{(n)}$ given $\mu_l^{(n)}$;
4: Define the step size $t_n = \frac{c}{a+n\cdot b}$, where the scalars $\{a, b, c\} > 0$;
5: Update $\mu_l^{(n+1)} = \mathcal{P}[\mu_l^{(n)} + t_n(\mathrm{Tr}(\boldsymbol{W}_l\boldsymbol{X}\boldsymbol{X}^H) - P_l)]$ by solving the problem

$$\min \sum_{l=1}^{L} \|\mathcal{P}[\mu_l] - \mu_l\|_2,$$
$$s.t.\ \sum_{l=1}^{L} \mathcal{P}[\mu_l]\boldsymbol{W}_l \succ \boldsymbol{0}, \quad \mathcal{P}[\mu_l] \geq 0. \quad (26)$$

6: Update $n = n + 1$
7: **until** $\mu_l^{(n-1)}(\mathrm{Tr}(\boldsymbol{W}_k\boldsymbol{X}^{(n-1)}(\boldsymbol{X}^{(n-1)})^H) - P_l) \leq \epsilon_l,\ \forall l$ or $n \leq I_{max}$, where $\epsilon_l > 0$ is sufficiently small.
8: **return** $\mu_l^{(n)}, \boldsymbol{X}^{(n)}, \forall\, l = 1, \cdots, L$.

as illustrated in simulations, the actual complexity of the proposed iterative SDP algorithm is much lower than the above worst case bound.

### C. THE LOW-COMPLEXITY UPPER BOUND OPTIMIZATION SCHEMES

#### 1) THE EQUIVALENCE TRANSFORMATION

From problem (18), it can be found that given statistical errors set $E$, the objective function of problem (18) can be classified into $f(\widetilde{\boldsymbol{C}}^*\boldsymbol{\Xi}\widetilde{\boldsymbol{C}}^T)$ with $\boldsymbol{\Xi} = \sigma_n^{-2}\boldsymbol{I}_L$. According to the literature [2], the function $f(\widetilde{\boldsymbol{C}}^*\boldsymbol{\Xi}\widetilde{\boldsymbol{C}}^T)$ is matrix monotonically decreasing w.r.t $\widetilde{\boldsymbol{C}}^*\boldsymbol{\Xi}\widetilde{\boldsymbol{C}}^T$. Then a useful **Theorem 1** in [58] is presented for the following analysis.

*Theorem 1:* For a matrix monotonically decreasing function $f(\boldsymbol{X}\boldsymbol{\Xi}\boldsymbol{X}^H)$, where $\boldsymbol{X} \in \mathbb{C}^{N\times M}$ and $\boldsymbol{\Xi} \in \mathbb{C}^{M\times M}$, the following problem

$$\min_{\boldsymbol{X}}\ f(\boldsymbol{X}\boldsymbol{\Xi}\boldsymbol{X}^H) \quad s.t.\ \mathrm{Tr}(\boldsymbol{W}_l\boldsymbol{X}\boldsymbol{X}^H) \leq P_l, \quad l = 1, \cdots, L_p, \quad (24)$$

with $\boldsymbol{W}_l \succeq \boldsymbol{0} \in \mathbb{C}^{N\times N}$, is equivalent to

$$\min_{\boldsymbol{X}}\ f(\boldsymbol{X}\boldsymbol{\Xi}\boldsymbol{X}^H) \quad s.t.\ \mathrm{Tr}(\boldsymbol{W}\boldsymbol{X}\boldsymbol{X}^H) \leq P, \quad (25)$$

where $P = \sum_{l=1}^{L_p} P_l$ and $\boldsymbol{W} = \sum_{l=1}^{L_p} u_l\boldsymbol{W}_l$. Note that $\boldsymbol{W} \succ \boldsymbol{0}$ and this constraint guarantees that the original problem (24) is feasible. Besides, $u_l's$ are nonnegative scalars satisfying $u_l(\mathrm{Tr}(\boldsymbol{W}_l\boldsymbol{X}\boldsymbol{X}^H) - P_l) = 0, \forall\, l$, and can be found using the subgradient method [60], which is briefly introduced in **Algorithm 1**.

Based on **Theorem 1** and defining $\boldsymbol{R}_{\widetilde{C}} = \widetilde{\boldsymbol{C}}^*\widetilde{\boldsymbol{C}}^T$, the robust training optimization problem (18) with general power constraints is equivalent to

$$\begin{aligned}\min_{\boldsymbol{R}_{\widetilde{C}}}\max_{E\in\xi}\ &\mathrm{Tr}\Bigg\{\bigg(\sigma_h^{-2}(\widehat{\boldsymbol{\Sigma}}_S^T + \boldsymbol{E}_S^T)^{-1} \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D)^{-1} \\ &+ \sigma_n^{-2}\boldsymbol{R}_{\widetilde{C}} \otimes (\widehat{\boldsymbol{\Psi}}_D + \boldsymbol{E}_{ND})^{-1}\bigg)^{-1}\Bigg\} \\ &+ \mathrm{Tr}\Bigg\{\bigg((\hat{\sigma}_h^2\widehat{\boldsymbol{\Sigma}}_S^T + \widetilde{\boldsymbol{E}}_S^T)^{-1} \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D)^{-1} \\ &+ \sigma_n^{-2}\boldsymbol{R}_{\widetilde{C}} \otimes \left(\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D + \widehat{\boldsymbol{\Psi}}_D + \widetilde{\boldsymbol{E}}_{ND}\right)^{-1}\bigg)^{-1}\Bigg\} \\ s.t.\ &\mathrm{Tr}(\boldsymbol{W}^*\boldsymbol{R}_{\widetilde{C}}) \leq P'_S, \quad \boldsymbol{R}_{\widetilde{C}} \succeq \boldsymbol{0},\end{aligned} \quad (27)$$

where $\boldsymbol{W}^* = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{W}^{\frac{1}{2}} = \sum_{l=1}^{L_p} \mu_l\boldsymbol{W}_l^* \succ \boldsymbol{0}$ and $P'_S = \sum_{l=1}^{L_p} P'_{S_l}$. Unfortunately, the problem (27) is still nonconvex due to the coupled variables. In the following in order to solve the problem (27) effectively, two upper bound optimization algorithms are proposed.

#### 2) WEIGHTING-SCALED-SCHEME

The first alternative scheme is based on the relaxation of the weighting related matrix $\boldsymbol{W}$. To be specific, we firstly scale the power constraint $\mathrm{Tr}(\boldsymbol{W}^*\boldsymbol{R}_{\widetilde{C}}) \leq P'_S$ in problem (27) up to $\mathrm{Tr}(\ \lambda_{\max}(\boldsymbol{W}^*)\boldsymbol{R}_{\widetilde{C}}) \leq P'_S$. Then by defining $\widetilde{\boldsymbol{R}}_{\widetilde{C}} = \lambda_{\max}(\boldsymbol{W}^*)\boldsymbol{R}_{\widetilde{C}}$, the weighting-scaled problem is formulated as

$$\begin{aligned}\min_{\widetilde{\boldsymbol{R}}_{\widetilde{C}}}\max_{E\in\xi}\ &\mathrm{Tr}\Bigg\{\bigg(\sigma_h^{-2}(\widehat{\boldsymbol{\Sigma}}_S^T + \boldsymbol{E}_S^T)^{-1} \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D)^{-1} \\ &+ \sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\widetilde{\boldsymbol{R}}_{\widetilde{C}} \otimes (\widehat{\boldsymbol{\Psi}}_D + \boldsymbol{E}_{ND})^{-1}\bigg)^{-1}\Bigg\} \\ &+ \mathrm{Tr}\Bigg\{\bigg((\hat{\sigma}_h^2\widehat{\boldsymbol{\Sigma}}_S^T + \widetilde{\boldsymbol{E}}_S^T)^{-1} \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D)^{-1} \\ &+ \sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\widetilde{\boldsymbol{R}}_{\widetilde{C}} \\ &\otimes (\widehat{\boldsymbol{\Psi}}_D + \hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D + \widetilde{\boldsymbol{E}}_{ND})^{-1}\bigg)^{-1}\Bigg\} \\ s.t.\ &\mathrm{Tr}(\widetilde{\boldsymbol{R}}_{\widetilde{C}}) \leq P'_S,\ \widetilde{\boldsymbol{R}}_{\widetilde{C}} \succeq \boldsymbol{0}.\end{aligned} \quad (28)$$

In fact, the optimal analytical structure of $\{\widetilde{\boldsymbol{R}}_{\widetilde{C}}, E\}$ for the problem (28) can be proved utilizing the EVDs of $\{\widehat{\boldsymbol{\Sigma}}_S, \widehat{\boldsymbol{\Sigma}}_D, \widehat{\boldsymbol{\Psi}}_D\}$. So we firstly define the following EVDs

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_S^T &= \widehat{\boldsymbol{U}}_S\boldsymbol{\Lambda}_S\widehat{\boldsymbol{U}}_S^H, \\ \widehat{\boldsymbol{\Sigma}}_D &= \widehat{\boldsymbol{U}}_D\boldsymbol{\Lambda}_D\widehat{\boldsymbol{U}}_D^H, \\ \widehat{\boldsymbol{\Psi}}_D &= \widehat{\boldsymbol{U}}_D\boldsymbol{\Omega}_D\widehat{\boldsymbol{U}}_D^H,\end{aligned} \quad (29)$$

where $\widehat{\boldsymbol{U}}_S$ and $\widehat{\boldsymbol{U}}_D$ are the eigenvector unitary matrix of $\widehat{\boldsymbol{\Sigma}}_S^T$ and $\widehat{\boldsymbol{\Psi}}_D$, respectively. The diagonal elements of diagonal matrices $\{\boldsymbol{\Lambda}_S, \boldsymbol{\Lambda}_D, \boldsymbol{\Omega}_D\}$ are the eigenvalues of $\{\widehat{\boldsymbol{\Sigma}}_S, \widehat{\boldsymbol{\Sigma}}_D, \widehat{\boldsymbol{\Psi}}_D\}$, respectively, with $\boldsymbol{\Lambda}_S = \mathrm{diag}[\lambda_{S,1}, \cdots, \lambda_{S,N_S}]$, $\boldsymbol{\Lambda}_D = \mathrm{diag}[\lambda_{D,1}, \cdots, \lambda_{D,N_D}]$ and $\boldsymbol{\Omega}_D = \mathrm{diag}[\Omega_{D,1}, \cdots, \Omega_{D,N_D}]$. From (29), it can be observed that $\widehat{\boldsymbol{\Psi}}_D$ and $\widehat{\boldsymbol{\Sigma}}_D$ have identical eigenvectors [61]. It means that the colored noise shares the spatial structure of the MIMO channels. In other words, the colored noise impinges on the receiver antenna array from the same spatial directions as that of desired signals [61]. This model is widely used in the existing training designs [30]–[32], [49], [61]. Based on the structure derived in (29), the problem (28) is further reformulated as

$$\begin{aligned}\min_{\widetilde{\boldsymbol{R}}_{\widetilde{C}}}\max_{E'\in\xi}\ &\mathrm{Tr}\Big\{\Big(\sigma_h^{-2}(\boldsymbol{\Lambda}_S+\boldsymbol{E}'_S)^{-1}\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}'_D)^{-1}\\&+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\widehat{\boldsymbol{U}}_S^H\widetilde{\boldsymbol{R}}_{\widetilde{C}}\widehat{\boldsymbol{U}}_S\otimes(\boldsymbol{\Omega}_D+\boldsymbol{E}'_{ND})^{-1}\Big)^{-1}\Big\}\\&+\mathrm{Tr}\Big\{\Big((\hat{\sigma}_h^2\boldsymbol{\Lambda}_S+\widetilde{\boldsymbol{E}}'_S)^{-1}\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}'_D)^{-1}+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\\&\times\widehat{\boldsymbol{U}}_S^H\widetilde{\boldsymbol{R}}_{\widetilde{C}}\widehat{\boldsymbol{U}}_S\otimes(\boldsymbol{\Omega}_D+\hat{\sigma}_r^2\boldsymbol{\Lambda}_D+\widetilde{\boldsymbol{E}}'_{ND})^{-1}\Big)^{-1}\Big\}\\ s.t.\ &\mathrm{Tr}(\widetilde{\boldsymbol{R}}_{\widetilde{C}})\le P'_S,\ \ \widetilde{\boldsymbol{R}}_{\widetilde{C}}\succeq\boldsymbol{0},\end{aligned}\tag{30}$$

where $E'=\{\boldsymbol{E}'_D,\boldsymbol{E}'_S,\widetilde{\boldsymbol{E}}'_S,\boldsymbol{E}'_{ND},\widetilde{\boldsymbol{E}}'_{ND}\}$, $\boldsymbol{E}'_D=\widehat{\boldsymbol{U}}_D^H\boldsymbol{E}_D\widehat{\boldsymbol{U}}_D$ and $\boldsymbol{E}'_S=\widehat{\boldsymbol{U}}_S^H\boldsymbol{E}_S^T\widehat{\boldsymbol{U}}_S,\widetilde{\boldsymbol{E}}'_S=\widehat{\boldsymbol{U}}_S^H\widetilde{\boldsymbol{E}}_S^T\widehat{\boldsymbol{U}}_S$ as well as $\boldsymbol{E}'_{ND}=\widehat{\boldsymbol{U}}_D^H\boldsymbol{E}_{ND}\widehat{\boldsymbol{U}}_D,\widetilde{\boldsymbol{E}}'_{ND}=\widehat{\boldsymbol{U}}_D^H\widetilde{\boldsymbol{E}}_{ND}\widehat{\boldsymbol{U}}_D$. To simplify the considered optimization problems, the optimal diagonalizable structures of $\{\widehat{\boldsymbol{U}}_S^H\widetilde{\boldsymbol{R}}_{\widetilde{C}}\widehat{\boldsymbol{U}}_S,E'\}$ of the problem (30) are derived and the results are demonstrated **Theorem 2**.

