Training Sequence Design for MIMO Channels: An Application-Oriented Approach

Dimitrios Katselis,[†⋆] Cristian R. Rojas,[†] Mats Bengtsson, Emil Björnson, Xavier Bombois, Nafiseh Shariati, Magnus Jansson and Håkan Hjalmarsson

***Abstract*— In this paper, the problem of training optimization for estimating a multiple-input multiple-output (MIMO) flat fading channel in the presence of spatially and temporally correlated Gaussian noise is studied in an application-oriented setup. So far, the problem of MIMO channel estimation has mostly been treated within the context of minimizing the mean square error (MSE) of the channel estimate subject to various constraints, such as an upper bound on the available training energy. We introduce a more general framework for the task of training sequence design in MIMO systems, which can treat not only the minimization of channel estimator's MSE, but also the optimization of a final performance metric of interest related to the use of the channel estimate in the communication system. First, we show that the proposed framework can be used to minimize the training energy budget subject to a quality constraint on the MSE of the channel estimator. A deterministic version of the "dual" problem is also provided. We then focus on four specific applications, where the training sequence can be optimized with respect to the classical channel estimation MSE, a weighted channel estimation MSE and the MSE of the equalization error due to the use of an equalizer at the receiver or an appropriate linear precoder at the transmitter. In this way, the intended use of the channel estimate is explicitly accounted for. The superiority of the proposed designs over existing methods is demonstrated via numerical simulations.**

***Index Terms*— Channel equalization, L-optimality criterion, MIMO channels, system identification, training sequence design.**

I. Introduction

An important factor in the performance of multiple antenna systems is the accuracy of the channel state information (CSI) [1]. CSI is primarily used at the receiver side for purposes of coherent or semi-coherent detection, but it can be also used at the transmitter side, e.g., for precoding and adaptive modulation. Since in communication systems the maximization of spectral efficiency is an objective of interest, the training duration and energy should be minimized. Most current systems use training signals that are white, both spatially and temporally, which is known to be a good choice according to several criteria [2], [3]. However, in case that some prior knowledge of the channel or noise statistics is available, it is possible to tailor the training signal and to obtain a significantly improved performance. Especially, several authors have studied scenarios where long-term CSI in the form of a covariance matrix over the short-term fading is available. So far, most proposed algorithms have been designed to minimize the squared error of the channel estimate, e.g., [4]–[9]. Alternative design criteria are used in [5] and [10], where the channel entropy is minimized given the received training signal. In [11], the resulting capacity in the case of a single-input single-output (SISO) channel is considered, while [12] focuses on the pairwise error probability.

Herein, a generic context is described, drawing from similar techniques that have been recently proposed for training signal design in system identification [13]–[15]. This context aims at providing a unified theoretical framework, that can be used to treat the MIMO training optimization problem in various scenarios. Furthermore, it provides a different way of looking at the aforementioned problem, that could be adjusted to a wide variety of estimation-related problems in communication systems. First, we show how the problem of minimizing the training energy subject to a quality constraint can be solved, while a "dual" deterministic (average design) problem is considered[1]. In the sequel, we show that by a suitable definition of the performance measure the problem of optimizing the training for minimizing the channel MSE can be treated as a special case. We also consider a weighted version of the channel MSE, which relates to the well-known L-optimality criterion [17]. Moreover, we explicitly consider how the channel estimate will be used and attempt to optimize the end performance of the data transmission, which is not necessarily equivalent to minimizing the mean square error (MSE) of the channel estimate. Specifically, we study two uses of the channel estimate: channel equalization at the receiver using a minimum mean square error (MMSE) equalizer and channel inversion (zero-forcing precoding) at the transmitter, and derive the corresponding optimal training signals for each case. In the case of MMSE equalization, separate approximations are provided for the high and low SNR regimes. Finally, the resulting performance is illustrated based on numerical simulations. Compared to related results in the control literature, here we directly design a finite-length training signal and consider not only deterministic channel

[†] Equally contributing authors. The order of their names is alphabetical.

[⋆] Corresponding author. Email: dimitrik@kth.se.

D. Katselis, C. R. Rojas, M. Bengtsson, E. Björnson, N. Shariati, M. Jansson and H. Hjalmarsson are with ACCESS Linnaeus Center, School of Electrical Engineering, KTH Royal Institute of Technology, SE 100-44, Stockholm, Sweden. E-mail: dimitrik@kth.se, cristian.rojas@ee.kth.se, mats.bengtsson@ee.kth.se, emil.bjornson@ee.kth.se, nafiseh@kth.se, magnus.jansson@ee.kth.se, hjalmars@kth.se.

X. Bombois is with Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, The Netherlands. E-mail: X.J.A.Bombois@tudelft.nl.

This work was partially supported by the Swedish Research Council under contract 621-2010-5819.

[1]The word "dual" in this paper defers from the Lagrangian duality studied in the context of convex optimization theory. See [16] for more details on this type of duality.

parameters, but also a Bayesian channel estimation framework. A related pilot design strategy has been proposed in [18] for the problem of jointly estimating the frequency offset and the channel impulse response in single antenna transmissions.

Implementing an adaptive choice of pilot signals in a practical system would require a feedback signalling overhead, since both the transmitter and the receiver have to agree on the choice of the pilots. Just as previous studies in the area, the current paper is primarily intended to provide a theoretical benchmark on the resulting performance of such a scheme. Directly considering the end performance in the pilot design is a step into making the results more relevant. The data model used in [4]–[10] is based on a questionable assumption, namely that the channel is frequency flat, but that the noise is allowed to be frequency selective. Such an assumption might be relevant in systems that share spectrum with other radio interfaces using a narrower bandwidth and possibly in situations where channel coding introduces a temporal correlation in interfering signals. In order to focus on the main principles of our proposed strategy and to keep the mathematical derivations as simple as possible, the same model has been used in the current paper.

As a final comment, the novelty of this paper is on introducing the application-oriented framework as the appropriate context for training sequence design in communication systems. To this end, Hermitian form-like approximations of performance metrics are addressed here because they usually are good approximations of many performance metrics of interest, as well as, for simplicity purposes and comprehensiveness of presentation. To illustrate the framework, we have for simplicity chosen to study performance metrics related to the MSE of the information carrying signal after equalization. Directly designing for performance metrics like bit error rate (BER) would be even more relevant but would involve more technical complications. Also, the BER is with good approximation monotonically increasing in the MSE of the input to the detector and we illustrate numerically that our design outperforms previous state-of-the-art also in terms of BER.

This paper is organized as follows: Section II introduces the basic MIMO received signal model and specific assumptions on the structure of channel and noise covariance matrices. Section III presents the optimal channel estimators, when the channel is considered to be either a deterministic or a random matrix. Section IV presents the application-oriented optimal training designs in a guaranteed performance context, based on confidence ellipsoids and Markov bound relaxations. Moreover, Section V focuses on four specific applications, namely that of MSE channel estimation, channel estimation based on the L-optimality criterion and finally channel estimation for MMSE equalization and ZF precoding. Numerical simulations are provided in Section VI, while Section VII concludes this paper.

Notations: Boldface (lower case) is used for column vectors, $\mathbf{x}$, and (upper case) for matrices, $\mathbf{X}$. Moreover, $\mathbf{X}^T$, $\mathbf{X}^H$, $\mathbf{X}^*$ and $\mathbf{X}^\dagger$ denote the transpose, the conjugate transpose, the conjugate and the Moore-Penrose pseudoinverse of $\mathbf{X}$, respectively. The trace of $\mathbf{X}$ is denoted as $\mathrm{tr}(\mathbf{X})$ and $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semidefinite. $\mathrm{vec}(\mathbf{X})$ is the vector produced by stacking the columns of $\mathbf{X}$, and $(\mathbf{X})_{i,j}$ is the (i,j)-th element of $\mathbf{X}$. $[\mathbf{X}]_+$ means that all negative eigenvalues of $\mathbf{X}$ are replaced by zeros (i.e., $[\mathbf{X}]_+ \succeq \mathbf{0}$). $\mathcal{CN}(\bar{\mathbf{x}}, \mathbf{Q})$ stands for circularly symmetric complex Gaussian random vectors, where $\bar{\mathbf{x}}$ is the mean and $\mathbf{Q}$ the covariance matrix. Finally, $\alpha!$ denotes the factorial of the nonnegative integer α and $\mathrm{mod}(a,b)$ the modulo operation between the integers a, b.

II. System Model

We consider a MIMO communication system with n_T antennas at the transmitter and n_R antennas at the receiver. The received signal at time t is modelled as

$$\mathbf{y}(t) = \mathbf{H}\mathbf{x}(t) + \mathbf{n}(t)$$

where $\mathbf{x}(t) \in \mathbb{C}^{n_T}$ and $\mathbf{y}(t) \in \mathbb{C}^{n_R}$ are the baseband representations of the transmitted and received signals, respectively. The impact of background noise and interference from adjacent communication links is represented by the additive term $\mathbf{n}(t) \in \mathbb{C}^{n_R}$. We will further assume that $\mathbf{x}(t)$ and $\mathbf{n}(t)$ are independent (weakly) stationary signals. The channel response is modeled by $\mathbf{H} \in \mathbb{C}^{n_R \times n_T}$, which is assumed constant during the transmission of one block of data and independent between blocks; that is, we are assuming frequency flat block fading. Two different models of the channel will be considered:

i) A deterministic model.
ii) A stochastic Rayleigh fading model[2], i.e., $\mathrm{vec}(\mathbf{H}) \in \mathcal{CN}(\mathbf{0}, \mathbf{R})$, where, for mathematical tractability, we will assume that the known covariance matrix $\mathbf{R}$ possesses the Kronecker model used, e.g., in [7], [10]:

$$\mathbf{R} = \mathbf{R}_T^T \otimes \mathbf{R}_R \tag{1}$$

where $\mathbf{R}_T \in \mathbb{C}^{n_T \times n_T}$ and $\mathbf{R}_R \in \mathbb{C}^{n_R \times n_R}$ are the spatial covariance matrices at the transmitter and receiver side, respectively. This model has been experimentally verified in [19], [20] and further motivated in [21], [22].

We consider training signals of arbitrary length B, represented by $\mathbf{P} \in \mathbb{C}^{n_T \times B}$, whose columns are the transmitted signal vectors during training. Placing the received vectors in $\mathbf{Y} = \begin{bmatrix}\mathbf{y}(1) & \dots & \mathbf{y}(B)\end{bmatrix} \in \mathbb{C}^{n_R \times B}$, we have

$$\mathbf{Y} = \mathbf{H}\mathbf{P} + \mathbf{N},$$

where $\mathbf{N} = \begin{bmatrix}\mathbf{n}(1) & \dots & \mathbf{n}(B)\end{bmatrix} \in \mathbb{C}^{n_R \times B}$ is the combined noise and interference matrix.

Defining $\widetilde{\mathbf{P}} = \mathbf{P}^T \otimes \mathbf{I}$, we can then write

$$\mathrm{vec}(\mathbf{Y}) = \widetilde{\mathbf{P}}\,\mathrm{vec}(\mathbf{H}) + \mathrm{vec}(\mathbf{N}). \tag{2}$$

As, for example, in [7], [10], we assume that $\mathrm{vec}(\mathbf{N}) \in \mathcal{CN}(\mathbf{0}, \mathbf{S})$, where the covariance matrix $\mathbf{S}$ also possesses a Kronecker structure

$$\mathbf{S} = \mathbf{S}_Q^T \otimes \mathbf{S}_R. \tag{3}$$

[2]For simplicity, we have assumed a zero-mean channel, but it is straightforward to extend the results to Rician fading channels, similarly to [9].

Here, $\mathbf{S}_Q \in \mathbb{C}^{B\times B}$ represents the temporal covariance matrix[3] and $\mathbf{S}_R \in \mathbb{C}^{n_R\times n_R}$ represents the received spatial covariance matrix.

The channel and noise statistics will be assumed known to the receiver during estimation. Statistics can often be achieved by long-term estimation and tracking [23].

For the data transmission phase, we will assume that the transmit signal $\{\mathbf{x}(t)\}$ is a zero-mean, weakly stationary process, which is both temporally and spatially white, i.e., its spectrum is $\mathbf{\Phi}_x(\omega) = \lambda_x \mathbf{I}$.

III. Channel Matrix Estimation

A. Deterministic Channel Estimation

The minimum variance unbiased (MVU) channel estimator for the signal model (2), subject to a deterministic channel (Assumption i) in Section II), is given by [24]

$$\operatorname{vec}(\widehat{\mathbf{H}}_{\text{MVU}}) = (\widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \widetilde{\mathbf{P}})^{-1} \widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \operatorname{vec}(\mathbf{Y}). \tag{4}$$

This estimate has the distribution

$$\operatorname{vec}(\widehat{\mathbf{H}}_{\text{MVU}}) \in \mathcal{CN}(\operatorname{vec}(\mathbf{H}), \boldsymbol{\mathcal{I}}^{-1}_{\text{F,MVU}}), \tag{5}$$

where $\boldsymbol{\mathcal{I}}_{\text{F,MVU}}$ is the inverse covariance matrix

$$\boldsymbol{\mathcal{I}}_{\text{F,MVU}} = \widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \widetilde{\mathbf{P}}. \tag{6}$$

From this, it follows that the estimation error $\widetilde{\mathbf{H}} \triangleq \widehat{\mathbf{H}}_{\text{MVU}} - \mathbf{H}$ will, with probability α, belong to the uncertainty set

$$\mathcal{D}_D = \left\{ \widetilde{\mathbf{H}} \ : \ \operatorname{vec}^H(\widetilde{\mathbf{H}}) \boldsymbol{\mathcal{I}}_{\text{F,MVU}} \operatorname{vec}(\widetilde{\mathbf{H}}) \leq \frac{1}{2}\chi^2_\alpha(2n_T n_R) \right\}, \tag{7}$$

where $\chi^2_\alpha(n)$ is the α percentile of the $\chi^2(n)$ distribution [15].

B. Bayesian Channel Estimation

For the case of a stochastic channel model (Assumption ii) in Section II), the posterior channel distribution becomes (see [24])

$$\operatorname{vec}(\mathbf{H}) | \mathbf{Y}, \mathbf{P} \in \mathcal{CN}(\operatorname{vec}(\widehat{\mathbf{H}}_{\text{MMSE}}), \mathbf{C}_{\text{MMSE}}), \tag{8}$$

where the first and second moments are

$$\begin{aligned} \operatorname{vec}(\widehat{\mathbf{H}}_{\text{MMSE}}) &= (\mathbf{R}^{-1} + \widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \widetilde{\mathbf{P}})^{-1} \widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \operatorname{vec}(\mathbf{Y}), \\ \mathbf{C}_{\text{MMSE}} &= (\mathbf{R}^{-1} + \widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \widetilde{\mathbf{P}})^{-1}. \end{aligned} \tag{9}$$

Thus, the estimation error $\widetilde{\mathbf{H}} \triangleq \widehat{\mathbf{H}}_{\text{MMSE}} - \mathbf{H}$ will, with probability α, belong to the uncertainty set

$$\mathcal{D}_B = \left\{ \widetilde{\mathbf{H}} \ : \ \operatorname{vec}^H(\widetilde{\mathbf{H}}) \boldsymbol{\mathcal{I}}_{\text{F,MMSE}} \operatorname{vec}(\widetilde{\mathbf{H}}) \leq \frac{1}{2}\chi^2_\alpha(2n_T n_R) \right\}, \tag{10}$$

where $\boldsymbol{\mathcal{I}}_{\text{F,MMSE}} \triangleq \mathbf{C}^{-1}_{\text{MMSE}}$ is the inverse covariance matrix in the MMSE case [15].