*Theorem 2:* For the problem (30), the optimal variables still satisfy the following diagonalizable structure

$$\{\widehat{\boldsymbol{U}}_S^H\widetilde{\boldsymbol{R}}_{\widetilde{C}}^{opt}\widehat{\boldsymbol{U}}_S,E'^{opt}\}=\{\boldsymbol{\Lambda}_C,\boldsymbol{\Lambda}_{E'_D}\boldsymbol{\Lambda}_{E'_S},\boldsymbol{\Lambda}_{\widetilde{E}'_S},\boldsymbol{\Lambda}_{E'_{ND}},\boldsymbol{\Lambda}_{\widetilde{E}'_{ND}}\}.\tag{31}$$

where $\boldsymbol{\Lambda}_C=\mathrm{diag}[\lambda_{C,1},\cdots,\lambda_{C,N_S}]$, $\boldsymbol{\Lambda}_{E'_D}=\mathrm{diag}[\lambda_{E'_D,1},\cdots,\lambda_{E'_D,N_D}]$, $\boldsymbol{\Lambda}_{E'_S}=\mathrm{diag}[\lambda_{E'_S,1},\cdots,\lambda_{E'_S,N_S}]$, $\boldsymbol{\Lambda}_{\widetilde{E}'_S}=\mathrm{diag}[\lambda_{\widetilde{E}'_S,1},\cdots,\lambda_{\widetilde{E}'_S,N_S}]$, $\boldsymbol{\Lambda}_{E'_{ND}}=\mathrm{diag}[\lambda_{E'_{ND},1},\cdots,\lambda_{E'_{ND},N_D}]$ and $\boldsymbol{\Lambda}_{\widetilde{E}'_{ND}}=\mathrm{diag}[\lambda_{\widetilde{E}'_{ND},1},\cdots,\lambda_{\widetilde{E}'_{ND},N_D}]$. Then the matrix-variable problem (30) can be simplified as the following optimization problem

$$\begin{aligned}\min_{\boldsymbol{\Lambda}_C}\max_{\Lambda_E\in\xi}\ &\sum_{i=1}^{N_S}\sum_{j=1}^{N_D}\Big(\sigma_h^{-2}(\lambda_{S,i}+\lambda_{E'_S,i})^{-1}(\lambda_{D,j}+\lambda_{E'_D,j})^{-1}\\&+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\lambda_{C,i}(\Omega_{D,j}+\lambda_{E'_{ND},j})^{-1}\Big)^{-1}\\&+\sum_{i=1}^{N_S}\sum_{j=1}^{N_D}\Big((\hat{\sigma}_h^2\lambda_{S,i}+\lambda_{\widetilde{E}'_S,i})^{-1}(\lambda_{D,j}+\lambda_{E'_D,j})^{-1}\\&+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\lambda_{C,i}(\Omega_{D,j}+\hat{\sigma}_r^2\lambda_{D,j}\\&+\lambda_{\widetilde{E}'_{ND},j})^{-1}\Big)^{-1}\\ s.t.\ &\sum_{i=1}^{N_S}\lambda_{C,i}\le P'_S,\quad \lambda_{C,i}\ge 0,\ i=1,\cdots,N_S.\end{aligned}\tag{32}$$

where $\Lambda_E=\{\boldsymbol{\Lambda}_{E'_D}\boldsymbol{\Lambda}_{E'_S},\boldsymbol{\Lambda}_{\widetilde{E}'_S},\boldsymbol{\Lambda}_{E'_{ND}},\boldsymbol{\Lambda}_{\widetilde{E}'_{ND}}\}$. By recalling that $\boldsymbol{C}=\widetilde{\boldsymbol{C}}_1(\boldsymbol{Q}_H^{[1:N_S]})^H$, $\boldsymbol{R}_{\widetilde{C}}=\widetilde{\boldsymbol{C}}_1^*\widetilde{\boldsymbol{C}}_1^T$ and $\widetilde{\boldsymbol{R}}_{\widetilde{C}}=\lambda_{\max}(\boldsymbol{W}^*)\boldsymbol{R}_{\widetilde{C}}$, we can finally derive the semi-closed-form $\{\boldsymbol{C},E\}$ of the weighting-scaled-scheme as

$$\begin{aligned}\boldsymbol{C}&=\lambda_{\max}(\boldsymbol{W}^*)^{-\frac{1}{2}}\widehat{\boldsymbol{U}}_S^*\boldsymbol{\Lambda}_C^{\frac{1}{2}}\boldsymbol{Q}_A(\boldsymbol{Q}_H^{[1:N_S]})^T,\\ \boldsymbol{E}_D&=\widehat{\boldsymbol{U}}_D\boldsymbol{\Lambda}_{E'_D}\widehat{\boldsymbol{U}}_D^H,\\ \boldsymbol{E}_S^T&=\widehat{\boldsymbol{U}}_S\boldsymbol{\Lambda}_{E'_S}\widehat{\boldsymbol{U}}_S^H,\\ \widetilde{\boldsymbol{E}}_S^T&=\widehat{\boldsymbol{U}}_S\boldsymbol{\Lambda}_{\widetilde{E}'_S}\widehat{\boldsymbol{U}}_S^H,\\ \boldsymbol{E}_{ND}&=\widehat{\boldsymbol{U}}_D\boldsymbol{\Lambda}_{E'_{ND}}\widehat{\boldsymbol{U}}_D^H,\\ \widetilde{\boldsymbol{E}}_{ND}&=\widehat{\boldsymbol{U}}_D\boldsymbol{\Lambda}_{\widetilde{E}'_{ND}}\widehat{\boldsymbol{U}}_D^H.\end{aligned}\tag{33}$$

where $\boldsymbol{Q}_A\in\mathbb{C}^{N_S\times N_S}$ is an arbitrary orthogonal unitary matrix satisfying $\boldsymbol{Q}_A\boldsymbol{Q}_A^H=\boldsymbol{I}_{N_S}$.

*Proof:* Please see Appendix B.

Particularly, the proposed weighting-scaled-scheme is tight when $\boldsymbol{W}_l=\boldsymbol{I}_{N_S},\forall\, l=1,\cdots,L$ holds corresponding to the total power constraint. In other words, for the case of $\boldsymbol{W}_l=\boldsymbol{I}_{N_S},\forall\, l=1,\cdots,L$, the analytical solutions presented in **Theorem 2** are actually optimal to the original robust training problem (16).

### 3) CORRELATION-SCALED-SCHEME

Firstly, based on the definition $\overline{\boldsymbol{R}}_{\widetilde{C}}=\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{R}_{\widetilde{C}}\boldsymbol{W}^{\frac{1}{2}}$, the robust training problem (27) can be reformulated as

$$\begin{aligned}\min_{\overline{\boldsymbol{R}}_{\widetilde{C}}}\max_{\widetilde{E}\in\xi}\ &\mathrm{Tr}\Big\{(\boldsymbol{W}\otimes\boldsymbol{I}_{N_D})\Big(\sigma_h^{-2}(\boldsymbol{W}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_S^T\boldsymbol{W}^{-\frac{1}{2}}+\overline{\boldsymbol{E}}_S)^{-1}\\&\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1}+\sigma_n^{-2}\overline{\boldsymbol{R}}_{\widetilde{C}}\otimes(\widehat{\boldsymbol{\Psi}}_D+\boldsymbol{E}_{ND})^{-1}\Big)^{-1}\Big\}\\&+\mathrm{Tr}\Big\{(\boldsymbol{W}\otimes\boldsymbol{I}_{N_D})\big((\hat{\sigma}_h^2\boldsymbol{W}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_S^T\boldsymbol{W}^{-\frac{1}{2}}+\overline{\overline{\boldsymbol{E}}}_S)^{-1}\\&\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1}+\sigma_n^{-2}\overline{\boldsymbol{R}}_{\widetilde{C}}\otimes(\widehat{\boldsymbol{\Psi}}_D\\&+\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D+\widetilde{\boldsymbol{E}}_{ND})^{-1}\Big)^{-1}\Big\}\\ s.t.\ &\mathrm{Tr}(\overline{\boldsymbol{R}}_{\widetilde{C}})\le P'_S,\ \ \overline{\boldsymbol{R}}_{\widetilde{C}}\succeq\boldsymbol{0}\end{aligned}\tag{34}$$

where

$$\begin{aligned}\widetilde{E}&=\{\overline{\boldsymbol{E}}_S,\overline{\overline{\boldsymbol{E}}}_S,\boldsymbol{E}_D,\boldsymbol{E}_{ND},\widetilde{\boldsymbol{E}}_{ND}\},\\ \overline{\boldsymbol{E}}_S&=\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{E}_S^T\boldsymbol{W}^{-\frac{1}{2}},\\ \overline{\overline{\boldsymbol{E}}}_S&=\boldsymbol{W}^{-\frac{1}{2}}\widetilde{\boldsymbol{E}}_S^T\boldsymbol{W}^{-\frac{1}{2}}.\end{aligned}\tag{35}$$

Following the philosophy of worst case robustness optimization, in the sequel, we consider the new unitarily-invariant channel errors $\{\overline{\boldsymbol{E}}_S,\overline{\overline{\boldsymbol{E}}}_S\}$ instead of $\{\boldsymbol{E}_S,\widetilde{\boldsymbol{E}}_S\}$ for the problem (34). Particularly, the uncertainty size of $\{\overline{\boldsymbol{E}}_S,\overline{\overline{\boldsymbol{E}}}_S\}$ is mainly determined by the specific type of the unitarily-invariant uncertainty set $\{\boldsymbol{E}_S,\widetilde{\boldsymbol{E}}_S\}$. Taking the Frobenius norm constrained $\boldsymbol{E}_S$, i.e., $\|\boldsymbol{E}_S\|_F\le\xi_f$, as an example, we have $\|\overline{\boldsymbol{E}}_S\|_F\le\lambda_{\max}(\widetilde{\boldsymbol{W}})\xi_f$, where $\widetilde{\boldsymbol{W}}=(\boldsymbol{W}^{-1})^T\otimes\boldsymbol{W}^{-1}$.

In order to achieve the low-complexity upper bound optimization for problem (34), we also utilize the identity $\boldsymbol{W}\preceq\lambda_{\max}(\boldsymbol{W})\boldsymbol{I}$ to relax the correlation $\widehat{\boldsymbol{\Sigma}}_S^T$ related positive semidefinite matrix $\boldsymbol{W}_S=\boldsymbol{W}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_S^T\boldsymbol{W}^{-\frac{1}{2}}$ in problem (34) to be $\lambda_{max}(\boldsymbol{W}_S)\boldsymbol{I}_{N_S}$. Then the correlation-scaled upper bound optimization of the problem (34) is given by

$$\begin{aligned}
&\underset{\overline{\boldsymbol{R}}_{\widetilde{C}}}{min}\ \underset{\widetilde{E}\in\xi}{max}\ \operatorname{Tr}\Big\{(\boldsymbol{W}\otimes\boldsymbol{I}_{N_D})\Big(\sigma_h^{-2}(\lambda_{max}(\boldsymbol{W}_S)\boldsymbol{I}_{N_S}+\overline{\boldsymbol{E}}_S)^{-1}\\
&\quad\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1}+\sigma_n^{-2}\overline{\boldsymbol{R}}_{\widetilde{C}}\otimes(\widehat{\boldsymbol{\Psi}}_D+\boldsymbol{E}_{ND})^{-1}\Big)^{-1}\Big\}\\
&\quad+\operatorname{Tr}\Big\{(\boldsymbol{W}\otimes\boldsymbol{I}_{N_D})\Big((\hat{\sigma}_h^2\lambda_{max}(\boldsymbol{W}_S)\boldsymbol{I}_{N_S}+\overline{\overline{\boldsymbol{E}}}_S)^{-11}\\
&\quad\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1}+\sigma_n^{-2}\overline{\boldsymbol{R}}_{\widetilde{C}}\otimes(\widehat{\boldsymbol{\Psi}}_D+\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D\\
&\quad+\widetilde{\boldsymbol{E}}_{ND})^{-1}\Big)^{-1}\Big\}\\
&s.t.\quad \operatorname{Tr}(\overline{\boldsymbol{R}}_{\widetilde{C}})\le P'_S,\quad \overline{\boldsymbol{R}}_{\widetilde{C}}\succeq\boldsymbol{0}
\end{aligned}\tag{36}$$