[3]We set the subscript Q to $\mathbf{S}_Q$ to highlight its temporal nature and the fact that its size is $B \times B$. The matrices with subscript T in this paper share the common characteristic that they are $n_T \times n_T$, while those with subscript R are $n_R \times n_R$.

IV. Application-Oriented Optimal Training Design

In a communication system, an estimate of the channel, say $\widehat{\mathbf{H}}$, is needed at the receiver to detect the data symbols and may also be used at the transmitter to improve the performance. Let $J(\widetilde{\mathbf{H}}, \mathbf{H})$ be a scalar measure of the performance degradation at the receiver due to the estimation error $\widetilde{\mathbf{H}}$ for a channel $\mathbf{H}$. The objective of the training signal design is then to ensure that the resulting channel estimation error $\widetilde{\mathbf{H}}$ is such that

$$J(\widetilde{\mathbf{H}}, \mathbf{H}) \leq \frac{1}{\gamma} \tag{11}$$

for some parameter $\gamma > 0$, which we call *accuracy*. In our settings, (11) can not be typically ensured, since the channel estimation error is Gaussian distributed (see (5) and (8)) and, therefore, can be arbitrarily large. However, for the MVU estimator (4), we know that, with probability α, $\widetilde{\mathbf{H}}$ will belong to the set $\mathcal{D}_D$ defined in (7). Thus, we are led to training signal designs which guarantee (11) for all channel estimation errors $\widetilde{\mathbf{H}} \in \mathcal{D}_D$. One training design problem that is based on this concept is to minimize the required transmit energy budget subject to this constraint

$$\begin{aligned} \text{DGPP}: \quad & \underset{\mathbf{P}\in\mathbb{C}^{n_T\times B}}{\text{minimize}} \ \operatorname{tr}(\mathbf{P}\mathbf{P}^H) \\ & \text{s.t.} \ \ J(\widetilde{\mathbf{H}}, \mathbf{H}) \leq \tfrac{1}{\gamma} \quad \forall \widetilde{\mathbf{H}} \in \mathcal{D}_D. \end{aligned} \tag{12}$$

Similarly, for the MMSE estimator in Subsection III-B, the corresponding optimization problem is given as follows

$$\begin{aligned} \text{SGPP}: \quad & \underset{\mathbf{P}\in\mathbb{C}^{n_T\times B}}{\text{minimize}} \ \operatorname{tr}(\mathbf{P}\mathbf{P}^H) \\ & \text{s.t.} \ \ J(\widetilde{\mathbf{H}}, \mathbf{H}) \leq \tfrac{1}{\gamma} \quad \forall \widetilde{\mathbf{H}} \in \mathcal{D}_B, \end{aligned} \tag{13}$$

where $\mathcal{D}_B$ is defined in (10). We will call (12) and (13), the deterministic guaranteed performance problem (DGPP) and the stochastic guaranteed performance problem (SGPP), respectively. An alternative, "dual", problem is to maximize the accuracy γ subject to a constraint $\mathcal{P} > 0$ on the transmit energy budget. For the MVU estimator this can be written as

$$\begin{aligned} \text{DMPP}: \quad & \underset{\mathbf{P}\in\mathbb{C}^{n_T\times B}}{\text{maximize}} \ \gamma \\ & \text{s.t.} \ \ J(\widetilde{\mathbf{H}}, \mathbf{H}) \leq \tfrac{1}{\gamma} \quad \forall \widetilde{\mathbf{H}} \in \mathcal{D}_D, \\ & \qquad \operatorname{tr}(\mathbf{P}\mathbf{P}^H) \leq \mathcal{P}. \end{aligned} \tag{14}$$

We will call this problem the deterministic maximized performance problem (DMPP). The corresponding Bayesian problem will be denoted as the stochastic maximized performance problem (SMPP). We will study the DGPP/SGPP in detail in this contribution, but the DMPP/SMPP can be treated in similar ways. In fact, Theorem 3 in [16] suggests that the solutions to the DMPP/SMPP are the same as for DGPP/SGPP, save for a scaling factor.

The existing work on optimal training design for MIMO channels are, to the best of the authors knowledge, based upon standard measures on the quality of the channel estimate, rather than on the quality of the end-use of the channel. The framework presented in this section can be used to treat the existing results as special cases. Additionally, if an end performance metric is optimized, the DGPP/SGPP and DMPP/SMPP formulations better reflect the ultimate objective of the training

design. This type of optimal training design formulations has already been used in the control literature, but mainly for large sample sizes [13], [14], [25], [26], yielding an enhanced performance with respect to conventional estimation-theoretic approaches. A reasonable question is to examine if such a performance gain can be achieved in the case of training sequence design for MIMO channel estimation, where the sample sizes would be very small.

Remark: Ensuring (11) can be translated into a chance constraint of the form

$$\Pr\left\{ J(\widetilde{\mathbf{H}}, \mathbf{H}) \leq \frac{1}{\gamma} \right\} \geq 1 - \varepsilon \tag{15}$$

for some $\varepsilon \in [0, 1]$. Problems (12), (13) and (14) correspond to a convex relaxation of this chance constraint based on confidence ellipsoids [27], as we show in the next subsection.

A. Approximating the Training Design Problems

A key issue regarding the above training signal design problems is their computational tractability. In general, they are highly non-linear and non-convex. However, for performance metrics that are sufficiently smooth functions of the estimation error and have a minimum when the estimation error is zero, Taylor's theorem shows that they can be well approximated by a constant plus a quadratic term in $\widetilde{\mathbf{H}}$. Therefore, we consider performance metrics that can be approximated by

$$J(\widetilde{\mathbf{H}}, \mathbf{H}) \approx \mathrm{vec}^H(\widetilde{\mathbf{H}}) \boldsymbol{\mathcal{I}}_{\text{adm}} \mathrm{vec}(\widetilde{\mathbf{H}}). \tag{16}$$

For mathematical tractability, we will further assume that the Hermitian positive definite matrix $\boldsymbol{\mathcal{I}}_{\text{adm}}$ can be written in Kronecker product form as $\boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R$ for some matrices $\boldsymbol{\mathcal{I}}_T$ and $\boldsymbol{\mathcal{I}}_R$. In Section V, we will show several examples of practically relevant performance metrics that can be approximated in this form. This means that we can approximate the set $\{\widetilde{\mathbf{H}} : J(\widetilde{\mathbf{H}}, \mathbf{H}) \leq 1/\gamma\}$ of all admissible estimation errors $\widetilde{\mathbf{H}}$ by a (complex) ellipsoid in the parameter space

$$\mathcal{D}_{adm} = \{\widetilde{\mathbf{H}} \ : \ \mathrm{vec}^H(\widetilde{\mathbf{H}}) \gamma \boldsymbol{\mathcal{I}}_{\text{adm}} \mathrm{vec}(\widetilde{\mathbf{H}}) \leq 1\}. \tag{17}$$

Consequently, the DGPP (12) can be approximated by

$$\text{ADGPP}: \ \begin{array}{ll} \underset{\mathbf{P} \in \mathbb{C}^{n_T \times B}}{\text{minimize}} & \mathrm{tr}(\mathbf{P}\mathbf{P}^H) \\ \text{s.t.} & \mathcal{D}_D \subseteq \mathcal{D}_{adm}. \end{array} \tag{18}$$

We call this problem the approximative DGPP (ADGPP). Both $\mathcal{D}_D$ and $\mathcal{D}_{adm}$ are level sets of quadratic functions of the channel estimation error. Rewriting (7) so that we have the same level as in (17), we obtain

$$\mathcal{D}_D = \left\{ \widetilde{\mathbf{H}} \ : \ \mathrm{vec}^H(\widetilde{\mathbf{H}}) \frac{2\boldsymbol{\mathcal{I}}_{\text{F,MVU}}}{\chi^2_\alpha(2n_T n_R)} \mathrm{vec}(\widetilde{\mathbf{H}}) \leq 1 \right\}.$$

Comparing this expression with (17) gives that $\mathcal{D}_D \subseteq \mathcal{D}_{adm}$ if and only if

$$\frac{2\boldsymbol{\mathcal{I}}_{\text{F,MVU}}}{\chi^2_\alpha(2n_T n_R)} \succeq \gamma \boldsymbol{\mathcal{I}}_{\text{adm}}$$

(for a more general result see [15, Theorem 3.1]).

When $\boldsymbol{\mathcal{I}}_{\text{adm}}$ has the form $\boldsymbol{\mathcal{I}}_{\text{adm}} = \boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R$, with $\boldsymbol{\mathcal{I}}_T \in \mathbb{C}^{n_T \times n_T}$ and $\boldsymbol{\mathcal{I}}_R \in \mathbb{C}^{n_R \times n_R}$, the ADGPP (18) can then be written as

$$\begin{array}{ll} \underset{\mathbf{P} \in \mathbb{C}^{n_T \times B}}{\text{minimize}} & \mathrm{tr}(\mathbf{P}\mathbf{P}^H) \\ \text{s.t.} & \underbrace{\widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \widetilde{\mathbf{P}}}_{\boldsymbol{\mathcal{I}}_{\text{F,MVU}}} \succeq \frac{\gamma \chi^2_\alpha(2n_T n_R)}{2} \boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R. \end{array} \tag{19}$$

Similarly, by observing that $\mathcal{D}_{adm}$ only depends on the channel estimation error, and following the derivations above, the SGPP can be approximated by the following formulation

$$\begin{array}{ll} \underset{\mathbf{P} \in \mathbb{C}^{n_T \times B}}{\text{minimize}} & \mathrm{tr}(\mathbf{P}\mathbf{P}^H) \\ \text{s.t.} & \underbrace{\mathbf{R}^{-1} + \widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \widetilde{\mathbf{P}}}_{\boldsymbol{\mathcal{I}}_{\text{F,MMSE}}} \succeq \frac{\gamma \chi^2_\alpha(2n_T n_R)}{2} \boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R. \end{array} \tag{20}$$

We call the last problem approximative SGPP (ASGPP).

Remarks:

1) Several examples of the approximation (16) are presented in Section V. The approximation (16) is not possible for the performance metric of every application. Therefore, in some applications, alternative convex approximations of the corresponding performance metrics may have to be found.
2) The quality of the approximation (16) is characterized by its corresponding tightness to the true performance metric. For our purposes, when the tightness of the aforementioned approximation is acceptable, such an approximation will be desirable because it corresponds to a Hermitian form, therefore offering nice mathematical properties and tractability.
3) The sizes of $\mathcal{D}_D$ and $\mathcal{D}_{adm}$ critically depend on the parameter α. In practice, requiring α to have a value close to 1 corresponds to adequately representing the uncertainty set in which (approximately) all possible channel estimation errors lie.

B. The Deterministic Guaranteed Performance Problem

The problem formulations for ADGPP and ASGPP in (19) and (20), respectively, are similar in structure. The solutions to these problems (and to other approximative guaranteed performance problems) can be obtained from the following general theorem.

Theorem 1: Consider the optimization problem

$$\begin{array}{ll} \underset{\mathbf{P} \in \mathbb{C}^{n \times N}}{\text{minimize}} & \mathrm{tr}(\mathbf{P}\mathbf{P}^H) \\ \text{s.t.} & \mathbf{P}\mathbf{A}^{-1}\mathbf{P}^H \succeq \mathbf{B} \end{array} \tag{21}$$

where $\mathbf{A} \in \mathbb{C}^{N \times N}$ is Hermitian positive definite, $\mathbf{B} \in \mathbb{C}^{n \times n}$ is Hermitian positive semi-definite, and $N \geq \mathrm{rank}(\mathbf{B})$. An optimal solution to (21) is

$$\mathbf{P}^{\text{opt}} = \mathbf{U}_B \mathbf{D}_P \mathbf{U}_A^H \tag{22}$$

where $\mathbf{D}_P \in \mathbb{C}^{n \times N}$ is a rectangular diagonal matrix with $\sqrt{(\mathbf{D}_A)_{1,1}(\mathbf{D}_B)_{1,1}}, \ldots, \sqrt{(\mathbf{D}_A)_{m,m}(\mathbf{D}_B)_{m,m}}$ on the main diagonal. Here, $m = \min(n, N)$, while $\mathbf{U}_A$ and $\mathbf{U}_B$ are

unitary matrices that originate from the eigendecompositions of $\mathbf{A}$ and $\mathbf{B}$, respectively, i.e.,

$$\begin{aligned}\mathbf{A} &= \mathbf{U}_A\mathbf{D}_A\mathbf{U}_A^H \\ \mathbf{B} &= \mathbf{U}_B\mathbf{D}_B\mathbf{U}_B^H\end{aligned} \tag{23}$$

and $\mathbf{D}_A, \mathbf{D}_B$ are real-valued diagonal matrices, with their diagonal elements sorted in ascending and descending order, respectively; that is, $0 < (\mathbf{D}_A)_{1,1} \leq \ldots \leq (\mathbf{D}_A)_{N,N}$ and $(\mathbf{D}_B)_{1,1} \geq \ldots \geq (\mathbf{D}_B)_{n,n} \geq 0$.

If the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$ are distinct and strictly positive, then the solution (22) is unique up to the multiplication of the columns of $\mathbf{U}_A$ and $\mathbf{U}_B$ by complex unit-norm scalars.

Proof: The proof is given in Appendix II. ∎

By the right choice of $\mathbf{A}$ and $\mathbf{B}$, Theorem 1 will solve the ADGPP in (19). This is shown by the next theorem (recall that we have assumed that $\mathbf{S} = \mathbf{S}_Q^T \otimes \mathbf{S}_R$).

Theorem 2: Consider the optimization problem

$$\begin{aligned}\underset{\mathbf{P}\in\mathbb{C}^{n_T\times B}}{\text{minimize}} \quad & \text{tr}(\mathbf{P}\mathbf{P}^H) \\ \text{s.t.} \quad & \widetilde{\mathbf{P}}^H(\mathbf{S}_Q^T \otimes \mathbf{S}_R)^{-1}\widetilde{\mathbf{P}} \succeq c\boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R\end{aligned} \tag{24}$$

where $\widetilde{\mathbf{P}} = \mathbf{P}^T \otimes \mathbf{I}$, $\mathbf{S}_Q \in \mathbb{C}^{B\times B}$, $\mathbf{S}_R \in \mathbb{C}^{n_R\times n_R}$ are Hermitian positive definite, and $\boldsymbol{\mathcal{I}}_T \in \mathbb{C}^{n_T\times n_T}$, $\boldsymbol{\mathcal{I}}_R \in \mathbb{C}^{n_R\times n_R}$ are Hermitian positive semi-definite, and c is a positive constant.

If $B \geq \text{rank}(\boldsymbol{\mathcal{I}}_T)$, this problem is equivalent to (21) in Theorem 1 for $\mathbf{A} = \mathbf{S}_Q$ and $\mathbf{B} = c\lambda_{\max}(\mathbf{S}_R\boldsymbol{\mathcal{I}}_R)\boldsymbol{\mathcal{I}}_T$, where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue.

Proof: The proof is given in Appendix III. ∎

C. The Stochastic Guaranteed Performance Problem

Next, we will see that Theorem 1 can be also used to solve the ASGPP in (20). In order to obtain closed-form solutions, we need some equality relation between the Kronecker blocks of $\mathbf{R} = \mathbf{R}_T^T \otimes \mathbf{R}_R$ and of either $\mathbf{S} = \mathbf{S}_Q^T \otimes \mathbf{S}_R$ or $\boldsymbol{\mathcal{I}}_{\text{adm}} = \boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R$. For instance, it can be $\mathbf{R}_R = \mathbf{S}_R$, which may be satisfied if the receive antennas are spatially uncorrelated or if the signal and interference are received from the same main direction. See [7] for details on the interpretations of these assumptions.