Then based on the EVDs in (29) and the EVD $\boldsymbol{W}=\boldsymbol{U}_W\boldsymbol{\Lambda}_W\boldsymbol{U}_W^H$, where $\boldsymbol{U}_W\in\mathbb{C}^{N_S\times N_S}$ is the unitary eigenvector matrix and the diagonal matrix $\boldsymbol{\Lambda}_W=\operatorname{diag}[\lambda_{W,1},\cdots,\lambda_{W,N_S}]$ contains the $N_S$ eigenvalues of $\boldsymbol{W}$, the problem (36) is equivalent to the following one

$$\begin{aligned}
&\underset{\overline{\boldsymbol{R}}_{\widetilde{C}}}{min}\ \underset{\widetilde{E}\in\xi'}{max}\ \operatorname{Tr}\Big\{(\boldsymbol{\Lambda}_W\otimes\boldsymbol{I}_{N_D})(\sigma_h^{-2}(\lambda_{max}(\boldsymbol{W}_S)\boldsymbol{I}_{N_S}+\overline{\boldsymbol{E}}'_S)^{-1}\\
&\quad\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}'_D)^{-1}+\sigma_n^{-2}\boldsymbol{U}_W^H\overline{\boldsymbol{R}}_{\widetilde{C}}\boldsymbol{U}_W\\
&\quad\otimes(\boldsymbol{\Omega}_D+\boldsymbol{E}'_{ND})^{-1}\Big)^{-1}\Big\}\\
&\quad+\operatorname{Tr}\Big\{(\boldsymbol{\Lambda}_W\otimes\boldsymbol{I}_{N_D})\Big((\hat{\sigma}_h^2\lambda_{max}(\boldsymbol{W}_S)\boldsymbol{I}_{N_S}+\overline{\overline{\boldsymbol{E}}}'_S)^{-1}\\
&\quad\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}'_D)^{-1}\\
&\quad+\sigma_n^{-2}\boldsymbol{U}_W^H\overline{\boldsymbol{R}}_{\widetilde{C}}\boldsymbol{U}_W\otimes(\boldsymbol{\Omega}_D+\hat{\sigma}_r^2\boldsymbol{\Lambda}_D+\widetilde{\boldsymbol{E}}'_{ND})^{-1}\Big)^{-1}\Big\}\\
&s.t.\quad \operatorname{Tr}(\overline{\boldsymbol{R}}_{\widetilde{C}})\le P'_S,\quad \overline{\boldsymbol{R}}_{\widetilde{C}}\succeq\boldsymbol{0}
\end{aligned}\tag{37}$$

where $\overline{\boldsymbol{E}}'_S=\boldsymbol{U}_W^H\overline{\boldsymbol{E}}_S\boldsymbol{U}_W$ and $\overline{\overline{\boldsymbol{E}}}'_S=\boldsymbol{U}_W^H\overline{\overline{\boldsymbol{E}}}_S\boldsymbol{U}_W$. Similarly, we can also prove that the optimal solutions of $\{\boldsymbol{U}_W^H\overline{\boldsymbol{R}}_{\widetilde{C}}\boldsymbol{U}_W,\boldsymbol{E}'_D,\ \overline{\boldsymbol{E}}'_S,\overline{\overline{\boldsymbol{E}}}'_S,\boldsymbol{E}'_{ND},\widetilde{\boldsymbol{E}}'_{ND}\}$ of problem (37) are diagonal, which is summarized in the following **Theorem 3**. It is worth noting that the proof for **Theorem 3** is similar to that for **Theorem 2**.

*Theorem 3:* For the problem (37), the optimal variables are also diagonal structured, which means

$$\{\boldsymbol{U}_W^H\overline{\boldsymbol{R}}_{\widetilde{C}}^{opt}\boldsymbol{U}_W,\widetilde{E}^{opt}\}=\{\boldsymbol{\Lambda}_C,\boldsymbol{\Lambda}_{E'_D},\boldsymbol{\Lambda}_{\overline{E}'_S},\boldsymbol{\Lambda}_{\overline{\overline{E}}'_S},\boldsymbol{\Lambda}_{E'_{ND}},\boldsymbol{\Lambda}_{\widetilde{E}'_{ND}}\}.\tag{38}$$

Therefore, the matrix-variable problem (37) can be significantly simplified as

$$\begin{aligned}
&\underset{\boldsymbol{\Lambda}_C}{min}\ \underset{\widetilde{\Lambda}_E\in\xi}{\max}\sum_{i=1}^{N_S}\sum_{j=1}^{N_D}\Big(\lambda_{W,i}(\sigma_h^{-2}(\lambda_{\max}(\boldsymbol{W}^*)+\lambda_{\overline{E}'_S,i})^{-1}(\lambda_{D,j}\\
&\quad+\lambda_{E'_D,j})^{-1}+\sigma_n^{-2}\lambda_{C,i}(\Omega_{D,j}+\lambda_{E'_{ND},j})^{-1})^{-1}\Big)\\
&\quad+\sum_{i=1}^{N_S}\sum_{j=1}^{N_D}\Big(\lambda_{W,i}((\hat{\sigma}_h^{-2}\lambda_{\max}(\boldsymbol{W}^*)+\lambda_{\overline{\overline{E}}'_S,i})^{-1}(\lambda_{D,j}\\
&\quad+\lambda_{E'_D,j})^{-1}+\sigma_n^{-2}\lambda_{C,i}(\Omega_{D,j}+\hat{\sigma}_r^{2}\lambda_{D,j}+\lambda_{\widetilde{E}'_{ND},j})^{-1}\Big)^{-1}\\
&s.t.\quad \sum_{i=1}^{N_S}\lambda_{C,i}\le P'_S,\quad \lambda_{C,i}\ge 0,\ i=1,\cdots,N_S.
\end{aligned}\tag{39}$$

where $\widetilde{\Lambda}_E=\{\boldsymbol{\Lambda}_{E'_D},\boldsymbol{\Lambda}_{\overline{E}'_S},\boldsymbol{\Lambda}_{\overline{\overline{E}}'_S},\boldsymbol{\Lambda}_{E'_{ND}},\boldsymbol{\Lambda}_{\widetilde{E}'_{ND}}\}$. By utilizing the identity (38), we can further derive the semi-closed-form $\{\boldsymbol{C},E\}$ of the correlation-scaled-scheme as

$$\begin{aligned}
\boldsymbol{C}&=(\boldsymbol{W}^{-\frac{1}{2}})^*\widehat{\boldsymbol{U}}_W^*\boldsymbol{\Lambda}_C^{\frac{1}{2}}\boldsymbol{Q}_A(\boldsymbol{Q}_H^{[1:N_S]})^T,\\
\boldsymbol{E}_D&=\widehat{\boldsymbol{U}}_D\boldsymbol{\Lambda}_{E'_D}\widehat{\boldsymbol{U}}_D^H,\\
\boldsymbol{E}_S&=\boldsymbol{W}^{\frac{1}{2}}\widehat{\boldsymbol{U}}_W\boldsymbol{\Lambda}_{\overline{E}'_S}\widehat{\boldsymbol{U}}_W^H\boldsymbol{W}^{\frac{1}{2}},\\
\widetilde{\boldsymbol{E}}_S^T&=\boldsymbol{W}^{\frac{1}{2}}\widehat{\boldsymbol{U}}_W\boldsymbol{\Lambda}_{\overline{\overline{E}}'_S}\widehat{\boldsymbol{U}}_W^H\boldsymbol{W}^{\frac{1}{2}}d,\\
\boldsymbol{E}_{ND}&=\widehat{\boldsymbol{U}}_D\boldsymbol{\Lambda}_{E'_{ND}}\widehat{\boldsymbol{U}}_D^H,\\
\widetilde{\boldsymbol{E}}_{ND}&=\widehat{\boldsymbol{U}}_D\boldsymbol{\Lambda}_{\widetilde{E}'_{ND}}\widehat{\boldsymbol{U}}_D^H.
\end{aligned}\tag{40}$$

*Proof:* Please see the proof for **Theorem 2** in Appendix B.

Based on **Theorem 2** and **3**, it is obvious that both the dimensions of optimization variables in the scalarized weighting-scaled problem (32) and the scalarized correlation-scaled problem (39) are significantly reduced compared to the proposed iterative SDP optimization algorithm, i.e., from $3N_S^2+3N_D^2$ to $3N_S+3N_D$, which realizes the low computational complexity. Besides, it is clear that the problems (32) and (39) are convex w.r.t the arbitrary variable in $\{\boldsymbol{\Lambda}_C,\Lambda_E\}$ and $\{\boldsymbol{\Lambda}_C,\widetilde{\Lambda}_E\}$, respectively, when the remaining variables are fixed. Therefore, the proposed iterative SDP optimization is still suitable to the simplified problems (32) and (39). Particularly, the low-complexity IBS algorithm in [52] can further be applied to derive the optimal training $\boldsymbol{C}$ for both two problems. To proceed, we take the problem (32) as an example to introduce the low-complexity IBS algorithm. Then the problem (39) can also be addressed by a similar logic. For the problem (32), we firstly utilize the lagrangian dual method to derive the optimal $\lambda_{C,i}$, $\forall i=1,\cdots,N_S$ as (41) as shown at the bottom of the next page, where $\eta\ge 0$ is the lagrangian multiplier of the weighted power constraint. Observing from (41), we find that $\eta$ is monotonically decreasing w.r.t $\lambda_{C,i}$, $\forall\ i=1,\cdots,N_S$, which implies that the IBS method in [52] can be applied to search the optimal $\lambda_{C,i}$ both satisfying (41) and (42), as shown at the bottom of the next page. The interested readers can refer to literature [52] for details.

*Remark 2:* Assuming that all considered statistical errors belong to the Nuclear norm constrained uncertainty set, i.e., $\operatorname{Tr}((\boldsymbol{\Lambda}\boldsymbol{\Lambda}^H)^{\frac{1}{2}})\le\zeta_u,\forall\boldsymbol{\Lambda}\in\Lambda_E/\widetilde{\Lambda}_E$, then the low-complexity IBS algorithm can also be applied to optimize the arbitrary statistical error term $\boldsymbol{\Lambda}\in\Lambda_E/\widetilde{\Lambda}_E$ for the simplified problem (32)/(39). Besides, to demonstrate the effectiveness of the proposed two low-complexity upper bound schemes, the training design with perfect channel and noise statistical information, i.e., $E=\emptyset$, is considered as a benchmark in our work, which is in essence an effective lower-bound for the LMMSE based robust training optimization.

## III. EXTENSION TO THE MUTUAL INFORMATION MAXIMIZATION

Another widely used criterion for training design is to maximize mutual information between the true channel $\boldsymbol{h}$ and the estimated channel $\hat{\boldsymbol{h}}$, i.e., $I(\boldsymbol{h}, \hat{\boldsymbol{h}})$. Because $I(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \log\det(\boldsymbol{\Psi}_M^{-1})$ [38], similarly to the formulation of the LMMSE based robust training optimization problem (14), the robust mutual information maximization problem under general power constraints is formulated as follows

$$\begin{aligned}&\max_{\boldsymbol{C}}\ \min_{E\in\xi}\ \log\det(\widetilde{\boldsymbol{R}}_h^{-1}+\bar{\boldsymbol{C}}^H\bar{\boldsymbol{Q}}_H\widetilde{\boldsymbol{R}}_n^{-1}\bar{\boldsymbol{Q}}_H^H\bar{\boldsymbol{C}}),\\ &s.t.\ \mathrm{Tr}\Big(\boldsymbol{W}_l(\boldsymbol{X}+\boldsymbol{C})(\boldsymbol{X}+\boldsymbol{C})^H\Big)\le P_{S_l},\quad \forall\, l=1,\cdots,L_p,\end{aligned}\tag{43}$$

Naturally, by performing the dimension-reduced operation and applying **Theorem 1**, the problem (43) can be equivalently simplified as

$$\begin{aligned}&\max_{\boldsymbol{R}_{\widetilde{C}}}\ \min_{E\in\xi}\ \log\det\big(\sigma_h^{-2}(\widehat{\boldsymbol{\Sigma}}_S^T+\boldsymbol{E}_S^T)^{-1}\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1}\\ &\qquad+\sigma_n^{-2}\boldsymbol{R}_{\widetilde{C}}\otimes(\widehat{\boldsymbol{\Psi}}_D+\boldsymbol{E}_{ND})^{-1}\big)\\ &\qquad+\log\det\Big((\hat{\sigma}_h^2\widehat{\boldsymbol{\Sigma}}_S^T+\widetilde{\boldsymbol{E}}_S^T)^{-1}\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1}\\ &\qquad+\sigma_n^{-2}\boldsymbol{R}_{\widetilde{C}}\otimes\big(\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D+\widehat{\boldsymbol{\Psi}}_D+\widetilde{\boldsymbol{E}}_{ND}\big)^{-1}\Big)\\ &s.t.\ \mathrm{Tr}(\boldsymbol{W}^*\boldsymbol{R}_{\widetilde{C}})\le P_S',\quad \boldsymbol{R}_{\widetilde{C}}\succeq\boldsymbol{0}.\end{aligned}\tag{44}$$