The solution to ASGPP in (20) is given by the next theorem.

Theorem 3: Consider the optimization problem

$$\begin{aligned}\underset{\mathbf{P}\in\mathbb{C}^{n_T\times B}}{\text{minimize}} \quad & \text{tr}(\mathbf{P}\mathbf{P}^H) \\ \text{s.t.} \quad & \mathbf{R}^{-1} + \widetilde{\mathbf{P}}^H\mathbf{S}^{-1}\widetilde{\mathbf{P}} \succeq c\boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R\end{aligned} \tag{25}$$

where $\widetilde{\mathbf{P}} = \mathbf{P}^T \otimes \mathbf{I}$, $\mathbf{R} = \mathbf{R}_T^T \otimes \mathbf{R}_R$, and $\mathbf{S} = \mathbf{S}_Q^T \otimes \mathbf{S}_R$. Here, $\mathbf{R}_T \in \mathbb{C}^{n_T\times n_T}$, $\mathbf{R}_R \in \mathbb{C}^{n_R\times n_R}$, $\mathbf{S}_Q \in \mathbb{C}^{B\times B}$, $\mathbf{S}_R \in \mathbb{C}^{n_R\times n_R}$ are Hermitian positive definite, and $\boldsymbol{\mathcal{I}}_T \in \mathbb{C}^{n_T\times n_T}$, $\boldsymbol{\mathcal{I}}_R \in \mathbb{C}^{n_R\times n_R}$ are Hermitian positive semi-definite, and c is a positive constant.

- If $\mathbf{R}_R = \mathbf{S}_R$ and $B \geq \text{rank}([c\lambda_{\max}(\mathbf{S}_R\boldsymbol{\mathcal{I}}_R)\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T^{-1}]_+)$, then the problem is equivalent to (21) in Theorem 1 for $\mathbf{A} = \mathbf{S}_Q$ and $\mathbf{B} = [c\lambda_{\max}(\mathbf{S}_R\boldsymbol{\mathcal{I}}_R)\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T^{-1}]_+$.
- If $\mathbf{R}_R^{-1} = \boldsymbol{\mathcal{I}}_R$ and $B \geq \text{rank}([c\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T^{-1}]_+)$, then the problem is equivalent to (21) in Theorem 1 for $\mathbf{A} = \mathbf{S}_Q$ and $\mathbf{B} = \lambda_{\max}(\mathbf{S}_R\boldsymbol{\mathcal{I}}_R)[c\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T^{-1}]_+$.
- If $\mathbf{R}_T^{-1} = \boldsymbol{\mathcal{I}}_T$ and $B \geq \text{rank}(\boldsymbol{\mathcal{I}}_T)$, then the problem is equivalent to (21) in Theorem 1 for $\mathbf{A} = \mathbf{S}_Q$ and $\mathbf{B} = \lambda_{\max}(\mathbf{S}_R[c\boldsymbol{\mathcal{I}}_R - \mathbf{R}_R]_+)\boldsymbol{\mathcal{I}}_T$.

Proof: The proof is given in Appendix III. ∎

The mathematical difference between ADGPP and ASGPP is the $\mathbf{R}^{-1}$ term that appears in the constraint of the latter. This term has a clear impact on the structure of the optimal ASGPP training matrix.

It is also worth noting that the solution for $\mathbf{R}_R = \mathbf{S}_R$ requires $B \geq \text{rank}([c\lambda_{\max}(\mathbf{S}_R\boldsymbol{\mathcal{I}}_R)\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T^{-1}]_+)$ which means that solutions can be achieved also for $B < n_T$ (i.e., when only the $B < n_T$ strongest eigendirections of the channel are excited by training). In certain cases, e.g., when the interference is temporally white ($\mathbf{S}_Q = \mathbf{I}$), it is optimal to have $B = \text{rank}([c\lambda_{\max}(\mathbf{S}_R\boldsymbol{\mathcal{I}}_R)\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T^{-1}]_+)$ as larger B will not decrease the training energy usage, cf. [9].

D. Optimizing the Average Performance

Except from the previously presented training designs, the application-oriented design can be alternatively given in the following deterministic "dual" context. If $\mathbf{H}$ is considered to be deterministic, then we can setup the following optimization problem

$$\begin{aligned}\underset{\mathbf{P}\in\mathbb{C}^{n_T\times B}}{\text{minimize}} \quad & \mathbb{E}_{\widetilde{\mathbf{H}}}\left\{J(\widetilde{\mathbf{H}}, \mathbf{H})\right\} \\ \text{s.t.} \quad & \text{tr}(\mathbf{P}\mathbf{P}^H) \leq \mathcal{P}.\end{aligned} \tag{26}$$

Clearly, for the MVU estimator

$$\mathbb{E}_{\widetilde{\mathbf{H}}}\left\{J(\widetilde{\mathbf{H}}, \mathbf{H})\right\} = \text{tr}\left\{\boldsymbol{\mathcal{I}}_{\text{adm}}(\widetilde{\mathbf{P}}^H\mathbf{S}^{-1}\widetilde{\mathbf{P}})^{-1}\right\},$$

so problem (26) is solved by the following theorem.

Theorem 4: Consider the optimization problem

$$\begin{aligned}\underset{\mathbf{P}\in\mathbb{C}^{n_T\times B}}{\text{minimize}} \quad & \text{tr}\left\{\boldsymbol{\mathcal{I}}_{\text{adm}}(\widetilde{\mathbf{P}}^H\mathbf{S}^{-1}\widetilde{\mathbf{P}})^{-1}\right\} \\ \text{s.t.} \quad & \text{tr}(\mathbf{P}\mathbf{P}^H) \leq \mathcal{P}\end{aligned} \tag{27}$$

where $\boldsymbol{\mathcal{I}}_{\text{adm}} = \boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R$ as before. Set $\boldsymbol{\mathcal{I}}_T' = \boldsymbol{\mathcal{I}}_T^T = \mathbf{U}_T\mathbf{D}_T\mathbf{U}_T^H$ and $\mathbf{S}_Q' = \mathbf{S}_Q^T = \mathbf{U}_Q\mathbf{D}_Q\mathbf{U}_Q^H$. Here, $\mathbf{U}_T \in \mathbb{C}^{n_T\times n_T}$, $\mathbf{U}_Q \in \mathbb{C}^{B\times B}$ are unitary matrices and $\mathbf{D}_T, \mathbf{D}_Q$ are diagonal $n_T \times n_T$ and $B \times B$ matrices containing the eigenvalues of $\boldsymbol{\mathcal{I}}_T'$ and $\mathbf{S}_Q'$ in descending and ascending order, respectively. Then, the optimal training matrix $\mathbf{P}$ equals $\left(\mathbf{U}_T\mathbf{D}_P\mathbf{U}_Q^H\right)^*$, where $\mathbf{D}_P$ is an $n_T \times B$ diagonal matrix with main diagonal entries equal to $(\mathbf{D}_P)_{i,i} = \sqrt{\mathcal{P}\sqrt{\alpha_i}/\sum_{j=1}^{n_T}\sqrt{\alpha_j}}, i = 1, 2, \ldots, n_T$ ($B \geq n_T$) and $\alpha_i = (\mathbf{D}_T)_{i,i}(\mathbf{D}_Q)_{i,i}, i = 1, 2, \ldots, n_T$ with the aforementioned ordering.

Proof: The proof is given in Appendix IV. ∎

Remarks:

1) In the general case of a non Kronecker-structured $\boldsymbol{\mathcal{I}}_{adm}$, the solution of the different designs, (19), (20) and (27) can be obtained using numerical methods like the semidefinite relaxation approach described in [28].
2) If $\boldsymbol{\mathcal{I}}_{\text{adm}}$ depends on $\mathbf{H}$, then in order to implement this design, the embedded $\mathbf{H}$ in $\boldsymbol{\mathcal{I}}_{\text{adm}}$ may be replaced by a previous channel estimate. This implies that this approach is possible whenever the channel variations

allow for such a design. This observation also applies to the designs in the previous subsections. See also [16], [29], where the same issue is discussed for other system identification applications.

The corresponding performance criterion for the case of the MMSE estimator is given by

$$\mathbb{E}_{\widetilde{\mathbf{H}},\mathbf{H}}\left\{J(\widetilde{\mathbf{H}},\mathbf{H})\right\} = \mathrm{tr}\left\{\boldsymbol{\mathcal{I}}_{\text{adm}}(\mathbf{R}^{-1}+\widetilde{\mathbf{P}}^H\mathbf{S}^{-1}\widetilde{\mathbf{P}})^{-1}\right\}.$$

In this case, we can derive closed form expressions for the optimal training under assumptions similar to those made in Theorem 3. We therefore have the following result:

Theorem 5: Consider the optimization problem

$$\begin{array}{ll}\underset{\mathbf{P}\in\mathbb{C}^{n_T\times B}}{\text{minimize}} & \mathrm{tr}\left\{\boldsymbol{\mathcal{I}}_{\text{adm}}(\mathbf{R}^{-1}+\widetilde{\mathbf{P}}^H\mathbf{S}^{-1}\widetilde{\mathbf{P}})^{-1}\right\} \\ \text{s.t.} & \mathrm{tr}(\mathbf{P}\mathbf{P}^H)\leq\mathcal{P}\end{array} \tag{28}$$

where $\boldsymbol{\mathcal{I}}_{\text{adm}} = \boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R$ as before. Set $\mathbf{S}'_Q = \mathbf{S}_Q^T = \mathbf{V}_Q\boldsymbol{\Lambda}_Q\mathbf{V}_Q^H$. Here, we assume that $\mathbf{V}_Q \in \mathbb{C}^{B\times B}$ is a unitary matrix and $\boldsymbol{\Lambda}_Q$ a diagonal $B\times B$ matrix containing the eigenvalues of $\mathbf{S}'_Q$ in arbitrary order. Assume also that $\mathbf{R}'_T = \mathbf{R}_T^T$ with eigenvalue decomposition $\mathbf{U}'_T\boldsymbol{\Lambda}'_T\mathbf{U}'^H_T$. The diagonal elements of $\boldsymbol{\Lambda}'_T$ are assumed to be arbitrarily ordered. Then, we have the following cases

- $\mathbf{R}_R = \mathbf{S}_R$: We further discriminate two cases
  - $\boldsymbol{\mathcal{I}}_T = \mathbf{I}$: Then the optimal training is given by a straightforward adaptation of Proposition 2 in [8].
  - $\mathbf{R}_T^{-1} = \boldsymbol{\mathcal{I}}_T$: Then, the optimal training matrix $\mathbf{P}$ equals $\left(\mathbf{U}'_T(\pi_{\text{opt}})\mathbf{D}_P\mathbf{V}_Q^H(\varpi_{\text{opt}})\right)^*$, where $\pi_{\text{opt}}, \varpi_{\text{opt}}$ stand for the optimal orderings of the eigenvalues of $\mathbf{R}'_T$ and $\mathbf{S}'_Q$, respectively. These optimal orderings are determined by Algorithm 1 in Appendix V. Additionally, define the parameter m_* as in eq. (69) (see Appendix V). Assuming in the following that, for simplicity of notation, $(\boldsymbol{\Lambda}'_T)_{i,i}$'s and $(\boldsymbol{\Lambda}_Q)_{i,i}$'s have the optimal ordering, the optimal $(\mathbf{D}_P)_{j,j}, j = 1,2,\ldots,m_*$ are given by the expression

$$\sqrt{\frac{\mathcal{P}+\sum_{i=1}^{m_*}\frac{(\boldsymbol{\Lambda}_Q)_{i,i}}{(\boldsymbol{\Lambda}'_T)_{i,i}}}{\sum_{i=1}^{m_*}\sqrt{\frac{(\boldsymbol{\Lambda}_Q)_{i,i}}{(\boldsymbol{\Lambda}'_T)_{i,i}}}}\sqrt{\frac{(\boldsymbol{\Lambda}_Q)_{j,j}}{(\boldsymbol{\Lambda}'_T)_{j,j}}}-\frac{(\boldsymbol{\Lambda}_Q)_{j,j}}{(\boldsymbol{\Lambda}'_T)_{j,j}}},$$

  while $(\mathbf{D}_P)_{j,j} = 0$ for $j = m_*+1,\ldots,n_T$.

Proof: The proof is given in Appendix V. ∎

Remarks: Two interesting additional cases complementing the last theorem are the following:

1) If the modal matrices of $\mathbf{R}_R$ and $\mathbf{S}_R$ are the same, $\boldsymbol{\mathcal{I}}_T = \mathbf{I}$ and $\boldsymbol{\mathcal{I}}_R = \mathbf{I}$, then the optimal training is given by [9].
2) In any other case (e.g., if $\mathbf{R}_R \neq \mathbf{S}_\mathbf{R}$), the training can be found using numerical methods like the semidefinite relaxation approach described in [28]. Note again that this approach can also handle general $\boldsymbol{\mathcal{I}}_{adm}$, not necessarily expressed as $\boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R$.

As a general conclusion, the objective function of the dual deterministic problems presented in this subsection can be shown to correspond to Markov bound approximations of the chance constraint (15). According to the analysis in [27], these approximations should be tighter than the approximations based on confidence ellipsoids presented in Subsections IV-A, IV-B and IV-C, for practically relevant values of ε.

V. Applications

A. Optimal Training for Channel Estimation

We now consider the channel estimation problem in its standard context, where the performance metric of interest is the (mean) square error of the corresponding channel estimator. Linear estimators for this task are given by (4), (9). The performance metric of interest is

$$J(\widetilde{\mathbf{H}},\mathbf{H}) = \mathrm{vec}^H(\widetilde{\mathbf{H}})\,\mathrm{vec}(\widetilde{\mathbf{H}}),$$

which corresponds to $\boldsymbol{\mathcal{I}}_{\text{adm}} = \mathbf{I}$, i.e., to $\boldsymbol{\mathcal{I}}_T = \mathbf{I}$ and $\boldsymbol{\mathcal{I}}_R = \mathbf{I}$. The ADGPP and ASGPP are given by (19) and (20), respectively, with the corresponding substitutions. Their solutions follow directly from Theorems 2 and 3, respectively. To the best of the authors' knowledge, such formulations for the classical MIMO training design problem are presented here for the first time. Furthermore, solutions to the standard approach of minimizing the channel MSE subject to a constraint on the training energy budget are provided by Theorems 4 and 5 as special cases.

Remark: Although the confidence ellipsoid and Markov bound approximations are generally different [27], in the simulation section we show that their performance is almost identical for reasonable operating γ-regimes in the specific case of standard channel estimation.

B. Optimal Training for the L-Optimality Criterion

Consider now a performance metric of the form

$$J_W(\widetilde{\mathbf{H}},\mathbf{H}) = \mathrm{vec}^H(\widetilde{\mathbf{H}})\mathbf{W}\,\mathrm{vec}(\widetilde{\mathbf{H}}),$$

for some positive semidefinite weighting matrix $\mathbf{W}$. Assume also that $\mathbf{W} = \mathbf{W}_1 \otimes \mathbf{W}_2$ for some positive semidefinite matrices $\mathbf{W}_1, \mathbf{W}_2$. Taking the expected value of this performance metric with respect to either $\widetilde{\mathbf{H}}$ or both $\widetilde{\mathbf{H}}$ and $\mathbf{H}$ leads to the well-known L-optimality criterion for optimal experiment design in statistics [17]. In this case, $\boldsymbol{\mathcal{I}}_T = \mathbf{W}_1^T$ and $\boldsymbol{\mathcal{I}}_R = \mathbf{W}_2$. In the context of MIMO communication systems, such a performance metric may arise, e.g., if we want to estimate the MIMO channel having some deficiencies in either the transmit and/or the receive antenna arrays. The simplest case would be both $\mathbf{W}_1$ and $\mathbf{W}_2$ being diagonal with nonzero entries in the interval $[0,1]$, $\mathbf{W}_1$ representing the deficiencies in the transmit antenna array and $\mathbf{W}_2$ in the receive array. More general matrices can be considered if we assume cross-couplings between the transmit and/or receive antenna elements.