It can be proved that the problem (44) is concave w.r.t. $\boldsymbol{R}_{\widetilde{C}}$ and the arbitrary variable in the set of $E$. As a result, the proposed iterative SDP optimization algorithm is still applicable to the problem (44). Besides, for mutual information maximization, both the weighting-scaled-scheme and the correlation-scaled scheme become an effective lower bound of the problem (44). In particular, the weighting-scaled-scheme can be directly extended to the problem (44) and the corresponding optimal solutions with diagonal structure are available similarly to that for the problem (30). Here, we omit this similar derivation process due to the space limitation. Nevertheless, considering the characteristics of function $\log\det(\boldsymbol{A})$, the relaxation of the matrix $\boldsymbol{W}_S = \boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{\Sigma}_S^T\boldsymbol{W}^{-\frac{1}{2}}$ is not required when applying the correlation-scaled-scheme to the problem (44). To be specific, similarly to the formulation of the problem (34), we firstly utilize the definition $\overline{\boldsymbol{R}}_{\widetilde{C}} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{R}_{\widetilde{C}}\boldsymbol{W}^{\frac{1}{2}}$ to equivalently rewrite the problem (44) as

$$\begin{aligned}&\max_{\overline{\boldsymbol{R}}_{\widetilde{C}}}\ \min_{\widetilde{E}\in\xi'}\ \log\det\Big(\sigma_h^{-2}(\boldsymbol{W}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_S^T\boldsymbol{W}^{-\frac{1}{2}}+\overline{\boldsymbol{E}}_S)^{-1}\\ &\qquad\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1}+\sigma_n^{-2}\overline{\boldsymbol{R}}_{\widetilde{C}}\otimes(\widehat{\boldsymbol{\Psi}}_D+\boldsymbol{E}_{ND})^{-1}\Big)\\ &\qquad+\log\det\Big((\hat{\sigma}_h^2\boldsymbol{W}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_S^T\boldsymbol{W}^{-\frac{1}{2}}+\widetilde{\overline{\boldsymbol{E}}}_S)^{-1}\\ &\qquad\otimes(\widehat{\boldsymbol{\Sigma}}_D+\boldsymbol{E}_D)^{-1}+\sigma_n^{-2}\overline{\boldsymbol{R}}_{\widetilde{C}}\otimes(\widehat{\boldsymbol{\Psi}}_D+\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D+\widetilde{\boldsymbol{E}}_{ND})^{-1}\Big)\\ &s.t.\ \mathrm{Tr}(\overline{\boldsymbol{R}}_{\widetilde{C}})\le P_S',\ \overline{\boldsymbol{R}}_{\widetilde{C}}\succeq\boldsymbol{0},\end{aligned}\tag{45}$$

where the constant term $\log\det(\boldsymbol{W}^{-1}\otimes\boldsymbol{I})$ has been removed and the new unitarily-invariant uncertainty sets $\{\overline{\boldsymbol{E}}_S, \widetilde{\overline{\boldsymbol{E}}}_S\}$ are also considered. Further by defining the EVD $\boldsymbol{W}_S = \boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{\Sigma}_S^T\boldsymbol{W}^{-\frac{1}{2}} = \boldsymbol{U}_{W_S}\boldsymbol{\Lambda}_{W_S}\boldsymbol{U}_{W_S}^H$, where $\boldsymbol{U}_{W_S}\in\mathbb{C}^{N_S\times N_S}$ is the unitary eigenvector matrix and the diagonal matrix $\boldsymbol{\Lambda}_{W_S}$ contains $N_S$ eigenvalues of $\boldsymbol{W}_S$, the problem (45) is equivalently transformed into

$$\begin{aligned}&\max_{\overline{\boldsymbol{R}}_{\widetilde{C}}}\ \min_{\widetilde{E}\in\xi'}\ \log\det\Big(\sigma_h^{-2}(\boldsymbol{\Lambda}_{W_S}+\overline{\boldsymbol{E}}_S')^{-1}\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}_D')^{-1}\\ &\qquad+\sigma_n^{-2}\boldsymbol{U}_{W_S}^H\overline{\boldsymbol{R}}_{\widetilde{C}}\boldsymbol{U}_{W_S}\otimes(\boldsymbol{\Omega}_D+\boldsymbol{E}_{ND}')^{-1}\Big)\\ &\qquad+\log\det\Big((\hat{\sigma}_h^2\boldsymbol{\Lambda}_{W_S}+\widetilde{\overline{\boldsymbol{E}}}_S')^{-1}\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}_D')^{-1}\\ &\qquad+\sigma_n^{-2}\boldsymbol{U}_{W_S}^H\overline{\boldsymbol{R}}_{\widetilde{C}}\boldsymbol{U}_{W_S}\otimes(\boldsymbol{\Omega}_D+\hat{\sigma}_r^2\boldsymbol{\Lambda}_D+\widetilde{\boldsymbol{E}}_{ND}')^{-1}\Big)\\ &s.t.\ \mathrm{Tr}(\overline{\boldsymbol{R}}_{\widetilde{C}})\le P_S',\quad \overline{\boldsymbol{R}}_{\widetilde{C}}\succeq\boldsymbol{0},\end{aligned}\tag{46}$$

where $\overline{\boldsymbol{E}}_S' = \boldsymbol{U}_{W_S}^H\overline{\boldsymbol{E}}_S\boldsymbol{U}_{W_S}$ and $\widetilde{\overline{\boldsymbol{E}}}_S' = \boldsymbol{U}_{W_S}^H\widetilde{\overline{\boldsymbol{E}}}_S\boldsymbol{U}_{W_S}$. Note that the diagonal structured optimal solutions of $\{\boldsymbol{U}_{W_S}^H\overline{\boldsymbol{R}}_{\widetilde{C}}\boldsymbol{U}_{W_S}, \widetilde{E}\}$ of problem (46) can also be obtained similarly to that in **Theorem 2**. Particularly, for the robust mutual information maximization, the training design with the perfect channel and noise statistical information becomes an effective upper bound of the problem (44).

---

$$\begin{aligned}&\sum_{j=1}^{N_D}\frac{\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}(\Omega_{D,j}+\lambda_{E_{ND,j}'})^{-1}}{\big(\sigma_h^{-2}(\lambda_{S,i}+\lambda_{E_{S,i}'})^{-1}(\lambda_{D,j}+\lambda_{E_{D,j}'})^{-1}+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\lambda_{C,i}(\Omega_{D,j}+\lambda_{E_{ND,j}'})^{-1}\big)^2}\\ &\quad+\sum_{j=1}^{N_D}\frac{\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}(\Omega_{D,j}+\hat{\sigma}_r^2\lambda_{D,j}+\lambda_{\widetilde{E}_{ND,j}'})^{-1}}{\big((\hat{\sigma}_h^2\lambda_{S,i}+\lambda_{\widetilde{E}_{S,i}'})^{-1}(\lambda_{D,j}+\lambda_{E_{D,j}'})^{-1}+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\lambda_{C,i}(\Omega_{D,j}+\hat{\sigma}_r^2\lambda_{D,j}+\lambda_{\widetilde{E}_{ND,j}'})^{-1}\big)^2}=\eta\end{aligned}\tag{41}$$

$$\sum_{i=1}^{N_S}\lambda_{C,i}=P_S'\tag{42}$$

## IV. NUMERICAL RESULTS AND DISCUSSIONS

In this section, some numerical results are given to assess the performance of the proposed robust training designs for cooperative communications in terms of both channel estimation MSE and mutual information. Without loss of generality, in our simulation the three nodes in the cooperative network are all equipped with 4 antennas, i.e., $N_S = N_R = N_D = 4$. The transmit and receive correlation matrices of the channel matrices are generated according to the widely used exponential model in our simulation. More specifically, the correlation matrices $\widehat{\boldsymbol{\Sigma}}_S$, $\widehat{\boldsymbol{\Sigma}}_R$ and $\widehat{\boldsymbol{\Sigma}}_D$ in the simulation setting are denoted as

$$[\widehat{\boldsymbol{\Sigma}}_S]_{i,j}=p_s^{j-i}; \quad [\widehat{\boldsymbol{\Sigma}}_R]_{i,j}=p_r^{j-i}, \ [\widehat{\boldsymbol{\Sigma}}_D]_{i,j}=p_d^{j-i}, \ \forall j \geq i, \quad (47)$$

where the correlation parameters $p_l$, $l \in \{s, r, d\}$ are the complex correlation coefficients with $|p_l| = 0.5$. The correlation matrices of the colored noises at the relay and receiver are modeled as $[\widehat{\boldsymbol{\Psi}}_R]_{i,j}=q_r^{j-i}$, $[\widehat{\boldsymbol{\Psi}}_T]_{i,j}=q_t^{j-i}$, $\forall j \geq i$, where $|q_l| = 0.5$, $l \in \{t, r\}$. As discussed above, both the eigenvectors of $\widehat{\boldsymbol{\Psi}}_D$ and $\widehat{\boldsymbol{\Sigma}}_D$ are assumed to be the same. In other words, the received colored noise has the identical spatial correlation as that of channel, i.e., $\widehat{\boldsymbol{\Psi}}_D=\widehat{\boldsymbol{U}}_D\boldsymbol{\Omega}_D\widehat{\boldsymbol{U}}_D^H$, where $\widehat{\boldsymbol{U}}_D$ is derived from $\widehat{\boldsymbol{\Sigma}}_D = \widehat{\boldsymbol{U}}_D\boldsymbol{\Lambda}_D\widehat{\boldsymbol{U}}_D^H$ and the diagonal $\boldsymbol{\Omega}_D$ is specified to guarantee the diagonal elements of $\widehat{\boldsymbol{\Psi}}_D$ being 1.

Besides, due to the fact that the general power constraints are considered in our work, we assume that the maximum transmit power corresponding to each weighted power constraint is $P_t$ dBW. Besides, the variance of all channel and noise elements are assumed to be $\sigma_h^2 = \sigma_n^2 = 1$, and thus the training SNR is defined as SNR $= P_t$. Particularly, in this simulation, the unitarily-invariant channel and noise statistical errors are specified as the Frobenius norm constrained uncertainty sets and the corresponding error threshold $\xi_f$ is initialized to be $\xi_f = 0.5$. Moreover, both the normalized worst case MSE w.r.t. the effective channel size $2N_DN_S$ and the robust mutual information are adopted as the channel estimation performance metrics. Further, to demonstrate the effectiveness and advantages of the proposed robust training designs, both the perfect training design and the non-robust training design are adopted as comparisons. Specifically, for the non-robust training design, we firstly solve the outer minimization subproblem of (16) with $E = \emptyset$, then substitute the obtained $\boldsymbol{C}$ and the worst case statistical errors derived from the proposed robust training designs into (16) to calculate the non-robust channel estimation MSE. While for the perfect training design, the min-max problem (16) is actually reduced to be a min problem considering $E = \emptyset$. Finally, in order to effectively solve the involved SDP problems (21) and (23), the famous matlab software toolbox CVX [60] is utilized.

Firstly, in Fig. 2, we show that the performance of worst case MSE as a function of training SNR under a total power constraint, i.e., $\boldsymbol{W}_l = \boldsymbol{I}_{N_S}, \forall\, l = 1, \cdots, 4$. It can be seen that the achievable worst case MSE decreases with SNR for all proposed training designs. Our proposed weighting-scaled-scheme in **Theorem 2** is tight and equivalent to the robust training design in [34] in essence. As mentioned before, literature [34] considered both the direct link and complete channel correlations for AF MIMO relaying networks. As a result, we also observe that the worst case MSE achieved by the proposed iterative SDP optimization and the weighting-scaled-scheme in **Theorem 2** are almost identical to that of the robust training design in [34]. Besides, the correlation-scaled-scheme in **Theorem 3** achieves higher worst case MSE than that of the above three schemes due to the relaxation in $\boldsymbol{\Sigma}_S$. As SNR increases, the correlation-scaled-scheme is closer to the iterative SDP optimization. However, all of them outperform the non-robust training design by considering the optimization of the worst case channel and noise statistical errors. In addition, as an effective lower bound, the perfect training design naturally realizes the minimum MSE among all proposed training designs.

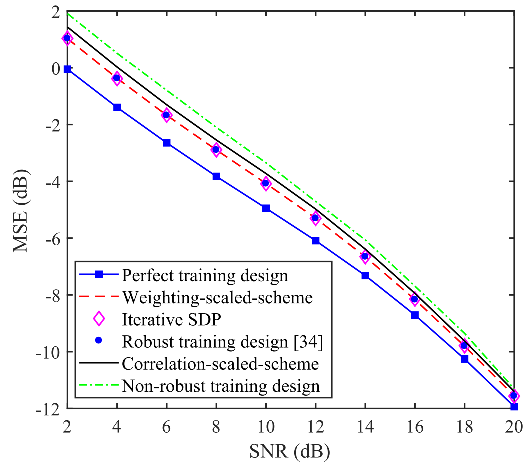


**FIGURE 2.** Worst case MSE performance versus SNR under total power constraint, where $E_l = I_{N_S}, \forall l = 1, \cdots, 4$.