Remark: The numerical approach of [28] mentioned after Theorems 4 and 5 can handle general weighting matrices $\mathbf{W}$, not necessarily Kronecker-structured.

C. Optimal Training for Channel Equalization

In this subsection we consider the problem of estimating a transmitted signal sequence $\{\mathbf{x}(t)\}$ from the corresponding received signal sequence $\{\mathbf{y}(t)\}$. Among a wide range of

methods that are available [30], [31], we will consider the MMSE equalizer and for mathematical tractability we will approximate it by the non-causal Wiener filter. Note that for reasonably long block lengths, the MMSE estimate becomes similar to the non-causal Wiener filter [32]. Thus, the optimal training design based on the non-causal Wiener filter should also provide good performance when using an MMSE equalizer.

1) Equalization using exact channel state information: Let us first assume that $\mathbf{H}$ is available. In this ideal case, and with the transmitted signal being weakly stationary with spectrum $\mathbf{\Phi}_x$, the MSE-optimal estimate of the transmitted signal $\mathbf{x}(t)$ from the received observations of $\mathbf{y}(t)$ can be obtained according to

$$\hat{\mathbf{x}}(t;\mathbf{H}) = \mathbf{F}(q;\mathbf{H})\mathbf{y}(t) \tag{29}$$

where q is the unit time shift operator, $[q\mathbf{x}(t) = \mathbf{x}(t+1)]$, and the non-causal Wiener filter $\mathbf{F}(e^{j\omega};\mathbf{H})$ is given by

$$\begin{aligned}\mathbf{F}(e^{j\omega};\mathbf{H}) &= \mathbf{\Phi}_{xy}(\omega)\mathbf{\Phi}_y^{-1}(\omega) \\ &= \mathbf{\Phi}_x(\omega)\mathbf{H}^H\left(\mathbf{H}\mathbf{\Phi}_x(\omega)\mathbf{H}^H + \mathbf{\Phi}_n(\omega)\right)^{-1}.\end{aligned} \tag{30}$$

Here, $\mathbf{\Phi}_{xy}(\omega) = \mathbf{\Phi}_x(\omega)\mathbf{H}^H$ denotes the cross-spectrum between $\mathbf{x}(t)$ and $\mathbf{y}(t)$, and

$$\mathbf{\Phi}_y(\omega) = \mathbf{H}\mathbf{\Phi}_x(\omega)\mathbf{H}^H + \mathbf{\Phi}_n(\omega) \tag{31}$$

is the spectral density of $\mathbf{y}(t)$. Using our assumption that $\mathbf{\Phi}_x(\omega) = \lambda_x\mathbf{I}$, we obtain the simplified expression

$$\mathbf{F}(e^{j\omega};\mathbf{H}) \;=\; \mathbf{H}^H\left(\mathbf{H}\mathbf{H}^H + \mathbf{\Phi}_n(\omega)/\lambda_x\right)^{-1}. \tag{32}$$

Remark: Assuming nonsingularity of $\mathbf{\Phi}_n(\omega)$ for every ω, the MMSE equalizer is applicable for all values of the pair (n_T, n_R).

2) Equalization using a channel estimate: Consider now the situation where the exact channel $\mathbf{H}$ is unavailable, but we only have an estimate $\widehat{\mathbf{H}}$. When we replace $\mathbf{H}$ by its estimate in the expressions above, the estimation error for the equalizer will increase. While the increase in the bit error rate would be a natural measure of the quality of the channel estimate $\widehat{\mathbf{H}}$, for simplicity we consider the total MSE of the difference, $\hat{\mathbf{x}}(t;\mathbf{H}+\widetilde{\mathbf{H}}) - \hat{\mathbf{x}}(t;\mathbf{H}) = \mathbf{\Delta}(q;\widetilde{\mathbf{H}},\mathbf{H})\mathbf{y}(t)$ (note that $\widehat{\mathbf{H}} = \mathbf{H} + \widetilde{\mathbf{H}}$), using the notation $\mathbf{\Delta}(q;\widetilde{\mathbf{H}},\mathbf{H}) \triangleq \mathbf{F}(q;\mathbf{H}+\widetilde{\mathbf{H}}) - \mathbf{F}(q;\mathbf{H})$. In view of this, we will use the channel equalization (CE) performance metric

$$\begin{aligned}J_{CE}(\widetilde{\mathbf{H}},\mathbf{H}) &= \mathbb{E}\left\{[\mathbf{\Delta}(q;\widetilde{\mathbf{H}},\mathbf{H})\mathbf{y}(t)]^H[\mathbf{\Delta}(q;\widetilde{\mathbf{H}},\mathbf{H})\mathbf{y}(t)]\right\} \\ &= \mathbb{E}\left\{\mathrm{tr}\left([\mathbf{\Delta}(q;\widetilde{\mathbf{H}},\mathbf{H})\mathbf{y}(t)][\mathbf{\Delta}(q;\widetilde{\mathbf{H}},\mathbf{H})\mathbf{y}(t)]^H\right)\right\} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{tr}\left(\mathbf{\Delta}(e^{j\omega};\widetilde{\mathbf{H}},\mathbf{H})\mathbf{\Phi}_y(\omega)\mathbf{\Delta}^H(e^{j\omega};\widetilde{\mathbf{H}},\mathbf{H})\right)\,d\omega.\end{aligned} \tag{33}$$

We see that the poorer the accuracy of the estimate, the larger the performance metric $J_{CE}(\widetilde{\mathbf{H}},\mathbf{H})$ and, thus, the larger the performance loss of the equalizer. Therefore, this performance metric is a reasonable candidate to use when formulating our training sequence design problem. Indeed, the Wiener equalizer based on the estimate $\widehat{\mathbf{H}} = \mathbf{H} + \widetilde{\mathbf{H}}$ of $\mathbf{H}$ can be deemed to have a satisfactory performance if $J_{CE}(\widetilde{\mathbf{H}},\mathbf{H})$ remains below some user-chosen threshold. Thus, we will use J_{CE} as J in problems (12) and (13). Though these problems are not convex, we show in Appendix I how they can be convexified, provided some approximations are made.

Remarks:

1) The excess MSE $J_{CE}(\widetilde{\mathbf{H}},\mathbf{H})$ quantifies the distance of the MMSE equalizer using the channel estimate $\widehat{\mathbf{H}}$ over the *clairvoyant* MMSE equalizer, i.e., the one using the true channel. This performance metric is not the same as the classical MSE in the equalization context, where the difference $\hat{\mathbf{x}}(t;\mathbf{H}+\widetilde{\mathbf{H}}) - \mathbf{x}(t)$ is considered instead of $\hat{\mathbf{x}}(t;\mathbf{H}+\widetilde{\mathbf{H}}) - \hat{\mathbf{x}}(t;\mathbf{H})$. However, since in practice the best transmit vector estimate that can be attained is the clairvoyant one, the choice of $J_{CE}(\widetilde{\mathbf{H}},\mathbf{H})$ is justified. This selection allows for a performance metric approximation given by (16).
2) There are certain cases of interest, where $J_{CE}(\widetilde{\mathbf{H}},\mathbf{H})$ approximately coincides with the classical equalization MSE. Such a case occurs when $n_R \geq n_T$, $\mathbf{H}$ is full column rank and the SNR is high during data transmission.

D. Optimal training for Zero-Forcing (ZF) Precoding

Apart from receiver side channel equalization, as another example of how to apply the channel estimate we consider point-to-point zero-forcing precoding, also known as channel inversion [33]. Here the channel estimate is fed back to the transmitter and its (pseudo-)inverse is used as a linear precoder. The data transmission is described by

$$\mathbf{y}(t) = \mathbf{H}\mathbf{\Psi}\mathbf{x}(t) + \mathbf{v}(t)$$

where the precoder is $\mathbf{\Psi} = \widehat{\mathbf{H}}^\dagger$, i.e., $\mathbf{\Psi} = \widehat{\mathbf{H}}^H(\widehat{\mathbf{H}}\widehat{\mathbf{H}}^H)^{-1}$ if we limit ourselves to the practically relevant case $n_T \geq n_R$ and assume that $\widehat{\mathbf{H}}$ is full rank. Note that $\mathbf{x}(t)$ is an $n_R \times 1$ vector in this case, but the transmitted vector is $\mathbf{\Psi}\mathbf{x}(t)$, which is $n_T \times 1$.

Under these assumptions, and following the same strategy and notation as in Appendix I, we get

$$\begin{aligned}\mathbf{y}(t;\widehat{\mathbf{H}}) - \mathbf{y}(t;\mathbf{H}) &= \mathbf{H}\widehat{\mathbf{H}}^\dagger\mathbf{x}(t) + \mathbf{v} - (\mathbf{H}\mathbf{H}^\dagger\mathbf{x}(t) + \mathbf{v}) \\ &= (\widehat{\mathbf{H}}\widehat{\mathbf{H}}^\dagger - \widetilde{\mathbf{H}}\widehat{\mathbf{H}}^\dagger - \mathbf{I})\mathbf{x}(t) \simeq -\widetilde{\mathbf{H}}\mathbf{H}^\dagger\mathbf{x}(t)\end{aligned} \tag{34}$$

Consequently, a quadratic approximation of the cost function is given by

$$\begin{aligned}J_{\mathrm{ZF}}(\widetilde{\mathbf{H}},\mathbf{H}) &= \mathbb{E}\left\{[\mathbf{y}(t;\widehat{\mathbf{H}}) - \mathbf{y}(t;\mathbf{H})]^H[\mathbf{y}(t;\widehat{\mathbf{H}}) - \mathbf{y}(t;\mathbf{H})]\right\} \\ &\simeq \lambda_x\,\mathrm{vec}^H(\widetilde{\mathbf{H}})\left((\mathbf{H}^\dagger(\mathbf{H}^\dagger)^H)^T \otimes \mathbf{I}\right)\mathrm{vec}(\widetilde{\mathbf{H}}) \\ &= \mathrm{vec}^H(\widetilde{\mathbf{H}})(\boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R)\,\mathrm{vec}(\widetilde{\mathbf{H}}),\end{aligned} \tag{35}$$

if we define $\boldsymbol{\mathcal{I}}_T = \lambda_x\mathbf{H}^\dagger(\mathbf{H}^\dagger)^H = \lambda_x\mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-2}\mathbf{H}$ and $\boldsymbol{\mathcal{I}}_R = \mathbf{I}$.

Remark: The cost functions of (27) and (28) reveal the fact that any performance-oriented training design is a compromise between the strict channel estimation accuracy and the desired accuracy related to the end performance metric at hand.

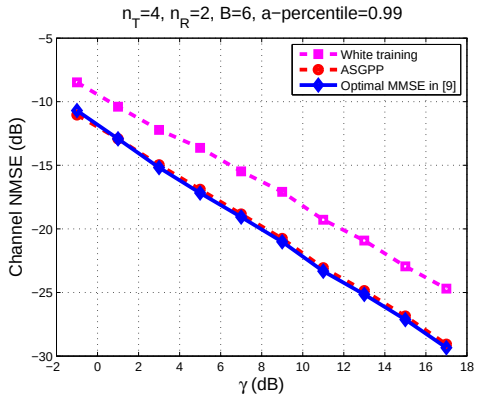

Fig. 1. $n_T = 4, n_R = 2, B = 6, a(\%) = 99$: Channel Estimation NMSE based on Subection V-A with $\mathbf{R}_R = \mathbf{S}_R$.

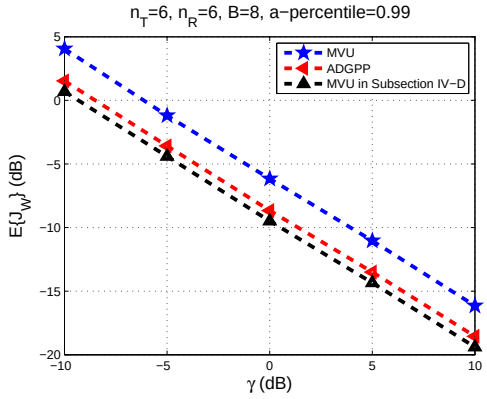

Fig. 2. $n_T = 6, n_R = 6, B = 8, a(\%) = 99$: L-optimality criterion with arbitrary but positive-semidefinite $\mathbf{W}_1, \mathbf{W}_2$ for the MVU estimator.

Caution is needed to identify cases where the performance-oriented design may severely degrade the channel estimation accuracy, annihilating all gains from such a design. In the case of ZF precoding, if $n_T > n_R$, $\boldsymbol{\mathcal{I}}_T$ will have rank at most n_R yielding a training matrix $\mathbf{P}$ with only n_R active eigendirections. This is in contrast to the secondary target, which is the channel estimation accuracy. Therefore, we expect ADGPP, ASGPP and the approaches in Subsection IV-D to behave abnormally in this case. Thus, we propose the performance-oriented design only when $n_T = n_R$ in the context of the ZF precoding.

VI. Numerical Examples

The purpose of this section is to examine the performance of optimal training sequence designs, and compare them with existing methods. For the channel estimation MSE figure, we plot the normalized MSE (NMSE), i.e., $\mathbb{E}(\|\mathbf{H} - \widehat{\mathbf{H}}\|^2/\|\mathbf{H}\|^2)$, versus the accuracy parameter γ. In all figures, fair comparison among the presented schemes is ensured via training energy equalization. Additionally, the matrices $\mathbf{R}_T, \mathbf{R}_R, \mathbf{S}_Q, \mathbf{S}_R$ follow the exponential model, that is, they are built according to

$$(\mathbf{R})_{i,j} = r^{j-i}, \quad j \geq i, \tag{36}$$

where r is the (complex) normalized correlation coefficient with magnitude $\rho = |r| < 1$. We choose to examine the high

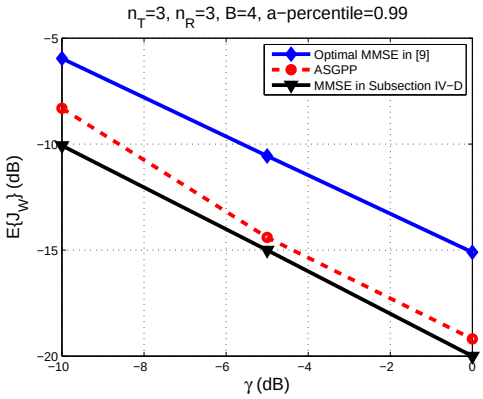

Fig. 3. $n_T = 3, n_R = 3, B = 4, a(\%) = 99$: L-optimality criterion with arbitrary but positive-semidefinite $\mathbf{W}_1, \mathbf{W}_2$ for the MMSE estimator with $\mathbf{R}_R = \mathbf{S}_R$.