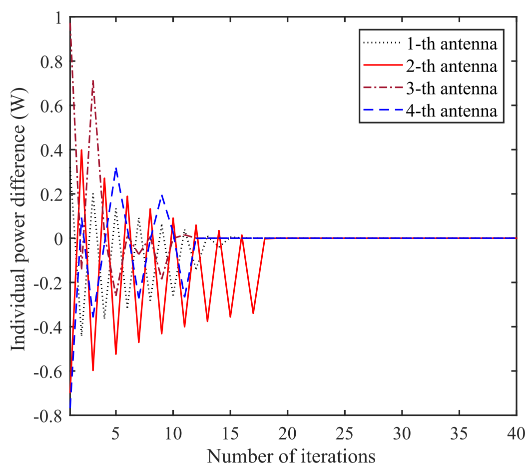


**FIGURE 3.** Convergence rate of the subgradient method in Theorem 1.

In order to demonstrate our proposed **Theorem 1**, in Fig. 3, we show the convergence performance of the subgradient method, which is used for searching $\mu_k's$ satisfying the general power constraints. To be specific,

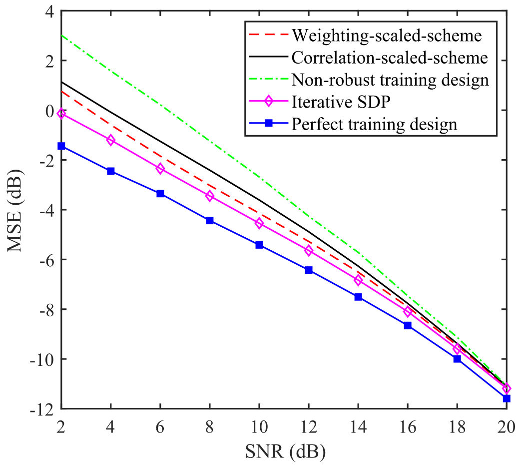


**FIGURE 4.** **Worst case MSE versus SNR under the combined total and individual power constraints, where $W_l = \mathrm{diag}[0_{1\times l-1}, 1, 0_{1\times 4-l}],\ \forall\, l = 1, \cdots, 4;\ W_5 = I_{N_S}$.**

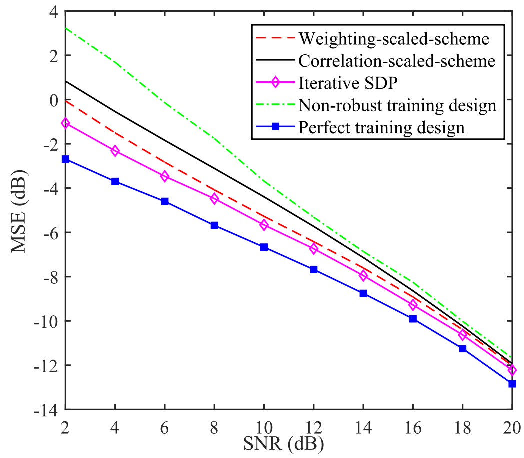


**FIGURE 5.** **Worst case MSE versus SNR under the more general power constraints,i.e., $W_1 = \mathrm{diag}[1, 0, 0, 0]$, $W_2 = \mathrm{diag}[0, 1, 0, 0]$, and $W_3 = \mathrm{diag}[1, 1, 0, 0]$.**

we consider the individual power constraints, i.e., $\boldsymbol{W}_l = \mathrm{diag}[0_{1\times l-1}, 1, 0_{1\times 4-l}]$, $\forall l = 1, \cdots, 4$, and adopt the weighting-scaled-scheme in **Theorem 2** as an example. Then in Fig. 3, the difference between the actual transmit power of each antenna and its corresponding maximum threshold, i.e., $\mathrm{Tr}(\boldsymbol{W}_l \widetilde{\boldsymbol{C}}\widetilde{\boldsymbol{C}}^H) - P'_{S_l}\ \forall l = 1, \cdots, 4$, is shown. We clearly observe that the subgradient method converges after around $10 \sim 20$ iterations with all individual power constraints being activated.

Following Fig. 3, in Fig. 4, we investigate the worst case MSE versus SNR under the combined individual power constraints and total power constraint, i.e., $\boldsymbol{W}_l = \mathrm{diag}[0_{1\times l-1}, 1, 0_{1\times 4-l}]$, $\forall\, l = 1, \cdots, 4$; $\boldsymbol{W}_5 = \boldsymbol{I}_{N_S}$. From this figure, it is observed that under the combined individual power constraints and total power constraint, the proposed iterative SDP algorithm outperforms the weighting-scaled-scheme in **Theorem 2** and the correlation-scaled-scheme in **Theorem 3**, which dues to the relaxation operation in the proposed low-complexity upper-bound optimization. Besides, the weighting-scaled-scheme in **Theorem 2** achieves the lower worst case MSE than that of the correlation-scaled-scheme in **Theorem 3**. However, with the increase of SNR, both the two upper-bound optimization schemes perform closer to the proposed iterative SDP optimization. Undoubtedly, among all studied robust training designs, the perfect training design still achieves the lowest channel estimation MSE, while the non-robust training design realizes the highest worst case MSE.

In fact, our proposed robust training designs are also applicable to the more general power constraints case. For example, a virtual MIMO transmitter consisting of three sources is considered, where two of the three sources are assumed to be single antenna, while the remaining one has two antennas. Then the total transmit power of each source is specified by $\boldsymbol{W}_1 = \mathrm{diag}[1, 0, 0, 0]$, $\boldsymbol{W}_2 = \mathrm{diag}[0, 1, 0, 0]$, and $\boldsymbol{W}_3 = \mathrm{diag}[1, 1, 0, 0]$, respectively. Based on this, the worst case

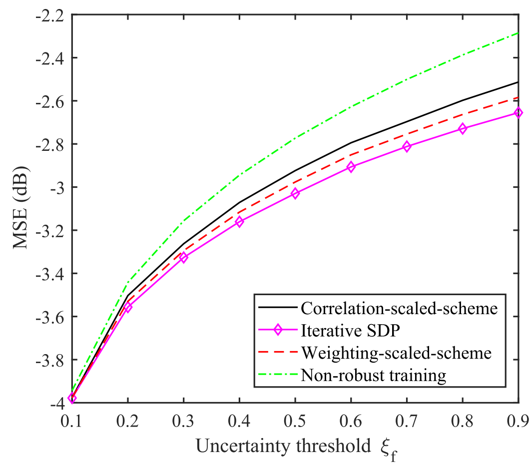


**FIGURE 6.** **Worst case MSE versus channel uncertainty parameter under total power constraint, where $E_l = I_{N_S}, \forall l = 1, \cdots, 4$.**

MSE versus SNR for all proposed robust training designs is shown in Fig. 5. Naturally, the almost identical worst case MSE performance as that of Fig. 4 is observed. Among the three proposed algorithms, the correlation-scaled-scheme outperforms the weighting-scaled-scheme algorithm, and the iterative SDP algorithm is optimal.

Further, Fig. 6 shows the worst case MSE as a function of the channel and noise statistical error threshold $\xi_f$, where the combined individual and total power constraints are considered. From Fig. 6, it is observed that the achievable worst case MSE increases with the uncertainty parameter $\xi_f$ for all studied robust training designs. As expected, among the proposed three robust training designs, the proposed iterative SDP optimization still performs best in terms of the worst case MSE, while the correlation-scaled-scheme in **Theorem 3** performs worst. In addition, the non-robust training design naturally achieves the highest worst case MSE especially under the large value of $\xi_f$.

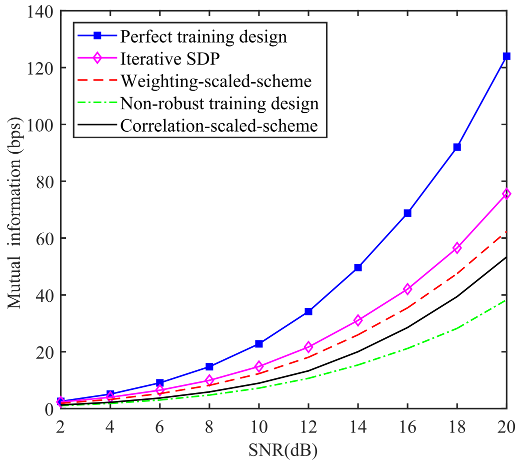


**FIGURE 7. Worst case mutual information versus SNR under the combined total and individual power constraints, where $W_l = \text{diag}[0_{1\times l-1}, 1, 0_{1\times 4-l}],\ \forall l = 1, \cdots, 4;\ W_5 = I_{N_S}$.**

Finally, we illustrate the worst case maximum mutual information versus SNR under the combined individual and total power constraints in Fig. 7. It can be seen that the achievable mutual information increases with SNR for all studied training designs. Similarly to that in Fig. 4, it is clearly seen that the the proposed iterative SDP optimization outperforms the weighting-scaled-scheme in **Theorem 2** and the error-scaled-scheme in **Theorem 3**. Besides, the perfect training design still performs best in terms of the achievable maximum mutual information, while the non-robust training design performs worst.

## V. CONCLUSIONS

In this paper, the robust training design for AF MIMO relaying channel estimation has been investigated. Specifically, we consider the channel and colored noise statistical errors subject to the unitarily-invariant uncertainty set, then the LMMSE based robust training optimization problem is formulated, which is generally non-convex and intractable. To solve this problem effectively, we propose an iterative SDP optimization and two low-complexity upper-bound optimization schemes, respectively. In particular, the diagonal structured optimal solutions of the proposed two upper-bound optimization schemes are available, and thus the semi-closed-form training matrix is can be obtained for effective channel estimation. Further, we extend our work into the robust mutual information maximization of the AF MIMO relaying channel, to which all proposed robust training designs are proved to be still applicable. Finally, numerical simulation results are conducted to indicate the effectiveness of the proposed various robust training designs in terms of channel estimation MSE and the achievable mutual information.

## APPENDIX A

Recalling the Kronecker structured channel models in (9), we can further model the imperfect spatial correlation matrices $\boldsymbol{\Sigma}_S$, $\boldsymbol{\Sigma}_R$ and $\boldsymbol{\Sigma}_D$ as

$$\begin{aligned}\boldsymbol{\Sigma}_S &= \widehat{\boldsymbol{\Sigma}}_S + \boldsymbol{E}_S,\\ \boldsymbol{\Sigma}_R &= \widehat{\boldsymbol{\Sigma}}_R + \boldsymbol{E}_R,\\ \boldsymbol{\Sigma}_D &= \widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D.\end{aligned} \tag{48}$$

where the positive-definite Hermitian matrices $\widehat{\boldsymbol{\Sigma}}_S \in \mathbb{C}^{N_S\times N_S}$, $\widehat{\boldsymbol{\Sigma}}_R \in \mathbb{C}^{N_R\times N_R}$ and $\widehat{\boldsymbol{\Sigma}}_D \in \mathbb{C}^{N_D\times N_D}$ are the estimated spatial correlation matrices, respectively. The matrices $\boldsymbol{E}_S \in \mathbb{C}^{N_S\times N_S}$, $\boldsymbol{E}_R \in \mathbb{C}^{N_R\times N_R}$ and $\boldsymbol{E}_D \in \mathbb{C}^{N_D\times N_D}$ denote the correlation error matrices corresponding to $\boldsymbol{\Sigma}_S$, $\boldsymbol{\Sigma}_R$ and $\boldsymbol{\Sigma}_D$, respectively. Then based on (48), the imperfect statistical model of the direct channel $\boldsymbol{H}_{SD}$ is given by

$$\begin{aligned}&\mathbb{E}[\text{vec}(\boldsymbol{H}_{SD})\text{vec}(\boldsymbol{H}_{SD})^H]\\ &\quad= \mathbb{E}[(\boldsymbol{\Sigma}_S^{\frac{1}{2}T} \otimes \boldsymbol{\Sigma}_D^{\frac{1}{2}})\text{vec}(\bar{\boldsymbol{H}}_{SD})\text{vec}(\bar{\boldsymbol{H}}_{SD})^H(\boldsymbol{\Sigma}_S^{\frac{1}{2}T} \otimes \boldsymbol{\Sigma}_D^{\frac{1}{2}})]\\ &\quad= \sigma_h^2(\widehat{\boldsymbol{\Sigma}}_S^T + \boldsymbol{E}_S^T) \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D),\end{aligned} \tag{49}$$