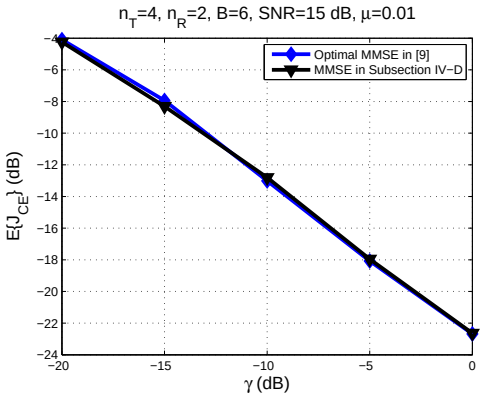

Fig. 4. $n_T = 4, n_R = 2, B = 6, \mathrm{SNR} = 15\mathrm{dB}, \mu = 0.01$: MMSE Channel Equalization with $\mathbf{R}_R \neq \mathbf{S}_\mathbf{R}$.

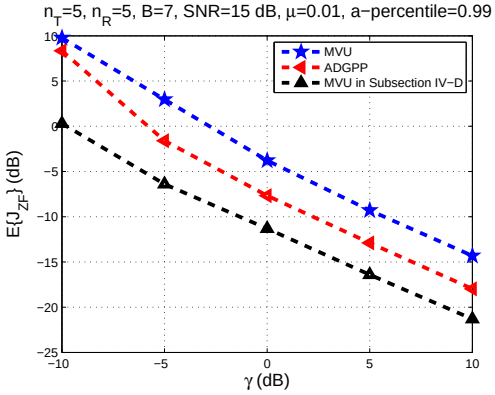

Fig. 5. $n_T = 5, n_R = 5, B = 7, \mathrm{SNR} = 15\mathrm{dB}, a(\%) = 99, \mu = 0.01$: ZF precoding based on Subection V-D for the MVU estimator. $\boldsymbol{\mathcal{I}}_{\text{adm}}$ is based on a previous channel estimate.

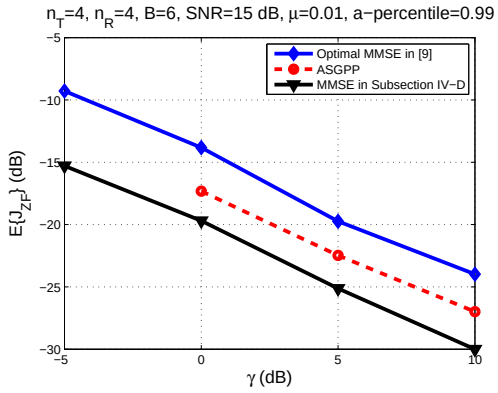

Fig. 6. $n_T = 4, n_R = 4, B = 6, \mathrm{SNR} = 15\mathrm{dB}, \mu = 0.01, a(\%) = 99$: ZF precoding MSE based on Subection V-D for the MMSE estimator with $\mathbf{R}_R = \mathbf{S}_R$. $\boldsymbol{\mathcal{I}}_{\text{adm}}$ is based on a previous channel estimate.

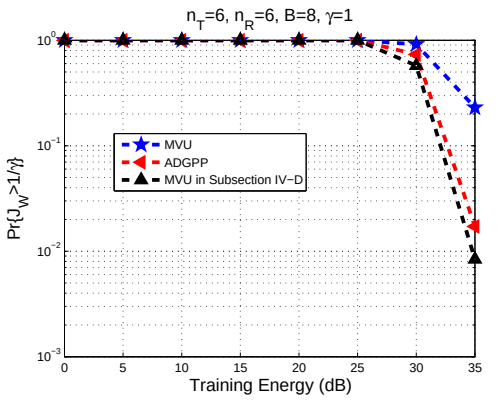

Fig. 7. $n_T = 6, n_R = 6, B = 8, \gamma = 1$: Outage probability for the L-optimality criterion with the MVU estimator. The accuracy parameter is $\gamma = 1$.

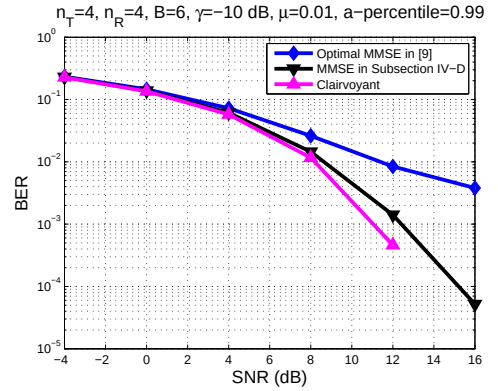

Fig. 8. $n_T = 4, n_R = 4, B = 6, \gamma = -10\mathrm{dB}, \mu = 0.01, a(\%) = 99$: BER performance using the signal estimates produced by the corresponding schemes in Fig. 6 with $\mathbf{R}_R = \mathbf{S}_R$ and $\gamma = -10$ dB. $\boldsymbol{\mathcal{I}}_{\text{adm}}$ is based on a previous channel estimate.

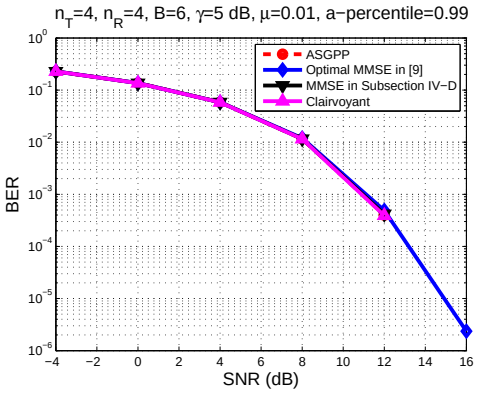

Fig. 9. $n_T = 4, n_R = 4, B = 6, \gamma = 5\mathrm{dB}, \mu = 0.01, a(\%) = 99$: BER performance using the signal estimates produced by the corresponding schemes in Fig. 6 with $\mathbf{R}_R = \mathbf{S}_R$ and $\gamma = 5$ dB. $\boldsymbol{\mathcal{I}}_{\text{adm}}$ is based on a previous channel estimate.

correlation scenario for all the presented schemes. Therefore, in all plots $|r| = 0.9$ for all matrices $\mathbf{R}_T, \mathbf{R}_R, \mathbf{S}_Q, \mathbf{S}_R$. Additionally, the transmit SNR during data transmission is chosen to be 15 dB, when channel equalization and ZF precoding are considered. High SNR expressions are therefore used for optimal training sequence designs. Since the optimal pilot sequences depend on the true channel, we have for these two applications additionally assumed that the channel changes from block to block according to the relationship $\mathbf{H}_i = \mathbf{H}_{i-1} + \mu\mathbf{E}_i$, where $\mathbf{E}_i$ has the same Kronecker structure as $\mathbf{H}$ and it is completely independent from $\mathbf{H}_{i-1}$. The estimated $\mathbf{H}_{i-1}$ is used in the pilot design. In Figs. 4, 5, 6, 8 and 9 the value of μ is 0.01.

In Fig. 1 the channel estimation NMSE performance versus the accuracy γ is presented for three different schemes. The scheme 'ASGPP' is the optimal Wiener filter together with the optimal guaranteed performance training matrix described in Subsection V-A. 'Optimal MMSE in [9]' is the scheme presented in [9], which solves the optimal training problem for the *vectorized* MMSE, operating on $\mathrm{vec}(\mathbf{Y})$. This solution is a special case in the statement of Theorem 5 for $\boldsymbol{\mathcal{I}}_{adm} = \mathbf{I}$, i.e., $\boldsymbol{\mathcal{I}}_T = \mathbf{I}$ and $\boldsymbol{\mathcal{I}}_R = \mathbf{I}$. Finally, the scheme 'White training' corresponds to the use of the vectorized MMSE filter at the receiver, with a white training matrix, i.e., one having equal singular values and arbitrary left and right singular matrices. This scheme is justified when the receiver knows the involved channel and noise statistics, but does not want to sacrifice bandwidth to feedback the optimal training matrix to the transmitter. This scheme is also justified in fast fading environments. In Fig. 1, we assume that $\mathbf{R}_R = \mathbf{S}_R$ and we implement the corresponding optimal training design for each scheme. 'ASGPP' is implemented first for a certain value of γ and the rest of the schemes are forced to have the same training energy. The 'Optimal MMSE in [9]' and 'ASGPP' schemes have the best and almost identical MSE performance. This indicates that for the problem of training design with the classical channel estimation MSE, the confidence ellipsoid relaxation of the chance constraint and the relaxation based on the Markov bound in Subsection IV-D deliver almost identical performances.

Figs. 2 and 3 demonstrate the L-optimality average performance metric $E\{J_W\}$ versus γ. Fig. 2 corresponds to the L-optimality criterion based on MVU estimators and Fig. 3 is based on MMSE estimators. In Fig. 2, the scheme 'MVU' corresponds to the optimal training for channel estimation when the MVU estimator is used. This training is given by Theorem 4 for $\boldsymbol{\mathcal{I}}_{adm} = \mathbf{I}$, i.e., $\boldsymbol{\mathcal{I}}_T = \mathbf{I}$ and $\boldsymbol{\mathcal{I}}_R = \mathbf{I}$. 'MVU in Subsection IV-D' is again the MVU estimator based on the same theorem but for the correct $\boldsymbol{\mathcal{I}}_{adm}$. The scheme 'MMSE in Subsection IV-D' is given by the numerical solution mentioned below Theorem 5, since $\mathbf{W}_1$ is different than the cases where a closed form solution is possible. Figs. 2 and 3 clearly show that both the confidence ellipsoid and Markov bound approximations are better than the optimal training for standard channel estimation. Therefore, for this problem the application-oriented training design is superior compared to training designs with respect to the quality of the channel estimate.

Fig. 4 demonstrates the performance of optimal training designs for the MMSE estimator in the context of MMSE channel equalization. We assume that $\mathbf{R}_R \neq \mathbf{S}_R$, since the high SNR expressions for $\boldsymbol{\mathcal{I}}_{adm}$ in the context of MMSE channel equalization in Appendix I indicate that $\boldsymbol{\mathcal{I}}_T = \mathbf{I}$ for this application and according to Theorem 5 the optimal training corresponds to the optimal training for channel estimation in [8]. We observe that the curves almost coincide. Moreover, it can be easily verified that for MMSE channel equalization with the MVU estimator, the optimal training designs given by Theorems 2 and 4 differ slightly only in the optimal power loading. These observations essentially show that the optimal training designs for the MVU and MMSE estimators in the classical channel estimation setup are nearly optimal for the application of MMSE channel equalization. This relies on the fact that for this particular application, $\boldsymbol{\mathcal{I}}_T = \mathbf{I}$ in the high data transmission SNR regime.

Figs. 5 and 6 present the corresponding performances in the case of the ZF precoding. The descriptions of the schemes are as before. In Fig. 6, we assume that $\mathbf{R}_R = \mathbf{S}_R$. The superiority of the application-oriented designs for the ZF precoding application is apparent in these plots. Here, $\boldsymbol{\mathcal{I}}_T \neq \mathbf{I}$ and this is why the optimal training for the channel estimate works less well in this application. Moreover, the "ASGPP" is plotted for $\gamma \geq 0$ dB, since for smaller values of γ all the eigenvalues of $\mathbf{B} = [c\lambda_{\max}(\mathbf{S}_R\boldsymbol{\mathcal{I}}_R)\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T^{-1}]_+$ are equal to zero for this particular set of parameters defining Fig. 6.

Fig. 7 presents an outage plot in the context of the L-optimality criterion for the MVU estimator. We assume that $\gamma = 1$. We plot $\Pr\{J_W > 1/\gamma\}$ versus the training power. This plot indirectly verifies that the confidence ellipsoid relaxation of the chance constraint given by the scheme "ASGPP" is not as tight as the Markov bound approximation given by the scheme "MVU in Subsection IV-D".

Finally, Figs. 8 and 9 present the BER performance of the nearest neighbor rule applied to the signal estimates produced by the corresponding schemes in Fig. 6, when the QPSK modulation is used. The "Clairvoyant" scheme corresponds to the ZF precoder with perfect channel knowledge. The channel estimates have been obtained for $\gamma = -10$ and 5 dB, respectively. Even if the application-oriented estimates are not optimized for the BER performance metric, they lead to better performance than the 'Optimal MMSE in [9]' scheme as is apparent in Fig. 8. In Fig. 9, the performances of all schemes approximately coincide. This is due to the fact that for $\gamma = 5$ dB all channel estimates are very good, thus leading to symbol MSE performance differences that have negligible impact on the BER performance.

VII. Conclusions

In this contribution, we have presented a quite general framework for MIMO training sequence design subject to flat and block fading, as well as spatially and temporally correlated Gaussian noise. The main contribution has been to incorporate the objective of the channel estimation into the design. We have shown that by a suitable approximation of $J(\widetilde{\mathbf{H}}, \mathbf{H})$, it is possible to solve this type of problem for several interesting applications such as standard MIMO channel estimation, L-optimality criterion, MMSE channel equalization and ZF precoding. For these problems, we have numerically demonstrated the superiority of the schemes derived in this paper. Additionally, the proposed framework is valuable since it provides a universal way of posing different estimation-related problems in communication systems. We have seen that it shows interesting promise for, e.g., ZF precoding and it may yield even greater end performance gains in estimation problems related to communication systems, when approximations can be avoided, depending on the end performance metric at hand.

Appendix I
Approximating the performance measure for MMSE Channel Equalization

In order to obtain the approximating set $\mathcal{D}_{adm}$, let us first denote the integrand in the performance metric (33) by

$$J'(\omega; \widetilde{\mathbf{H}}, \mathbf{H}) = \mathrm{tr}\left(\boldsymbol{\Delta}(e^{j\omega}; \widetilde{\mathbf{H}}, \mathbf{H})\boldsymbol{\Phi}_y(\omega)\boldsymbol{\Delta}^H(e^{j\omega}; \widetilde{\mathbf{H}}, \mathbf{H})\right). \tag{37}$$

In addition, let $\simeq$ denote an equality in which only dominating terms with respect to $||\widetilde{\mathbf{H}}||$ are retained. Then, using (32), we observe that

$$\begin{aligned}\boldsymbol{\Delta}(e^{j\omega}; \widetilde{\mathbf{H}}, \mathbf{H}) &= \mathbf{F}(e^{j\omega}; \mathbf{H} + \widetilde{\mathbf{H}}) - \mathbf{F}(e^{j\omega}; \mathbf{H}) \\ &\simeq \lambda_x \widetilde{\mathbf{H}}^H \boldsymbol{\Phi}_y^{-1} - \lambda_x^2 \mathbf{H}^H \boldsymbol{\Phi}_y^{-1}(\mathbf{H}\widetilde{\mathbf{H}}^H + \widetilde{\mathbf{H}}\mathbf{H}^H)\boldsymbol{\Phi}_y^{-1} \\ &= \lambda_x \Big(\underbrace{\left(\mathbf{I} - \lambda_x \mathbf{H}^H \boldsymbol{\Phi}_y^{-1}\mathbf{H}\right)}_{=\mathbf{Q}} \widetilde{\mathbf{H}}^H \boldsymbol{\Phi}_y^{-1} - \lambda_x \mathbf{H}^H \boldsymbol{\Phi}_y^{-1} \widetilde{\mathbf{H}} \mathbf{H}^H \boldsymbol{\Phi}_y^{-1} \Big)\end{aligned} \tag{38}$$

where we omitted the argument ω for simplicity. Inserting (38) in (37) results in the approximation

$$\begin{aligned}J'(\omega; \widetilde{\mathbf{H}}, \mathbf{H}) \simeq \lambda_x^2 \mathrm{tr}\Big(& \mathbf{Q}\widetilde{\mathbf{H}}^H \boldsymbol{\Phi}_y^{-1} \widetilde{\mathbf{H}} \mathbf{Q} \\ & + \lambda_x^2 \left(\mathbf{H}^H \boldsymbol{\Phi}_y^{-1} \widetilde{\mathbf{H}} \mathbf{H}^H \boldsymbol{\Phi}_y^{-1} \mathbf{H} \widetilde{\mathbf{H}}^H \boldsymbol{\Phi}_y^{-1} \mathbf{H} \right) \\ & - \lambda_x \mathbf{Q}\widetilde{\mathbf{H}}^H \boldsymbol{\Phi}_y^{-1} \mathbf{H} \widetilde{\mathbf{H}}^H \boldsymbol{\Phi}_y^{-1} \mathbf{H} \\ & - \lambda_x \mathbf{H}^H \boldsymbol{\Phi}_y^{-1} \widetilde{\mathbf{H}} \mathbf{H}^H \boldsymbol{\Phi}_y^{-1} \widetilde{\mathbf{H}} \mathbf{Q} \Big).\end{aligned} \tag{39}$$