Similarly, the imperfect statistical model of the compound channel $\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{H}_{SR}$ is

$$\begin{aligned}&\mathbb{E}[\text{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{H}_{SR})\text{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{H}_{SR})^H]\\ &\quad= \mathbb{E}\Big[\big(\boldsymbol{\Sigma}_S^{\frac{1}{2}T} \otimes (\boldsymbol{\Sigma}_D^{\frac{1}{2}}\bar{\boldsymbol{H}}_{RD}\boldsymbol{\Sigma}_R^{\frac{1}{2}}\boldsymbol{F}_R\boldsymbol{\Sigma}_R^{\frac{1}{2}})\big)\text{vec}(\bar{\boldsymbol{H}}_{SR})\text{vec}(\bar{\boldsymbol{H}}_{SR})^H\\ &\qquad\times\big(\boldsymbol{\Sigma}_S^{\frac{1}{2}T} \otimes (\boldsymbol{\Sigma}_R^{\frac{1}{2}}\boldsymbol{F}_R^H\boldsymbol{\Sigma}_R^{\frac{1}{2}}\bar{\boldsymbol{H}}_{RD}^H\boldsymbol{\Sigma}_D^{\frac{1}{2}})\big)\Big]\\ &\quad= \sigma_h^2\mathbb{E}\Big[\boldsymbol{\Sigma}_S^T \otimes \big(\boldsymbol{\Sigma}_D^{\frac{1}{2}}\bar{\boldsymbol{H}}_{RD}\boldsymbol{\Sigma}_R^{\frac{1}{2}}\boldsymbol{F}_R\boldsymbol{\Sigma}_R\boldsymbol{F}_R^H\boldsymbol{\Sigma}_R^{\frac{1}{2}}\bar{\boldsymbol{H}}_{RD}^H\boldsymbol{\Sigma}_D^{\frac{1}{2}}\big)\Big]\\ &\quad= \bar{\sigma}_h^2\boldsymbol{\Sigma}_S^T \otimes \boldsymbol{\Sigma}_D = \bar{\sigma}_h^2(\widehat{\boldsymbol{\Sigma}}_S^T + \boldsymbol{E}_S^T) \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D),\end{aligned} \tag{50}$$

where $\bar{\sigma}_h^2 = \sigma_h^4\text{Tr}\big((\widehat{\boldsymbol{\Sigma}}_R + \boldsymbol{E}_R)\boldsymbol{F}_R(\widehat{\boldsymbol{\Sigma}}_R + \boldsymbol{E}_R)\boldsymbol{F}_R^H\big)$. Clearly, the spatial correlation errors $\boldsymbol{E}_R$, $\boldsymbol{E}_S$ are coupled and intractable in the function $\bar{\sigma}_h^2(\widehat{\boldsymbol{\Sigma}}_S^T + \boldsymbol{E}_S^T)$. To facilitate the following analysis, we unify all coupled errors in $\bar{\sigma}_h^2(\widehat{\boldsymbol{\Sigma}}_S^T + \boldsymbol{E}_S^T)$ into a single error matrix $\widetilde{\boldsymbol{E}}_S \in \mathbb{C}^{N_S\times N_S}$. As a result, the formulation (50) can be rewritten as

$$\begin{aligned}&\mathbb{E}[\text{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{H}_{SR})\text{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{H}_{SR})^H]\\ &\qquad\qquad= (\hat{\sigma}_h^2\widehat{\boldsymbol{\Sigma}}_S^T + \widetilde{\boldsymbol{E}}_S^T) \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D).\end{aligned} \tag{51}$$

where $\hat{\sigma}_h^2 = \sigma_h^4\text{Tr}(\widehat{\boldsymbol{\Sigma}}_R\boldsymbol{F}_R\widehat{\boldsymbol{\Sigma}}_R\boldsymbol{F}_R^H)$. Based on (49) and (51), the imperfect channel covariance matrix $\widetilde{\boldsymbol{R}}_h$ in (8) is modeled as the formulation (52), as shown at the bottom of this page.

$$\begin{aligned}\widetilde{\boldsymbol{R}}_h &= \begin{bmatrix}\mathbb{E}[\text{vec}(\boldsymbol{H}_{SD})\text{vec}(\boldsymbol{H}_{SD})^H] & \boldsymbol{0}\\ \boldsymbol{0} & \mathbb{E}[\text{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{H}_{SR})\text{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{H}_{SR})^H]\end{bmatrix}\\ &= \begin{bmatrix}\sigma_h^2(\widehat{\boldsymbol{\Sigma}}_S^T + \boldsymbol{E}_S^T) \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D) & \boldsymbol{0}\\ \boldsymbol{0} & (\hat{\sigma}_h^2\widehat{\boldsymbol{\Sigma}}_S^T + \widetilde{\boldsymbol{E}}_S^T) \otimes (\widehat{\boldsymbol{\Sigma}}_D + \boldsymbol{E}_D)\end{bmatrix}\end{aligned} \tag{52}$$

$$\begin{aligned}\widetilde{\boldsymbol{R}}_n &= \begin{bmatrix}\mathbb{E}[\mathrm{vec}(\boldsymbol{N}_D^{[1]})\mathrm{vec}(\boldsymbol{N}_D^{[1]})^H] & \boldsymbol{0}\\ \boldsymbol{0} & \mathbb{E}[\mathrm{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{N}_R+\boldsymbol{N}_D^{[2]})\mathrm{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{N}_R+\boldsymbol{N}_D^{[2]})^H]\end{bmatrix}\\ &= \begin{bmatrix}\sigma_n^2\boldsymbol{I}_{K+L}\otimes(\widehat{\boldsymbol{\Psi}}_D+\boldsymbol{E}_{ND}) & \boldsymbol{0}\\ \boldsymbol{0} & \sigma_n^2\boldsymbol{I}_{K+L}\otimes(\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D+\widehat{\boldsymbol{\Psi}}_D+\widetilde{\boldsymbol{E}}_{ND})\end{bmatrix}\end{aligned} \tag{53}$$

Furthermore, taking advantage of the kronecker structured colored noise model in (11), the imperfect relay and destination noise correlation matrices $\boldsymbol{\Psi}_D$ and $\boldsymbol{\Psi}_R$ are expressed as

$$\boldsymbol{\Psi}_D = \widehat{\boldsymbol{\Psi}}_D + \boldsymbol{E}_{ND}, \quad \boldsymbol{\Psi}_R = \widehat{\boldsymbol{\Psi}}_R + \boldsymbol{E}_{NR}. \tag{54}$$

Based on (54), the imperfect statistical model of the destination noise $\boldsymbol{N}_D^{[1]}$ is given by

$$\begin{aligned}&\mathbb{E}[\mathrm{vec}(\boldsymbol{N}_D^{[1]})\mathrm{vec}(\boldsymbol{N}_D^{[1]})^H]\\ &\quad= \mathbb{E}[(\boldsymbol{I}_{K+L}\otimes\boldsymbol{\Psi}_D^{\frac{1}{2}})\mathrm{vec}(\bar{\boldsymbol{N}}_D^{[1]})\mathrm{vec}(\bar{\boldsymbol{N}}_D^{[1]})^H(\boldsymbol{I}_{K+L}\otimes\boldsymbol{\Psi}_D^{\frac{1}{2}})]\\ &\quad= \sigma_n^2\boldsymbol{I}_{K+L}\otimes\boldsymbol{\Psi}_D\\ &\quad= \sigma_n^2\boldsymbol{I}_{K+L}\otimes(\widehat{\boldsymbol{\Psi}}_D+\boldsymbol{E}_{ND}),\end{aligned} \tag{55}$$

Besides, similar to the derivation of the imperfect statistical model (50) for the compound channel, the imperfect compound noise $\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{N}_R + \boldsymbol{N}_D^{[2]}$ statistics can be derived to be

$$\begin{aligned}&\mathbb{E}[\mathrm{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{N}_R+\boldsymbol{N}_D^{[2]})\mathrm{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{N}_R+\boldsymbol{N}_D^{[2]})^H]\\ &\quad= \mathbb{E}[((\boldsymbol{\Sigma}_R^{\frac{1}{2}}\boldsymbol{F}_R\boldsymbol{N}_R)^T\otimes\boldsymbol{\Sigma}_D^{\frac{1}{2}})\mathrm{vec}(\bar{\boldsymbol{H}}_{RD})\mathrm{vec}(\bar{\boldsymbol{H}}_{RD})^H\\ &\qquad\times((\boldsymbol{\Sigma}_R^{\frac{1}{2}}\boldsymbol{F}_R\boldsymbol{N}_R)^*\otimes\boldsymbol{\Sigma}_D^{\frac{1}{2}})]+\mathbb{E}[\mathrm{vec}(\boldsymbol{N}_D^{[2]})\mathrm{vec}(\boldsymbol{N}_D^{[2]})^H]\\ &\quad= \sigma_h^2\mathbb{E}\big[(\bar{\boldsymbol{N}}_R^T\boldsymbol{\Psi}_R^{\frac{1}{2}T}\boldsymbol{F}_R^T\boldsymbol{\Sigma}_R^T\boldsymbol{F}_R^*\boldsymbol{\Psi}_R^{\frac{1}{2}T}\bar{\boldsymbol{N}}_R^*)\otimes\boldsymbol{\Sigma}_D\big]\\ &\qquad+\mathbb{E}\big[\mathrm{vec}(\boldsymbol{N}_D^{[2]})\mathrm{vec}(\boldsymbol{N}_D^{[2]})^H\big]\\ &\quad= \sigma_n^2\boldsymbol{I}_{K+L}\otimes(\bar{\sigma}_r^2\boldsymbol{\Sigma}_D+\boldsymbol{\Psi}_D)\\ &\quad= \sigma_n^2\boldsymbol{I}_{K+L}\otimes(\bar{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D+\bar{\sigma}_r^2\boldsymbol{E}_D+\widehat{\boldsymbol{\Psi}}_D+\boldsymbol{E}_{ND}).\end{aligned} \tag{56}$$

where $\bar{\sigma}_r^2 = \sigma_h^2\mathrm{Tr}((\widehat{\boldsymbol{\Psi}}_R^T+\boldsymbol{E}_{NR}^T)\boldsymbol{F}_R^T(\widehat{\boldsymbol{\Sigma}}_R^T+\boldsymbol{E}_R^T)\boldsymbol{F}_R^*)$. Naturally, we can also introduce a novel statistical error matrix $\widetilde{\boldsymbol{E}}_{ND}\in\mathbb{C}^{N_D\times N_D}$ to replace the coupled error terms $\{\boldsymbol{E}_{NR},\boldsymbol{E}_R,\boldsymbol{E}_D\}$ in function $(\bar{\sigma}_r^2\boldsymbol{\Sigma}_D+\boldsymbol{\Psi}_D)$. As a result, the formulation (56) can be rewritten as

$$\begin{aligned}&\mathbb{E}[\mathrm{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{N}_R+\boldsymbol{N}_D^{[2]})\mathrm{vec}(\boldsymbol{H}_{RD}\boldsymbol{F}_R\boldsymbol{N}_R+\boldsymbol{N}_D^{[2]})^H]\\ &\qquad= \sigma_n^2\boldsymbol{I}_{K+L}\otimes(\hat{\sigma}_r^2\widehat{\boldsymbol{\Sigma}}_D+\widehat{\boldsymbol{\Psi}}_D+\widetilde{\boldsymbol{E}}_{ND}),\end{aligned} \tag{57}$$

where $\hat{\sigma}_r^2 = \sigma_h^2\mathrm{Tr}(\widehat{\boldsymbol{\Psi}}_R^T\boldsymbol{F}_R^T\widehat{\boldsymbol{\Sigma}}_R^T\boldsymbol{F}_R^*)$. Further based on (55) and (57), the imperfect noise covariance matrix $\widetilde{\boldsymbol{R}}_n$ in (8) is finally derived as the formulation (53), as shown at the top of this page.

## APPENDIX B

*lemma 1:* [56] For a diagonal matrix $\boldsymbol{L}\in\mathbb{R}^{N\times N}$ with $\pm1$ being diagonal elements, then we have $\boldsymbol{L}^2=\boldsymbol{I}$ and $\boldsymbol{L}\boldsymbol{\Lambda}\boldsymbol{L}=\boldsymbol{\Lambda}$ for the arbitrary diagonal matrix $\boldsymbol{\Lambda}$. Based on this, the diagonal matrix $\boldsymbol{\Lambda}_X$ which shares the same diagonal elements with those of the matrix $\boldsymbol{X}$ can be expressed as $\boldsymbol{\Lambda}_X=\frac{1}{T}\sum_{\boldsymbol{L}\in\mathcal{L}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{L}$. Here, $\mathcal{L}$ represents the set of all $T=2^N$ possible forms of $\boldsymbol{L}$.