To rewrite this into a quadratic form in terms of $\mathrm{vec}(\widetilde{\mathbf{H}})$ we use the facts that $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}) = \mathrm{vec}^T(\mathbf{A}^T)\,\mathrm{vec}(\mathbf{B}) = \mathrm{vec}^H(\mathbf{A}^H)\,\mathrm{vec}(\mathbf{B})$ and $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\,\mathrm{vec}(\mathbf{B})$ for matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ of compatible dimensions. Hence, we can rewrite (39) as

$$\begin{aligned} J'(\omega;\widetilde{\mathbf{H}},\mathbf{H}) &\simeq \mathrm{vec}^H(\widetilde{\mathbf{H}})[\lambda_x^2 {\mathbf{Q}^2}^T \otimes \mathbf{\Phi}_y^{-1}]\,\mathrm{vec}(\widetilde{\mathbf{H}}) \\ &+ \mathrm{vec}^H(\widetilde{\mathbf{H}})[\lambda_x^4(\mathbf{H}^H\mathbf{\Phi}_y^{-1}\mathbf{H})^T \otimes \mathbf{\Phi}_y^{-1}\mathbf{H}\mathbf{H}^H\mathbf{\Phi}_y^{-1}]\,\mathrm{vec}(\widetilde{\mathbf{H}}) \\ &- \mathrm{vec}^H(\widetilde{\mathbf{H}})[\lambda_x^3(\mathbf{\Phi}_y^{-1}\mathbf{H}\mathbf{Q})^T \otimes \mathbf{\Phi}_y^{-1}\mathbf{H}]\,\mathrm{vec}(\widetilde{\mathbf{H}}^H) \\ &- \mathrm{vec}^H(\widetilde{\mathbf{H}}^H)[\lambda_x^3(\mathbf{Q}\mathbf{H}^H\mathbf{\Phi}_y^{-1})^T \otimes \mathbf{H}^H\mathbf{\Phi}_y^{-1}]\,\mathrm{vec}(\widetilde{\mathbf{H}}). \end{aligned} \tag{40}$$

In the next step, we introduce the permutation matrix $\mathbf{\Pi}$ defined such that $\mathrm{vec}(\widetilde{\mathbf{H}}^T) = \mathbf{\Pi}\,\mathrm{vec}(\widetilde{\mathbf{H}})$ for every $\widetilde{\mathbf{H}}$ to rewrite (40) as

$$\begin{aligned} J'(\omega;\widetilde{\mathbf{H}},\mathbf{H}) &\simeq \mathrm{vec}^H(\widetilde{\mathbf{H}})[\lambda_x^2 {\mathbf{Q}^2}^T \otimes \mathbf{\Phi}_y^{-1}]\,\mathrm{vec}(\widetilde{\mathbf{H}}) \\ &+ \mathrm{vec}^H(\widetilde{\mathbf{H}})[\lambda_x^4(\mathbf{H}^H\mathbf{\Phi}_y^{-1}\mathbf{H})^T \otimes \mathbf{\Phi}_y^{-1}\mathbf{H}\mathbf{H}^H\mathbf{\Phi}_y^{-1}]\,\mathrm{vec}(\widetilde{\mathbf{H}}) \\ &- \mathrm{vec}^H(\widetilde{\mathbf{H}})[\lambda_x^3(\mathbf{\Phi}_y^{-1}\mathbf{H}\mathbf{Q})^T \otimes \mathbf{\Phi}_y^{-1}\mathbf{H}]\mathbf{\Pi}\,\mathrm{vec}(\widetilde{\mathbf{H}}^*) \\ &- \mathrm{vec}^H(\widetilde{\mathbf{H}}^*)\mathbf{\Pi}^T[\lambda_x^3(\mathbf{Q}\mathbf{H}^H\mathbf{\Phi}_y^{-1})^T \otimes \mathbf{H}^H\mathbf{\Phi}_y^{-1}]\,\mathrm{vec}(\widetilde{\mathbf{H}}). \end{aligned} \tag{41}$$

We have now obtained a quadratic form. Note indeed that the last two terms are just complex conjugates of each other and thus we can write them as two times their real part.

A. High SNR analysis

In order to obtain a simpler expression for $\mathcal{I}_{\text{adm}}$, we will assume high SNR in the data transmission phase. We consider the practically relevant case where $\mathrm{rank}\,(\mathbf{H}) = \min(n_T, n_R)$. Depending on the rank of the channel matrix $\mathbf{H}$ we will have three different cases:

Case 1: $\mathrm{rank}(\mathbf{H}) = n_R < n_T$*:* Under this assumption, it can be shown that both the first and the second terms on the right hand side of (41) contribute to $\mathcal{I}_{\text{adm}}$. We have $\mathbf{Q} \to \mathbf{\Pi}^{\perp}_{\mathbf{H}^H}$ and $\lambda_x\mathbf{\Phi}_y^{-1} \to (\mathbf{H}\mathbf{H}^H)^{-1}$ for high SNR. Here, and in what follows, we use $\mathbf{\Pi}_{\mathbf{X}} = \mathbf{X}\mathbf{X}^{\dagger}$ to denote the orthogonal projection matrix on the range-space of $\mathbf{X}$ and $\mathbf{\Pi}^{\perp}_{\mathbf{X}} = \mathbf{I} - \mathbf{\Pi}_{\mathbf{X}}$ to denote the projection on the nullspace of $\mathbf{X}^H$. Moreover, $\lambda_x\mathbf{H}^H\mathbf{\Phi}_y^{-1}\mathbf{H} \to \mathbf{\Pi}_{\mathbf{H}^H}$ and $\lambda_x^2\mathbf{\Phi}_y^{-1}\mathbf{H}\mathbf{H}^H\mathbf{\Phi}_y^{-1} \to (\mathbf{H}\mathbf{H}^H)^{-1}$ for high SNR. As $\mathbf{\Pi}^{\perp}_{\mathbf{H}^H} + \mathbf{\Pi}_{\mathbf{H}^H} = \mathbf{I}$, summing the contributions from the first two terms in (41) finally gives the high SNR approximation

$$\mathcal{I}_{\text{adm}} = \lambda_x\mathbf{I} \otimes (\mathbf{H}\mathbf{H}^H)^{-1}. \tag{42}$$

Case 2: $\mathrm{rank}(\mathbf{H}) = n_R = n_T$*:* For the non-singular channel case, the second term on the right hand side of (41) dominates. Here, we have $\lambda_x\mathbf{H}^H\mathbf{\Phi}_y^{-1}\mathbf{H} \to \mathbf{I}$ and $\lambda_x^2\mathbf{\Phi}_y^{-1}\mathbf{H}\mathbf{H}^H\mathbf{\Phi}_y^{-1} \to (\mathbf{H}\mathbf{H}^H)^{-1}$ for high SNR. Clearly, this results in the same expression for $\mathcal{I}_{\text{adm}}$ as in Case 1, namely,

$$\mathcal{I}_{\text{adm}} = \lambda_x\mathbf{I} \otimes (\mathbf{H}\mathbf{H}^H)^{-1}. \tag{43}$$

Case 3: $\mathrm{rank}(\mathbf{H}) = n_T < n_R$*:* In this case, the second term on the right hand side of (41) dominates. When $\mathrm{rank}(\mathbf{H}) = n_T$ we get $\lambda_x\mathbf{H}^H\mathbf{\Phi}_y^{-1}\mathbf{H} \to \mathbf{I}$ and $\lambda_x^2\mathbf{\Phi}_y^{-1}\mathbf{H}\mathbf{H}^H\mathbf{\Phi}_y^{-1} \to \mathbf{\Phi}_n^{-1/2}[\mathbf{\Phi}_n^{-1/2}\mathbf{H}\mathbf{H}^H\mathbf{\Phi}_n^{-1/2}]^{\dagger}\mathbf{\Phi}_n^{-1/2}$ for high SNR. Using these approximations finally gives the high SNR approximation

$$\mathcal{I}_{\text{adm}} = \lambda_x\mathbf{I} \otimes \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbf{\Phi}_n^{-1/2}[\mathbf{\Phi}_n^{-1/2}\mathbf{H}\mathbf{H}^H\mathbf{\Phi}_n^{-1/2}]^{\dagger}\mathbf{\Phi}_n^{-1/2}\; d\omega\right).$$

B. Low SNR analysis

For the low SNR regime, we do not need to differentiate our analysis for the cases $n_T \geq n_R$ and $n_T < n_R$, because now $\mathbf{\Phi}_y \to \mathbf{\Phi}_n$. It can be shown that the first term on the right hand side of (41) dominates; that is, the term involving

$$\lambda_x^2\;\left((\mathbf{Q}^2)^T \otimes \mathbf{\Phi}_y^{-1}\right).$$

Moreover, $\mathbf{Q} \to \mathbf{I}$ and $\mathbf{\Phi}_y^{-1} \to \mathbf{\Phi}_n^{-1}$. This yields

$$\mathcal{I}_{\text{adm}} = \mathbf{I} \otimes \left(\frac{\lambda_x^2}{2\pi}\int_{-\pi}^{\pi} \mathbf{\Phi}_n^{-1}\; d\omega\right). \tag{44}$$

Appendix II
Proof of Theorem 1

For the proof of Theorem 1, we require some preliminary results. Lemma 1 and Lemma 2 will be used to establish the uniqueness part of Theorem 1, and Lemma 3 is an extension of a standard result in majorization theory, which is used in the main part of the proof.

Lemma 1: Let $\mathbf{D} \in \mathbb{R}^{n\times n}$ be a diagonal matrix with elements $d_{1,1} > \cdots > d_{n,n} > 0$. If $\mathbf{U} \in \mathbb{C}^{n\times n}$ is a unitary matrix such that $\mathbf{U}\mathbf{D}\mathbf{U}^H$ has diagonal $(d_{1,1},\ldots,d_{n,n})$, then $\mathbf{U}$ is of the form $\mathbf{U} = \mathrm{diag}(u_{1,1},\ldots,u_{n,n})$, where $|u_{i,i}| = 1$ for $i = 1,\ldots,n$. This also implies that $\mathbf{U}\mathbf{D}\mathbf{U}^H = \mathbf{D}$.

Proof: Let $\mathbf{V} = \mathbf{U}\mathbf{D}\mathbf{U}^H$. The equation for $(\mathbf{V})_{i,i}$ is

$$\sum_{k=1}^{n} d_{k,k}|u_{i,k}|^2 = d_{i,i}$$

from which we have, by the orthonormality of the columns of $\mathbf{U}$, that

$$\sum_{k=1}^{n} \frac{d_{k,k}}{d_{i,i}}|u_{i,k}|^2 = 1 = \sum_{k=1}^{n}|u_{i,k}|^2. \tag{45}$$

We now proceed by induction on $i = 1,\ldots,n$ to show that the ith column of $\mathbf{U}$ is $[0\ \cdots\ 0\ u_{i,i}\ 0\ \cdots\ 0]^T$ with $|u_{i,i}| = 1$. For $i = 1$, it follows from (45) and the fact that $\mathbf{U}$ is unitary that

$$\begin{aligned} |u_{1,1}|^2 + \left|\frac{d_{2,2}}{d_{1,1}}u_{2,1}\right|^2 + \cdots + \left|\frac{d_{n,n}}{d_{1,1}}u_{n,1}\right|^2 \\ = |u_{1,1}|^2 + \cdots + |u_{n,1}|^2 = 1. \end{aligned}$$

However, since $d_{1,1} > \cdots > d_{n,n} > 0$, the only way to satisfy this equation is to have $|u_{1,1}| = 1$ and $u_{i,1} = 0$ for $i = 2,\ldots,n$. Now, if the assertion holds for $i = 1,\ldots,k$, the orthogonality of the columns of $\mathbf{U}$ implies that $u_{i,k+1} = 0$ for $i = 1,\ldots,k$, and by following a similar reasoning as for the case $i = 1$ we deduce that $|u_{k+1,k+1}| = 1$ and $u_{i,k+1} = 0$ for $i = k+2,\ldots,n$. ∎

Lemma 2: Let $\mathbf{D} \in \mathbb{R}^{N\times N}$ be a diagonal matrix with elements $d_{1,1} > \cdots > d_{N,N} > 0$. If $\mathbf{U} \in \mathbb{C}^{N\times n}$, with

$n \leq N$, is such that $\mathbf{U}^H\mathbf{U} = I$ and $\mathbf{V} = \widetilde{\mathbf{D}}\mathbf{U}\widetilde{\mathbf{D}}^{-1}$ (where $\widetilde{\mathbf{D}} = \mathrm{diag}(d_{1,1}, \ldots, d_{n,n})$) also satisfies $\mathbf{V}^H\mathbf{V} = \mathbf{I}$, then $\mathbf{U}$ is of the form $\mathbf{U} = [\mathrm{diag}(u_{1,1}, \ldots, u_{n,n}) \quad \mathbf{0}_{N-m,n}]^T$, where $|u_{i,i}| = 1$ for $i = 1, \ldots, n$.

Proof: The idea is similar to the proof of Lemma 1. We proceed by induction on the ith column of $\mathbf{V}$. For the first column of $\mathbf{V}$ we have, by the orthonormality of the columns of $\mathbf{U}$ and $\mathbf{V}$, that

$$\begin{aligned}|u_{1,1}|^2 + \left|\frac{d_{2,2}}{d_{1,1}}u_{2,1}\right|^2 + \cdots + \left|\frac{d_{N,N}}{d_{1,1}}u_{N,1}\right|^2 &= 1 \\ &= |u_{1,1}|^2 + \cdots + |u_{N,1}|^2.\end{aligned}$$

Since $d_{1,1} > \cdots > d_{N,N} > 0$, the only way to satisfy this equation is to have $|u_{1,1}| = 1$ and $u_{i,1} = 0$ for $i = 2, \ldots, N$. If now the assertion holds for columns 1 to k, the orthogonality of the columns of $\mathbf{U}$ implies that $u_{i,k+1} = 0$ for $i = 1, \ldots, k$, and by following a similar reasoning as for the first column of $\mathbf{U}$ we have that $|u_{k+1,k+1}| = 1$ and $u_{i,k+1} = 0$ for $i = k+2, \ldots, N$. ∎

Lemma 3: Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}$ be Hermitian matrices. Arrange the eigenvalues $a_1, \ldots, a_n$ of $\mathbf{A}$ in a descending order, and the eigenvalues $b_1, \ldots, b_n$ of $\mathbf{B}$ in an ascending order. Then $\mathrm{tr}(\mathbf{AB}) \geq \sum_{i=1}^n a_i b_i$. Furthermore, if $\mathbf{B} = \mathrm{diag}(b_1, \ldots, b_n)$ and both matrices have distinct eigenvalues, then $\mathrm{tr}(\mathbf{AB}) = \sum_{i=1}^n a_i b_i$ if and only if $\mathbf{A} = \mathrm{diag}(a_1, \ldots, a_n)$.