Recalling the problem (30), we define $\widetilde{\boldsymbol{R}}_S=\widehat{\boldsymbol{U}}_S^H\widetilde{\boldsymbol{R}}_{\widetilde{\boldsymbol{C}}}\widehat{\boldsymbol{U}}_S$ and then rexpress the objective function of (30) as

$$\begin{aligned}f(\widetilde{\boldsymbol{R}}_S)=\max_{\boldsymbol{E}'\in\xi}\ &\mathrm{Tr}\Big\{\Big(\sigma_h^{-2}(\boldsymbol{\Lambda}_S+\boldsymbol{E}_S')^{-1}\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}_D')^{-1}\\ &+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\widetilde{\boldsymbol{R}}_S\otimes(\boldsymbol{\Omega}_D+\boldsymbol{E}_{ND}')^{-1}\Big)^{-1}\Big\}\\ &+\mathrm{Tr}\Big\{\Big((\hat{\sigma}_h^2\boldsymbol{\Lambda}_S+\widetilde{\boldsymbol{E}}_S')^{-1}\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}_D')^{-1}\\ &+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\widetilde{\boldsymbol{R}}_S\otimes(\boldsymbol{\Omega}_D+\hat{\sigma}_r^2\boldsymbol{\Lambda}_D+\widetilde{\boldsymbol{E}}_{ND}')^{-1}\Big)^{-1}\Big\}\end{aligned} \tag{58}$$

Then based on **lemma 1**, we have the formulation (61), as shown at the top of the next page, where

$$\overrightarrow{\boldsymbol{E}}' = \{\overrightarrow{\boldsymbol{E}}'_S, \overrightarrow{\widetilde{\boldsymbol{E}}}'_S, \boldsymbol{E}_D', \boldsymbol{E}_{ND}', \widetilde{\boldsymbol{E}}_{ND}'\} \tag{59}$$

and $\boldsymbol{L}_{N_S}$ denotes the $N_S$-dimensional matrix $\boldsymbol{L}$. Besides, the last equality results from the fact that for the arbitrary unitarily-invariant statistical error $\boldsymbol{E}\in\xi$, $\boldsymbol{L}\boldsymbol{E}\boldsymbol{L}\in\xi$ also holds due to the unitary matrix $\boldsymbol{L}$. Further considering the convexity of $f(\widetilde{\boldsymbol{R}}_S)$ w.r.t $\widetilde{\boldsymbol{R}}_S$ because of the function $\mathrm{Tr}(\boldsymbol{A})^{-1}$ and utilizing the Jensen inequality, we have

$$\begin{aligned}f(\boldsymbol{\Lambda}_C) &= f\Big(\frac{1}{T}\sum_{\boldsymbol{L}_{N_S}\in\mathcal{L}}\boldsymbol{L}_{N_S}\widetilde{\boldsymbol{R}}_S\boldsymbol{L}_{N_S}\Big)\\ &\le \frac{1}{T}\sum_{\boldsymbol{L}_{N_S}\in\mathcal{L}} f(\boldsymbol{L}_{N_S}\widetilde{\boldsymbol{R}}_S\boldsymbol{L}_{N_S}) = f(\widetilde{\boldsymbol{R}}_S).\end{aligned} \tag{60}$$

where the diagonal matrix $\boldsymbol{\Lambda}_C$ has the same diagonal elements as that of $\widetilde{\boldsymbol{R}}_S = \widehat{\boldsymbol{U}}_S^H\widetilde{\boldsymbol{R}}_{\widetilde{\boldsymbol{C}}}\widehat{\boldsymbol{U}}_S$. The formulation (60) indicates that for any given $\widetilde{\boldsymbol{R}}_S$, its diagonal structured counterpart $\boldsymbol{\Lambda}_C$ achieves the smaller value of function $f(\widetilde{\boldsymbol{R}}_S)$, that is $f(\boldsymbol{\Lambda}_C)\le f(\widetilde{\boldsymbol{R}}_S)$. Besides, for the problem (30), we also have power constraints $\mathrm{Tr}(\widetilde{\boldsymbol{R}}_{\widetilde{\boldsymbol{C}}}) = \mathrm{Tr}(\widetilde{\boldsymbol{R}}_S) = \mathrm{Tr}(\boldsymbol{\Lambda}_C) \le P_S'$, which indicates that the optimal $\widetilde{\boldsymbol{R}}_S$ of problem (30) should be diagonal. Furthermore, recalling that $\boldsymbol{C} = \widetilde{\boldsymbol{C}}_1(\boldsymbol{Q}_H^{[1:N_S]})^H$ and $\boldsymbol{R}_{\widetilde{\boldsymbol{C}}} = \widetilde{\boldsymbol{C}}_1^*\widetilde{\boldsymbol{C}}_1^T$ as well as $\widetilde{\boldsymbol{R}}_{\widetilde{\boldsymbol{C}}} = \lambda_{\max}(\boldsymbol{W}^*)\boldsymbol{R}_{\widetilde{\boldsymbol{C}}}$ and $\widetilde{\boldsymbol{R}}_S = \boldsymbol{U}_S^H\widetilde{\boldsymbol{R}}_{\widetilde{\boldsymbol{C}}}\boldsymbol{U}_S$, the superimposed training matrix $\boldsymbol{C}$ is expressed as $\boldsymbol{C} = \lambda_{\max}(\boldsymbol{W}^*)^{-\frac{1}{2}}\widehat{\boldsymbol{U}}_S^*\boldsymbol{\Lambda}_C^{\frac{1}{2}}\boldsymbol{Q}_A(\boldsymbol{Q}_H^{[1:N_S]})^T$ in (33), where $\boldsymbol{Q}_A$ satisfying $\boldsymbol{Q}_A\boldsymbol{Q}_A^H = \boldsymbol{I}_{N_S}$ is an arbitrary orthogonal unitary matrix. As for the derivation of the worst case channel and noise statistical errors $\overrightarrow{\boldsymbol{E}}' = \{\overrightarrow{\boldsymbol{E}}'_S, \overrightarrow{\widetilde{\boldsymbol{E}}}'_S, \boldsymbol{E}_D', \boldsymbol{E}_{ND}', \widetilde{\boldsymbol{E}}_{ND}'\}$, we take $\{\boldsymbol{E}_{ND}', \widetilde{\boldsymbol{E}}_{ND}'\}$ as an example and substitute the diagonal

$$f(\boldsymbol{L}_{N_S}\widetilde{\boldsymbol{R}}_S\boldsymbol{L}_{N_S}) = \max_{\overrightarrow{E}'\in\xi} \mathrm{Tr}\Big\{\big(\sigma_h^{-2}(\boldsymbol{\Lambda}_S+\underbrace{\boldsymbol{L}_{N_S}\boldsymbol{E}'_S\boldsymbol{L}_{N_S}}_{\overrightarrow{E}'_S\in\xi})^{-1}\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}'_D)^{-1}+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\widetilde{\boldsymbol{R}}_S\otimes(\boldsymbol{\Omega}_D+\boldsymbol{E}'_{ND})^{-1}\big)^{-1}\Big\}$$
$$+\mathrm{Tr}\Big\{\big((\hat{\sigma}_h^2\boldsymbol{\Lambda}_S+\underbrace{\boldsymbol{L}_{N_S}\widetilde{\boldsymbol{E}}'_S\boldsymbol{L}_{N_S}}_{\overrightarrow{E}'_S\in\xi})^{-1}\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}'_D)^{-1}+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\widetilde{\boldsymbol{R}}_S\otimes(\boldsymbol{\Omega}_D+\hat{\sigma}_r^2\boldsymbol{\Lambda}_D+\widetilde{\boldsymbol{E}}'_{ND})^{-1}\big)^{-1}\Big\}=f(\widetilde{\boldsymbol{R}}_S) \tag{61}$$

$$f(\boldsymbol{L}_{N_D}\boldsymbol{E}'_{ND}\boldsymbol{L}_{N_D},\boldsymbol{L}_{N_D}\widetilde{\boldsymbol{E}}'_{ND}\boldsymbol{L}_{N_D})$$
$$=\max_{\widehat{E}'\in\xi}\mathrm{Tr}\Big\{\big(\sigma_h^{-2}(\boldsymbol{\Lambda}_S+\boldsymbol{E}'_S)^{-1}\otimes(\boldsymbol{\Lambda}_D+\underbrace{\boldsymbol{L}_{N_D}\boldsymbol{E}'_D\boldsymbol{L}_{N_D}}_{\overrightarrow{E}'_D\in\xi})^{-1}+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\boldsymbol{\Lambda}_C\otimes(\boldsymbol{\Omega}_D+\boldsymbol{E}'_{ND})^{-1}\big)^{-1}\Big\}$$
$$+\mathrm{Tr}\Big\{\big((\hat{\sigma}_h^2\boldsymbol{\Lambda}_S+\widetilde{\boldsymbol{E}}'_S)^{-1}\otimes(\boldsymbol{\Lambda}_D+\underbrace{\boldsymbol{L}_{N_D}\boldsymbol{E}'_D\boldsymbol{L}_{N_D}}_{\overrightarrow{E}'_D\in\xi})^{-1}+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\boldsymbol{\Lambda}_C\otimes(\boldsymbol{\Omega}_D+\hat{\sigma}_r^2\boldsymbol{\Lambda}_D+\widetilde{\boldsymbol{E}}'_{ND})^{-1}\big)^{-1}\Big\}=f(\boldsymbol{E}'_{ND},\widetilde{\boldsymbol{E}}'_{ND}) \tag{62}$$

$\boldsymbol{\Lambda}_C$ into the function $f(\widetilde{\boldsymbol{R}}_S)$ to redefine it as

$$\begin{aligned} &f(\boldsymbol{E}'_{ND},\widetilde{\boldsymbol{E}}'_{ND})\\ &=\max_{\left\{\substack{\boldsymbol{E}'_S,\\ \widetilde{\boldsymbol{E}}'_S,\boldsymbol{E}'_D}\right\}\in\xi}\mathrm{Tr}\Big\{\Big(\sigma_h^{-2}(\boldsymbol{\Lambda}_S+\boldsymbol{E}'_S)^{-1}\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}'_D)^{-1}\\ &\quad+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\boldsymbol{\Lambda}_C\otimes(\boldsymbol{\Omega}_D+\boldsymbol{E}'_{ND})^{-1}\Big)^{-1}\Big\}\\ &\quad+\mathrm{Tr}\Big\{\Big((\hat{\sigma}_h^2\boldsymbol{\Lambda}_S+\widetilde{\boldsymbol{E}}'_S)^{-1}\otimes(\boldsymbol{\Lambda}_D+\boldsymbol{E}'_D)^{-1}\\ &\quad+\sigma_n^{-2}\lambda_{\max}(\boldsymbol{W}^*)^{-1}\boldsymbol{\Lambda}_C\otimes(\boldsymbol{\Omega}_D+\hat{\sigma}_r^2\boldsymbol{\Lambda}_D+\widetilde{\boldsymbol{E}}'_{ND})^{-1}\Big)^{-1}\Big\}. \end{aligned} \tag{63}$$

Similarly to deriving $\boldsymbol{\Lambda}_C$, we firstly give the formulation (62), as shown at the top of this page, where the set of $\widehat{\boldsymbol{E}}' = \{\boldsymbol{E}'_S,\widetilde{\boldsymbol{E}}'_S,\ \overrightarrow{\boldsymbol{E}}'_D\}$ is still unitarily-invariant and $\boldsymbol{L}_{N_D}$ denotes the $N_D$-dimensional matrix $\boldsymbol{L}$. Due to the joint concavity of $f(\boldsymbol{L}_{N_D}\boldsymbol{E}'_{ND}\boldsymbol{L}_{N_D},\ \boldsymbol{L}_{N_D}\widetilde{\boldsymbol{E}}'_{ND}\boldsymbol{L}_{N_D})$ as indicated in Appendix A, we have

$$\begin{aligned} &f(\boldsymbol{\Lambda}_{\boldsymbol{E}'_{ND}},\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{E}}'_{ND}})\\ &=f\Big(\frac{1}{T}\sum_{\boldsymbol{L}_{N_D}\in\mathcal{L}}\boldsymbol{L}_{N_D}\boldsymbol{E}'_{ND}\boldsymbol{L}_{N_D},\ \frac{1}{T}\sum_{\boldsymbol{L}_{N_D}\in\mathcal{L}}\boldsymbol{L}_{N_D}\widetilde{\boldsymbol{E}}'_{ND}\boldsymbol{L}_{N_D}\Big)\\ &\geq\frac{1}{T}\sum_{\boldsymbol{L}_{N_D}\in\mathcal{L}}f(\boldsymbol{L}_{N_D}\boldsymbol{E}'_{ND}\boldsymbol{L}_{N_D},\boldsymbol{L}_{N_D}\widetilde{\boldsymbol{E}}'_{ND}\boldsymbol{L}_{N_D})\\ &=f(\boldsymbol{E}'_{ND},\widetilde{\boldsymbol{E}}'_{ND}). \end{aligned} \tag{64}$$

Therefore, for any given feasible $\{\boldsymbol{E}'_{ND},\widetilde{\boldsymbol{E}}'_{ND}\}$, it can be concluded that the diagonal matrices $\{\boldsymbol{\Lambda}_{\boldsymbol{E}'_{ND}},\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{E}}'_{ND}}\}$ achieve the maximum value of $f(\boldsymbol{E}'_{ND},\widetilde{\boldsymbol{E}}'_{ND})$. In order to compute the diagonal $\{\boldsymbol{E}'_{ND},\widetilde{\boldsymbol{E}}'_{ND}\}$, the noise uncertainties $\{\boldsymbol{E}_{ND},\widetilde{\boldsymbol{E}}_{ND}\}$ should satisfy $\boldsymbol{E}_{ND}=\widehat{\boldsymbol{U}}_D\boldsymbol{\Lambda}_{\boldsymbol{E}'_{ND}}\widehat{\boldsymbol{U}}_D^H$ and $\widetilde{\boldsymbol{E}}_{ND}=\widehat{\boldsymbol{U}}_D\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{E}}'_{ND}}\widehat{\boldsymbol{U}}_D^H$ by referring to $\boldsymbol{E}'_{ND}=\widehat{\boldsymbol{U}}_D^H\boldsymbol{E}_{ND}\widehat{\boldsymbol{U}}_D$ and $\widetilde{\boldsymbol{E}}'_{ND}=\widehat{\boldsymbol{U}}_D^H\widetilde{\boldsymbol{E}}_{ND}\widehat{\boldsymbol{U}}_D$. It is worth highlighting that the detailed proof of the diagonal $\{\boldsymbol{E}'_S,\widetilde{\boldsymbol{E}}'_S,\boldsymbol{E}'_D\}$ is almost the same as that of $\{\boldsymbol{E}'_{ND},\widetilde{\boldsymbol{E}}'_{ND}\}$. Thus the original $\{\boldsymbol{E}_S,\widetilde{\boldsymbol{E}}_S,\boldsymbol{E}_D\}$ can be derived as that in (33). Due to space limitation, it is omitted this paper.

## REFERENCES

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, no. 3, pp. 311–335, Mar. 1998.

[2] 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.

[3] W. Xu, X. Dong, and W.-S. Lu, "MIMO relaying broadcast channels with linear precoding and quantized channel state information feedback," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5233–5245, Oct. 2010.

[4] C. Xing, Y. Ma, Y. Zhou, and F. Gao, "Transceiver optimization for multi-hop communications with per-antenna power constraints," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1519–1534, Mar. 2016.

[5] W. Xu and X. Dong, "Optimized one-way relaying strategy with outdated CSI quantization for spatial multiplexing," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4458–4464, Aug. 2012.

[6] J. Zhang and M. Haardt, "Energy efficient two-way non-regenerative relaying for relays with multiple antennas," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1079–1083, Aug. 2015.

[7] C. Xing, F. Gao, and Y. Zhou, "A framework for transceiver designs for multi-hop communications with covariance shaping constraints," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3930–3945, Feb. 2015.

[8] L. Liang, W. Xu, and X. Dong, "Limited Feedback-Based Multi-Antenna Relay Broadcast Channels with Block Diagonalization," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4092–4101, Aug. 2013.

[9] Z. Peng, W. Xu, L.-C. Wang, and C. Zhao, "Achievable rate analysis and feedback design for multiuser MIMO relay with imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 780–793, Feb. 2014.

[10] C. Xing, S. Ma, and Y.-C. Wu, "Robust joint design of linear relay precoder and destination equalizer for dual-hop amplify-and-forward MIMO relay systems," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2273–2283, Apr. 2010.

[11] S. Gong, C. Xing, S. Chen, and Z. Fei, "Polarization sensitive array based physical-layer security," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 3964–3981, May 2018.

[12] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[13] S. Gong, S. Ma, C. Xing, and G. Yang, "Optimal beamforming and time allocation for partially wireless powered sensor networks with downlink SWIPT," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3197–3212, Jun. 2019.

[14] S. Sugiura, S. Chen, and L. Hanzo, "MIMO-aided near-capacity turbo transceivers: Taxonomy and performance versus complexity," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 2, pp. 421–442, May 2012.

[15] S. Gong, C. Xing, S. Chen, and Z. Fei, "Secure communications for dual-polarized MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4177–4192, Aug. 2017.

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

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

[18] S. Gong, S. Wang, S. Chen, C. Xing, and X. Wei, "Time-invariant joint transmit and receive beampattern optimization for polarization-subarray based frequency diverse array radar," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5364–5379, Oct. 2018.

[19] S. Sugiura, S. Chen, H. Haas, P. M. Grant, and L. Hanzo, "Coherent versus non-coherent decode-and-forward relaying aided cooperative space-time shift keying," *IEEE Trans. Commun.*, vol. 59, no. 6, pp. 1707–1719, Jun. 2011.

[20] X. Lin, M. Tao, and Y. Xu, "MIMO two-way compress-and-forward relaying with approximate joint Eigen-decomposition," *IEEE Commun. Lett.*, vol. 17, no. 9, pp. 1750–1753, Sep. 2013.

[21] Y. Song and N. Devroye, "Lattice codes for the Gaussian relay channel: Decode-and-forward and compress-and-forward," *IEEE Trans. Inf. Theory*, vol. 59, no. 8, pp. 4927–4948, Aug. 2013.

[22] H.-B. Kong, H. M. Shin, T. Oh, and I. Lee, "Joint MMSE transceiver designs for MIMO AF relaying systems with direct link," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3547–3560, Jun. 2017.

[23] C. Xing, S. Ma, Z. Fei, Y.-C. Wu, and H. V. Poor, "A general robust linear transceiver design for multi-hop amplify-and-forward MIMO relaying systems," *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1196–1209, Mar. 2013.

[24] H. Shen, W. Xu, J. Wang, and C. Zhao, "A worst-case robust beamforming design for multi-antenna AF relaying," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 713–716, Apr. 2013.

[25] H. Shen, W. Xu, and C. Zhao, "Transceiver optimization for full-duplex massive MIMO AF relaying with direct link," *IEEE Access*, vol. 4, pp. 8857–8864, 2016.

[26] X. Yang, X. Liang, T. Song, W. Xu, and S. Jin, "Multiuser massive MIMO AF relaying: Spectral efficiency and power allocation," *IEEE Access*, vol. 6, pp. 18894–18906, 2018.

[27] Y. Dai and X. Dong, "Power allocation for multi-pair massive MIMO two-way AF relaying with linear processing," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 5932–5946, Sep. 2016.

[28] Y. Cai, Q. Shi, B. Champagne, and G. Y. Li, "Joint transceiver design for secure downlink communications over an amplify-and-forward MIMO relay," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3691–3704, Sep. 2017.

[29] 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.

[30] 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.

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

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

[33] J. Pang, G. Shen, D. Wang, L. Jiang, and W. Wang, "Channel estimation and optimal training design for amplify and forward MIMO relay channel under spatial fading correlation," in *Proc. IEEE 72nd Vehiculr Technol. Conf.* , Ottawa, ON, Canada, Sep. 2010, pp. 1–5.

[34] N. N. Tran, H. H. Nguyen, H. D. Tuan, and D. E. Dodds, "Training designs for amplify-and-forward relaying with spatially correlated antennas," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2864–2870, Jul. 2012.

[35] 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, Jun. 2010.

[36] G. Wang, F. Gao, W. Chen, and C. Tellambura, "Channel estimation and training design for two-way relay networks in time-selective fading environments," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2681–2691, Aug. 2011.

[37] B. Tang, J. Tang, and Y. Peng, "MIMO radar waveform design in colored noise based on information theory," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4684–4697, Sep. 2010.

[38] H. D. Tuan, H. H. Kha, and H. H. Nguyen, "Mutual information-based optimal training design for spatially correlated MIMO-OFDM," in *Proc. ISPACS*, Kanazawa, Japan, Jan. 2009, pp. 13–16.

[39] Y. Yang and R. S. Blum, "MIMO radar waveform design based on mutual information and minimum mean-square error estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 1, pp. 330–343, Jan. 2007.

[40] H. D. Tuan, H. K. Ha, H. H. Nguyen, and V.-J. Luong, "Optimized training sequences for spatially correlated MIMO-OFDM," *IEEE Trans. Wireless Commun.*, vol. 9, no. 9, pp. 2768–2778, Sep. 2010.

[41] J.-M. Kang and H.-M. Kim, "Both minimum MSE and maximum SNR channel training designs for MIMO AF multi-relay networks with spatially correlated fading," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 1888–1892, Nov. 2015.

[42] J. Ma, P. Orlik, J. Zhang, and G. Y. Li, "Pilot matrix design for estimating cascaded channels in two-hop MIMO amplify-and-forward relay systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1956–1965, Jun. 2011.

[43] Z. Xu, G. Y. Li, C. Yang, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient power allocation for pilots in training-based downlink OFDMA systems," *IEEE Trans. Commun.*, vol. 60, no. 10, pp. 3047–3058, Oct. 2012.

[44] S. Noh, M. D. Zoltowski, and D. J. Love, "Training sequence design for feedback assisted hybrid beamforming in massive MIMO systems," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 187–200, Jan. 2016.

[45] C. Xing, Z. Fei, Y. Zhou, Z. Pan, and H. Wang, "Transceiver designs with matrix-version water-filling architecture under mixed power constraints," *Sci. China Inf. Sci.*, vol. 59, no. 10, Oct. 2016, Art. no. 102312.

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

[47] S. Gong, C. Xing, Z. Fei, and S. Ma, "Millimeter-wave secrecy beamforming designs for two-way amplify-and-forward MIMO relaying networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2059–2071, Mar. 2017.

[48] 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.

[49] C.-T. Chiang and C. C. Fung, "Robust training sequence design for spatially correlated MIMO channel estimation," *IEEE Trans. Veh. Technol.*, vol. 60, no. 7, pp. 2882–2894, Jan. 2011.

[50] N. Shariati and M. Bengtsson, "Robust training sequence design for spatially correlated MIMO channels and arbitrary colored disturbance," in *Proc. IEEE 22nd Int. Symp. Pers., Indoor Mobile Radio Commun.*, Toronto, ON, Canada, Sep. 2011, pp. 1939–1943.

[51] B. Rong, S. Gong, and Z. Fei, "Robust superimposed training designs for MIMO AF relaying channels under total power constraint," in *Proc. IEEE/CIC Int. Conf. Commun. China*, Beijing, China, Aug. 2018, pp. 651–655.

[52] V. Nguyen, H. D. Tuan, H. H. Nguyen, and N. N. Tran, "Optimal superimposed training design for spatially correlated fading MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 3206–3217, Aug. 2008.

[53] A. G. Orozco-Lugo, M. M. Lara, and D. C. McLernon, "Channel estimation using implicit training," *IEEE Trans. Signal Process.*, vol. 52, no. 1, pp. 240–254, Jan. 2004.

[54] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory v.1*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[55] J.-P. Kermoal, L. Schumacher, K. I. Pedersen, P. E. Mogensen, and F. Frederiksen, "A stochastic MIMO radio channel model with experimental validation," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 6, pp. 1211–1226, Aug. 2002.

[56] Y. C. Eldar and N. Merhav, "A competitive minimax approach to robust estimation of random parameters," *IEEE Trans. Signal Process.*, vol. SP-52, no. 7, pp. 1931–1946, Jul. 2004.

[57] J. Wang and D. P. Palomar, "Worst-case robust MIMO transmission with imperfect channel knowledge," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3086–3100, Aug. 2009.
[58] S. Wang, S. Ma, C. Xing, S. Gong, J. An, and H. V. Poor, "Optimal training design for MIMO systems with general power constraints," *IEEE Trans. Signal Process.*, vol. 66, no. 14, pp. 3649–3664, Jul. 2018.
[59] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, vol. 2. SIAM, 2001.
[60] M. C. Grant and S. P. Boyd, *CVX Users' Guide*, document release 2.1, CVX Res., 2015.
[61] E. Björnson and B. Ottersten, "A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1807–1820, Mar. 2010.

**BEINI RONG** received the B.S. degree in information engineering from the Beijing Institute of Technology, Beijing, China, in 2016, where she is currently pursuing the M.S. degree with the School of Electronic and Information Engineering. Her research interests include signal processing and wireless communications.

**ZHONGSHAN ZHANG** received the B.E. and M.S. degrees in computer science and the Ph.D. degree in electrical engineering from the Beijing University of Posts and Telecommunications, in 1998, 2001, and 2004, respectively. In 2004, he joined the DoCoMo Beijing Laboratories as an Associate Researcher, and he was promoted to a Researcher, in 2005. In 2006, he joined the University of Alberta, Edmonton, AB, Canada, as a Postdoctoral Fellow. In 2009, he joined the Department of Research and Innovation, Alcatel-Lucent, Shanghai, as a Research Scientist. From 2010 to 2011, he was with NEC China Laboratories, as a Senior Researcher. He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology. His main research interests include statistical signal processing, self-organized networking, cognitive radio, and cooperative communications. He served or is serving as a Guest Editor and/or an Editor for several technical journals such as the *IEEE Communications Magazine* and the *KSII Transactions on Internet and Information Systems*.

**XIN ZHAO** received the B.S. degree in electronic engineering from the Beijing Institute of Technology, Beijing, China, in 2016, where he is currently pursuing the M.S. degree with the School of Electronic and Information Engineering. His research interests include signal processing, hybrid beamforming, nonlinear precoding, and convex optimization.

**XIAOYUN YU** received the B.S. degree in electronic engineering from Zhengzhou University, Zhengzhou, China, in 2016. She is currently pursuing the M.S. degree in information and communication engineering with the School of Information and Electronics, Beijing Institute of Technology, China. Her research interests include machine learning, convex optimization, and wireless communications.