Proof: See [34, Theorem 9.H.1.h] for the proof of the first assertion. For the second part, notice that if $\mathbf{B} = \mathrm{diag}(b_1, \ldots, b_n)$, then by [34, Theorem 6.A.3]

$$\mathrm{tr}(\mathbf{AB}) = \sum_{i=1}^n (\mathbf{A})_{i,i} b_i \geq \sum_{i=1}^n (\mathbf{A})_{[i,i]} b_i$$

where $\{(\mathbf{A})_{[i,i]}\}_{i=1,\ldots,n}$ denotes the ordered set $\{(\mathbf{A})_{1,1}, \ldots, (\mathbf{A})_{n,n}\}$ sorted in descending order. Since $\{(\mathbf{A})_{[i,i]}\}_{i=1,\ldots,n}$ is *majorized* by $\{a_1, \ldots, a_n\}$, and the b_i's are distinct, we can use [34, Theorem 3.A.2] to show that

$$\sum_{i=1}^n (\mathbf{A})_{[i,i]} b_i > \sum_{i=1}^n a_i b_i$$

unless $(\mathbf{A})_{[i,i]} = a_i$ for every $i = 1, \ldots, n$. Therefore, $\mathrm{tr}(\mathbf{AB}) = \sum_{i=1}^n a_i b_i$ if and only if the diagonal of $\mathbf{A}$ is $(a_1, \ldots, a_n)$. Now we have to prove that $\mathbf{A}$ is actually diagonal, but this follows from Lemma 1. ∎

Proof of Theorem 1. First, we simplify the expressions in (21). Using the eigendecompositions in (23) of $\mathbf{A}$ and $\mathbf{B}$, we see that

$$\begin{aligned}\mathbf{PA}^{-1}\mathbf{P}^H \succeq \mathbf{B} \quad &\Leftrightarrow \quad \mathbf{PU}_A\mathbf{D}_A^{-1}\mathbf{U}_A^H\mathbf{P}^H \succeq \mathbf{U}_B\mathbf{D}_B\mathbf{U}_B^H \\ &\Leftrightarrow \quad \mathbf{U}_B^H\mathbf{PU}_A\mathbf{D}_A^{-1}\mathbf{U}_A^H\mathbf{P}^H\mathbf{U}_B \succeq \mathbf{D}_B.\end{aligned}$$

Now, define $\bar{\mathbf{P}} = \mathbf{U}_B^H\mathbf{PU}_A\mathbf{D}_A^{-1/2}$ and observe that

$$\begin{aligned}\mathrm{tr}(\mathbf{PP}^H) &= \mathrm{tr}\left((\mathbf{U}_B\bar{\mathbf{P}}\mathbf{D}_A^{-H/2}\mathbf{U}_A^H)(\mathbf{U}_B\bar{\mathbf{P}}\mathbf{D}_A^{-H/2}\mathbf{U}_A^H)^H\right) \\ &= \mathrm{tr}(\mathbf{U}_B\bar{\mathbf{P}}\mathbf{D}_A^{-1}\bar{\mathbf{P}}^H\mathbf{U}_B^H) = \mathrm{tr}(\bar{\mathbf{P}}^H\bar{\mathbf{P}}\mathbf{D}_A^{-1}).\end{aligned}$$

Therefore, (21) is equivalent to

$$\begin{array}{ll}\underset{\mathbf{P}\in\mathbb{C}^{n\times N}}{\text{minimize}} & \mathrm{tr}(\bar{\mathbf{P}}^H\bar{\mathbf{P}}\mathbf{D}_A^{-1}) \\ \text{s.t.} & \bar{\mathbf{P}}\bar{\mathbf{P}}^H \succeq \mathbf{D}_B.\end{array} \tag{46}$$

To further simplify our problem, consider the singular value decomposition $\bar{\mathbf{P}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$, where $\mathbf{U} \in \mathbb{C}^{n\times n}$ and $\mathbf{V} \in \mathbb{C}^{N\times N}$ are unitary matrices and $\mathbf{\Sigma}$ has the structure

$$\mathbf{\Sigma} = \begin{bmatrix}\sigma_1 & & 0 & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & \sigma_m & 0 & \cdots & 0\end{bmatrix} \text{ or } \mathbf{\Sigma} = \begin{bmatrix}\sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_m \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0\end{bmatrix}$$

depending on whether $N \geq n$ or $N < n$. The singular values are ordered such that $\sigma_1 \geq \cdots \geq \sigma_m > 0$. Now, observe that (46) is equivalent to

$$\begin{array}{ll}\underset{\mathbf{P}\in\mathbb{C}^{n\times N}}{\text{minimize}} & \mathrm{tr}(\mathbf{V}^H\mathbf{\Sigma}^H\mathbf{\Sigma}\mathbf{V}^H\mathbf{D}_A^{-1}) \\ \text{s.t.} & \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^H\mathbf{U}^H \succeq \mathbf{D}_B.\end{array} \tag{47}$$

With this problem formulation, it follows (from Sylvester's law of inertia [35]) that we need $m \geq \mathrm{rank}(\mathbf{D}_B)$ to achieve feasibility in the constraint (i.e., having at least as many non-zero singular values of $\mathbf{\Sigma}$ as non-zero eigenvalues in $\mathbf{D}_B$). This corresponds to the condition $N \geq \mathrm{rank}(\mathbf{B})$ in the theorem.

Now we will show that $\mathbf{U}$ and $\mathbf{V}$ can be taken to be the identity matrices. Using Lemma 3, the cost function can be lower bounded as

$$\begin{aligned}\mathrm{tr}(\mathbf{V}\mathbf{\Sigma}^H\mathbf{\Sigma}\mathbf{V}^H\mathbf{D}_A^{-1}) &\geq \sum_{j=1}^n \lambda_{n-j+1}(\mathbf{D}_A)\lambda_j(\mathbf{V}\mathbf{\Sigma}^H\mathbf{\Sigma}\mathbf{V}^H) \\ &= \sum_{j=1}^m (\mathbf{D}_A)_{jj}\sigma_j^2\end{aligned} \tag{48}$$

where $\lambda_j(\cdot)$ denotes the jth largest eigenvalue. The equality is achieved if $\mathbf{V} = \mathbf{I}$, and observe that we can select $\mathbf{V}$ in this manner without affecting the constraint.

To show that $\mathbf{U}$ can also be taken as the identity matrix, notice that the cost function in (47) does not depend on $\mathbf{U}$, while the constraint implies (by looking at the diagonal elements of the inequality and recalling that $\mathbf{U}$ is unitary) that

$$\sigma_i^2 \geq (\mathbf{D}_B)_{i,i}, \qquad i = 1, \ldots, m, \tag{49}$$

requiring $m \geq \mathrm{rank}(\mathbf{D}_B)$. Suppose that $\bar{\mathbf{U}}$ and $\bar{\mathbf{\Sigma}}$ minimize the cost. Then, we can replace $\bar{\mathbf{U}}$ by $\mathbf{I}$ and satisfy the constraint, without affecting the cost in (48). This means that there exists an optimal solution with $\mathbf{U} = \mathbf{I}$.

With $\mathbf{U} = \mathbf{I}$ and $\mathbf{V} = \mathbf{I}$, the problem (47) is equivalent (in terms of $\mathbf{\Sigma}$) to

$$\begin{array}{ll}\underset{\sigma_1\geq 0,\ldots,\sigma_m\geq 0}{\text{minimize}} & \sum_{i=1}^m \sigma_i^2(\mathbf{D}_A)_{i,i} \\ \text{s.t.} & \sigma_i^2 \geq (\mathbf{D}_B)_{i,i}, \quad i = 1, \ldots, m.\end{array}$$

It is easy to see that the optimal solution for this problem is $\sigma_i^{\text{opt}} = \sqrt{(\mathbf{D}_B)_{i,i}}, i = 1, \ldots, m$. By creating an optimal $\mathbf{\Sigma}$,

denoted as $\mathbf{\Sigma}^{\text{opt}}$, with the singular values $\sigma_1^{\text{opt}}, \ldots, \sigma_m^{\text{opt}}$, we achieve an optimal solution

$$\mathbf{P}^{\text{opt}} = \mathbf{U}_B \bar{\mathbf{P}} \mathbf{D}_A^{1/2} \mathbf{U}_A^H = \mathbf{U}_B \mathbf{\Sigma}^{\text{opt}} \mathbf{D}_A^{1/2} \mathbf{U}_A^H = \mathbf{U}_B \mathbf{D}_P \mathbf{U}_A^H$$

with $\mathbf{D}_P$ as stated in the theorem.

Finally, we will show how to characterize all optimal solutions for the case when $\mathbf{A}$ and $\mathbf{B}$ have distinct non-zero eigenvalues (thus, $m = n$). The optimal solutions need to give equality in (48) and thus Lemma 3 gives that $\mathbf{V}\mathbf{\Sigma}\mathbf{\Sigma}^H\mathbf{V}^H$ is diagonal and equal to $\mathbf{\Sigma}\mathbf{\Sigma}^H$. Lemma 1 then implies that $\mathbf{V} = \text{diag}(v_{1,1}, \ldots, v_{n,n})$ with $|v_{i,i}| = 1$ for $i = 1, \ldots, n$.

For the optimal $\mathbf{\Sigma}$, we have that $\sigma_i^2 = (\mathbf{D}_B)_{i,i}$ for $i = 1, \ldots, n$, so the diagonal elements of $\mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^H\mathbf{U}^H - \mathbf{D}_B$ are zero. Since $\mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^H\mathbf{U}^H - \mathbf{D}_B \succeq 0$ for every feasible solution of (47), $\mathbf{U}$ has to satisfy $\mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^H\mathbf{U}^H = \mathbf{D}_B$. Lemma 2 then establishes that the first n columns of $\mathbf{U}$ are of the form $[\text{diag}(u_{1,1}, \ldots, u_{n,n}) \quad \mathbf{0}_{N-m,n}]^T$, where $|u_{i,i}| = 1$ for $i = 1, \ldots, n$. Since $\mathbf{U}$ has to be unitary, and its last $N-n+1$ columns play no role in $\bar{\mathbf{P}}$ (due to the form of $\mathbf{\Sigma}$), we can take them as $[\mathbf{0}_{n,N-m+1} \quad \mathbf{I}_{N-m+1}]^T$ without loss of generality.

Summarizing, an optimal solution is given by (23). When $\mathbf{A}$ and $\mathbf{B}$ have distinct eigenvalues, $\mathbf{V}$ and $\mathbf{U}$ can only multiply the columns of $\mathbf{U}_A$ and $\mathbf{U}_B$, respectively, by complex scalars of unit magnitude.

Appendix III
Proof of Theorem 2 and Theorem 3

Before proving Theorem 2 and 3, a lemma will be given that characterizes equivalences between different sets of feasible training matrices $\mathbf{P}$.

Lemma 4: Let $\mathbf{B} \in \mathbb{C}^{n\times n}$ and $\mathbf{C} \in \mathbb{C}^{m\times m}$ be Hermitian matrices, and $f : \mathbb{C}^{n\times N} \to \mathbb{C}^{n\times n}$ be such that $f(\mathbf{P}) = f(\mathbf{P})^H$. Then, the following sets are equivalent

$$\{\mathbf{P} | f(\mathbf{P}) \otimes \mathbf{I} \succeq \mathbf{B} \otimes \mathbf{C}\} = \{\mathbf{P} | f(\mathbf{P}) \succeq \lambda_{\max}(\mathbf{C})\mathbf{B}\}. \quad (50)$$

Proof: The equivalence will be proved by showing that the left hand side (LHS) is a subset of right hand side (RHS), and *vice versa*. First, assume that $f(\mathbf{P}) \succeq \lambda_{\max}(\mathbf{C})\mathbf{B}$, then

$$\begin{aligned} f(\mathbf{P}) \otimes \mathbf{I} &\succeq \lambda_{\max}(\mathbf{C})\mathbf{B} \otimes \mathbf{I} \\ &= (\mathbf{B} \otimes \lambda_{\max}(\mathbf{C})\mathbf{I}) \succeq (\mathbf{B} \otimes \mathbf{C}). \end{aligned} \quad (51)$$

Hence, RHS $\subseteq$ LHS.

Next, assume that $f(\mathbf{P}) \otimes \mathbf{I} \succeq \mathbf{B} \otimes \mathbf{C}$, but for the purpose of contradiction that $f(\mathbf{P}) \not\succeq \lambda_{\max}(\mathbf{C})\mathbf{B}$. Then, there exists a vector $\mathbf{x}$ such that $\mathbf{x}^H(f(\mathbf{P}) - \lambda_{\max}(\mathbf{C})\mathbf{B})\mathbf{x} < 0$. Let $\mathbf{v}$ be an eigenvector of $\mathbf{C}$ that corresponds to $\lambda_{\max}(\mathbf{C})$ and define $\mathbf{y} = \mathbf{x} \otimes \mathbf{v}$. Then

$$\begin{aligned} &\mathbf{y}(f(\mathbf{P}) \otimes \mathbf{I} - \mathbf{B} \otimes \mathbf{C})\mathbf{y} \\ &\quad = (\mathbf{x}^H f(\mathbf{P})\mathbf{x})\|\mathbf{v}\|^2 - (\mathbf{x}^H\mathbf{B}\mathbf{x})(\mathbf{v}^H\mathbf{C}\mathbf{v}) \\ &\quad = \mathbf{x}^H(f(\mathbf{P}) - \lambda_{\max}(\mathbf{C})\mathbf{B})\mathbf{x}\|\mathbf{v}\|^2 < 0 \end{aligned} \quad (52)$$

which is a contradiction. Hence, LHS $\subseteq$ RHS. ∎

Proof of Theorem 2. Rewrite the constraint as

$$\begin{aligned} &\widetilde{\mathbf{P}}^H(\mathbf{S}_Q^T \otimes \mathbf{S}_R)^{-1}\widetilde{\mathbf{P}} \succeq c\boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R \\ \Leftrightarrow \quad &(\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H)^T \otimes \mathbf{S}_R^{-1} \succeq c\boldsymbol{\mathcal{I}}_T^T \otimes \boldsymbol{\mathcal{I}}_R \\ \Leftrightarrow \quad &(\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H) \otimes \mathbf{I} \succeq c\boldsymbol{\mathcal{I}}_T \otimes \mathbf{S}_R\boldsymbol{\mathcal{I}}_R. \end{aligned} \quad (53)$$

Let $f(\mathbf{P}) = \mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H$. Then Lemma 4 gives that the set of feasible $\mathbf{P}$ is equivalent to the set of feasible $\mathbf{P}$ with the constraint

$$(\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H) \succeq c\lambda_{\max}(\mathbf{S}_R\boldsymbol{\mathcal{I}}_R)\boldsymbol{\mathcal{I}}_T. \quad (54)$$

Proof of Theorem 3.

In the case that $\mathbf{R}_R = \mathbf{S}_R$, the constraint can be rewritten as

$$(\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H + \mathbf{R}_T^{-1})^T \otimes \mathbf{I} \succeq c\boldsymbol{\mathcal{I}}_T^T \otimes \mathbf{S}_R\boldsymbol{\mathcal{I}}_R. \quad (55)$$

With $f(\mathbf{P}) = \mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H + \mathbf{R}_T^{-1}$, Lemma 4 can be applied to achieve the equivalent constraint

$$\begin{aligned} &\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H + \mathbf{R}_T^{-1} \succeq c\lambda_{\max}(\mathbf{S}_R\boldsymbol{\mathcal{I}}_R)\boldsymbol{\mathcal{I}}_T \\ \Leftrightarrow \quad &\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H \succeq c\lambda_{\max}(\mathbf{S}_R\boldsymbol{\mathcal{I}}_R)\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T^{-1} \\ \Leftrightarrow \quad &\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H \succeq [c\lambda_{\max}(\mathbf{S}_R\boldsymbol{\mathcal{I}}_R)\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T^{-1}]_+ \end{aligned} \quad (56)$$

where the last equality follows from the fact that the left hand side is positive semi-definite.

In the case that $\mathbf{R}_R^{-1} = \boldsymbol{\mathcal{I}}_R$, the constraint can be rewritten as

$$\begin{aligned} &(\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H)^T \otimes \mathbf{S}_R^{-1} \succeq (c\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T)^T \otimes \boldsymbol{\mathcal{I}}_R \\ \Leftrightarrow \quad &(\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H)^T \otimes \mathbf{S}_R^{-1} \succeq [c\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T]_+^T \otimes \boldsymbol{\mathcal{I}}_R. \end{aligned} \quad (57)$$

Observe that this expression is identical to the constraint in (24), except that the positive semi-definite $\boldsymbol{\mathcal{I}}_T$ has been replaced by $[c\boldsymbol{\mathcal{I}}_T - \mathbf{R}_T]_+$. Thus, the equivalence follows directly from Theorem 2.

In the case $\mathbf{R}_T^{-1} = \boldsymbol{\mathcal{I}}_T$, the constraint can be rewritten as

$$\begin{aligned} &(\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H)^T \otimes \mathbf{S}_R^{-1} \succeq \boldsymbol{\mathcal{I}}_T^T \otimes (c\boldsymbol{\mathcal{I}}_R - \mathbf{R}_R) \\ \Leftrightarrow \quad &(\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H)^T \otimes \mathbf{S}_R^{-1} \succeq \boldsymbol{\mathcal{I}}_T^T \otimes [c\boldsymbol{\mathcal{I}}_R - \mathbf{R}_R]_+. \end{aligned} \quad (58)$$

As in the previous case, the equivalence follows directly from Theorem 2.

Appendix IV
Proof of Theorem 4

Our basic assumption is that $\boldsymbol{\mathcal{I}}_T, \boldsymbol{\mathcal{I}}_R$ are both Hermitian matrices, which is encountered in the applications presented in this paper. Denoting by $\mathbf{P}'$ the matrix $\mathbf{P}^T$ and using the fact that[4] $\boldsymbol{\mathcal{I}}_{\text{adm}} = \left(\boldsymbol{\mathcal{I}}_T' \otimes \boldsymbol{\mathcal{I}}_R\right)^{1/2}\left(\boldsymbol{\mathcal{I}}_T' \otimes \boldsymbol{\mathcal{I}}_R\right)^{1/2}$, it can be seen that our optimization problem takes the following form

$$\begin{aligned} \underset{\mathbf{P}' \in \mathbb{C}^{B\times n_T}}{\text{minimize}} \quad & J(\mathbf{H}) \\ \text{s.t.} \quad & \text{tr}(\mathbf{P}'\mathbf{P}'^H) \leq \mathcal{P} \end{aligned} \quad (59)$$

where $J(\mathbf{H}) = \mathbb{E}_{\widetilde{\mathbf{H}}}\left\{J(\widetilde{\mathbf{H}}, \mathbf{H})\right\}$ is given by the expression

$$\begin{aligned} &\text{tr}\left\{\left[\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{P}'^H\mathbf{S}_Q'^{-1}\mathbf{P}'\boldsymbol{\mathcal{I}}_T'^{-1/2} \otimes \boldsymbol{\mathcal{I}}_R^{-1/2}\mathbf{S}_R^{-1}\boldsymbol{\mathcal{I}}_R^{-1/2}\right]^{-1}\right\} \\ &= \text{tr}\left\{\left[\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{P}'^H\mathbf{S}_Q'^{-1}\mathbf{P}'\boldsymbol{\mathcal{I}}_T'^{-1/2}\right]^{-1} \otimes \boldsymbol{\mathcal{I}}_R^{1/2}\mathbf{S}_R\boldsymbol{\mathcal{I}}_R^{1/2}\right\}. \end{aligned}$$

[4]For a Hermitian positive semidefinite matrix $\mathbf{A}$, we consider here that $\mathbf{A}^{1/2}$ is the matrix with the same eigenvectors as $\mathbf{A}$ and eigenvalues the square roots of the corresponding eigenvalues of $\mathbf{A}$. With this definition of the square root of a Hermitian positive semidefinite matrix, it is clear that $\mathbf{A}^{1/2} = \mathbf{A}^{H/2}$, leading to $\mathbf{A} = \mathbf{A}^{1/2}\mathbf{A}^{H/2} = \mathbf{A}^{H/2}\mathbf{A}^{1/2}$.

Using the fact that $\mathrm{tr}\left(\mathbf{A}\otimes\mathbf{B}\right)=\mathrm{tr}\left(\mathbf{A}\right)\mathrm{tr}\left(\mathbf{B}\right)$ for square matrices $\mathbf{A}$ and $\mathbf{B}$, it is clear from the last expression that the optimal training matrix can be found by minimizing

$$\mathrm{tr}\left\{\left[\mathbf{V}_T^H\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{P}'^H\mathbf{S}'^{-1}_Q\mathbf{P}'\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{V}_T\right]^{-1}\right\}, \tag{60}$$

where $\mathbf{V}_T$ denotes the modal matrix of $\boldsymbol{\mathcal{I}}_T'$ corresponding to an arbitrary ordering of its eigenvalues. Here, we have used the invariance of the trace operator under unitary transformations. First, note that for an arbitrary Hermitian positive definite matrix $\mathbf{A}$, $\mathrm{tr}\left(\mathbf{A}^{-1}\right)=\sum_i 1/\lambda_i\left(\mathbf{A}\right)$, where $\lambda_i\left(\mathbf{A}\right)$ is the ith eigenvalue of $\mathbf{A}$. Since the function $1/x$ is strictly convex for $x>0$, $\mathrm{tr}\left(\mathbf{A}^{-1}\right)$ is a Schur-convex function with respect to the eigenvalues of $\mathbf{A}$ [34]. Additionally, for any Hermitian matrix $\mathbf{A}$, the vector of its diagonal entries is majorized by the vector of its eigenvalues [34]. Combining the last two results, it follows that $\mathrm{tr}\left(\mathbf{A}^{-1}\right)$ is minimized when $\mathbf{A}$ is diagonal. Therefore, we may choose the modal matrices of $\mathbf{P}'$ in such a way that $\mathbf{V}_T^H\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{P}'^H\mathbf{S}'^{-1}_Q\mathbf{P}'\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{V}_T$ is diagonalized. Suppose that the singular value decomposition (SVD) of $\mathbf{P}'^H$ is $\mathbf{U}\mathbf{D}_{P'}\mathbf{V}^H$ and that the modal matrix of $\mathbf{S}'_Q$, corresponding to arbitrary ordering of its eigenvalues, is $\mathbf{V}_Q$. Setting $\mathbf{U}=\mathbf{V}_T$ and $\mathbf{V}=\mathbf{V}_Q$, $\mathbf{V}_T^H\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{P}'^H\mathbf{S}'^{-1}_Q\mathbf{P}'\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{V}_T$ is diagonalized and is given by the expression

$$\boldsymbol{\Lambda}_T^{-1/2}\mathbf{D}_{P'}\boldsymbol{\Lambda}_Q^{-1}\mathbf{D}_{P'}\boldsymbol{\Lambda}_T^{-1/2}.$$

Here, $\boldsymbol{\Lambda}_T$ and $\boldsymbol{\Lambda}_Q$ are the diagonal eigenvalue matrices containing the eigenvalues of $\boldsymbol{\mathcal{I}}_T'$ and $\mathbf{S}'_Q$, respectively, in their main diagonals. The ordering of the eigenvalues corresponds to $\mathbf{V}_T$ and $\mathbf{V}_Q$. Clearly, by reordering the columns of $\mathbf{V}_T$ and $\mathbf{V}_Q$, we can reorder the eigenvalues in $\boldsymbol{\Lambda}_T$ and $\boldsymbol{\Lambda}_Q$. Assume that there are two different permutations π,ϖ such that $\pi\left(\left(\boldsymbol{\Lambda}_T\right)_{1,1}\right),\ldots,\pi\left(\left(\boldsymbol{\Lambda}_T\right)_{n_T,n_T}\right)$ and $\varpi\left(\left(\boldsymbol{\Lambda}_Q\right)_{1,1}\right),\ldots,\varpi\left(\left(\boldsymbol{\Lambda}_Q\right)_{B,B}\right)$ minimize $J(\mathbf{H})$ subject to our training energy constraint. Then, the entries of the corresponding eigenvalue matrix of $\mathbf{V}_T^H\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{P}'^H\mathbf{S}'^{-1}_Q\mathbf{P}'\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{V}_T$ are $\left(\mathbf{D}_{P'}\right)_{i,i}^2/\left(\pi\left(\left(\boldsymbol{\Lambda}_T\right)_{i,i}\right)\varpi\left(\left(\boldsymbol{\Lambda}_Q\right)_{i,i}\right)\right), i=1,2,\ldots,n_T$ ($B\geq n_T$). Setting $\left(\mathbf{D}_{P'}\right)_{i,i}^2=\kappa_i, i=1,2,\ldots,n_T$, the optimization problem (59) results in

$$\begin{array}{ll}\underset{\pi,\varpi,\kappa_i,i=1,2,\ldots,n_T}{\text{minimize}} & \sum_{i=1}^{n_T}\frac{1}{\frac{\kappa_i}{\pi\left(\left(\boldsymbol{\Lambda}_T\right)_{i,i}\right)\varpi\left(\left(\boldsymbol{\Lambda}_Q\right)_{i,i}\right)}}\\ \text{s.t.} & \sum_{i=1}^{n_T}\kappa_i\leq\mathcal{P}\end{array} \tag{61}$$

which leads to

$$\begin{array}{ll}\underset{\pi,\varpi,\kappa_i,i=1,2,\ldots,n_T}{\text{minimize}} & \sum_{i=1}^{n_T}\frac{\alpha_i}{\kappa_i}\\ \text{s.t.} & \sum_{i=1}^{n_T}\kappa_i\leq\mathcal{P}\end{array} \tag{62}$$

where $\alpha_i=\pi\left(\left(\boldsymbol{\Lambda}_T\right)_{i,i}\right)\varpi\left(\left(\boldsymbol{\Lambda}_Q\right)_{i,i}\right), i=1,2,\ldots,n_T$. Forming the Lagrangian of the last problem, it can be seen that

$$\left(\mathbf{D}_{P'}\right)_{i,i}=\sqrt{\frac{\mathcal{P}\sqrt{\alpha_i}}{\sum_{j=1}^{n_T}\sqrt{\alpha_j}}}, i=1,2,\ldots,n_T$$

while the objective value equals to $\left(\sum_{i=1}^{n_T}\sqrt{\alpha_i}\right)^2/\mathcal{P}$. Using Lemma 3, it can be seen that π and ϖ should correspond to opposite orderings of $\left(\boldsymbol{\Lambda}_T\right)_{i,i},\left(\boldsymbol{\Lambda}_Q\right)_{j,j}, i=1,2,\ldots,n_T, j=1,2,\ldots,B$, respectively. Since B can be greater than n_T, the eigenvalues of $\boldsymbol{\mathcal{I}}_T'$ must be set in decreasing order and those of $\mathbf{S}'_Q$ in increasing order.

Appendix V
Proof of Theorem 5

Using the factorization $\boldsymbol{\mathcal{I}}_{\text{adm}}=\left(\boldsymbol{\mathcal{I}}_T'\otimes\boldsymbol{\mathcal{I}}_R\right)^{1/2}\left(\boldsymbol{\mathcal{I}}_T'\otimes\boldsymbol{\mathcal{I}}_R\right)^{1/2}$, we can see that $E\left\{J(\tilde{\mathbf{H}},\mathbf{H})\right\}$ is given by the expression

$$\begin{aligned}\mathrm{tr}\quad&\left\{\left[\left(\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{R}_T'^{-1}\boldsymbol{\mathcal{I}}_T'^{-1/2}\otimes\boldsymbol{\mathcal{I}}_R^{-1/2}\mathbf{R}_R^{-1}\boldsymbol{\mathcal{I}}_R^{-1/2}\right)\right.\right.\\ +\quad&\left.\left.\left(\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{P}'^H\mathbf{S}'^{-1}_Q\mathbf{P}'\boldsymbol{\mathcal{I}}_T'^{-1/2}\otimes\boldsymbol{\mathcal{I}}_R^{-1/2}\mathbf{S}_R^{-1}\boldsymbol{\mathcal{I}}_R^{-1/2}\right)\right]^{-1}\right\},\end{aligned} \tag{63}$$

where $\mathbf{R}_T'=\mathbf{R}_T^T$ with eigenvalue decomposition $\mathbf{U}_T'\boldsymbol{\Lambda}_T'\mathbf{U}_T'^H$. This objective function subject to the training energy constraint $\mathrm{tr}(\mathbf{P}'\mathbf{P}'^H)\leq\mathcal{P}$ seems very difficult to minimize analytically unless special assumptions are made.

- $\mathbf{R}_R=\mathbf{S}_\mathbf{R}$: Then, (63) becomes

$$\begin{aligned}\mathrm{tr}\quad&\left\{\left(\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{R}_T'^{-1}\boldsymbol{\mathcal{I}}_T'^{-1/2}+\boldsymbol{\mathcal{I}}_T'^{-1/2}\mathbf{P}'^H\mathbf{S}'^{-1}_Q\mathbf{P}'\boldsymbol{\mathcal{I}}_T'^{-1/2}\right)^{-1}\right.\\ \otimes\quad&\left.\boldsymbol{\mathcal{I}}_R^{1/2}\mathbf{R}_R\boldsymbol{\mathcal{I}}_R^{1/2}\right\}.\end{aligned} \tag{64}$$

  Using once more the fact that $\mathrm{tr}\left(\mathbf{A}\otimes\mathbf{B}\right)=\mathrm{tr}\left(\mathbf{A}\right)\mathrm{tr}\left(\mathbf{B}\right)$ for square matrices $\mathbf{A}$ and $\mathbf{B}$, it is clear from (64) that the optimal training matrix can be found by minimizing

$$\mathrm{tr}\left\{\left(\mathbf{R}_T'^{-1}+\mathbf{P}'^H\mathbf{S}'^{-1}_Q\mathbf{P}'\right)^{-1}\boldsymbol{\mathcal{I}}_T'\right\}. \tag{65}$$

  Again, here some special assumptions may be of interest.

  - $\boldsymbol{\mathcal{I}}_T=\mathbf{I}$: Then the optimal training matrix can be found by straightforward adjustment of Proposition 2 in [8].
  - $\mathbf{R}_T^{-1}=\boldsymbol{\mathcal{I}}_T$: Then (65) takes the form

$$\mathrm{tr}\left\{\left(\mathbf{I}+\mathbf{R}_T'^{1/2}\mathbf{P}'^H\mathbf{S}'^{-1}_Q\mathbf{P}'\mathbf{R}_T'^{1/2}\right)^{-1}\right\}. \tag{66}$$

    Using the same majorization argument as in the previous Appendix for $\mathrm{tr}\left(\mathbf{A}^{-1}\right)=\sum_i 1/\lambda_i\left(\mathbf{A}\right)$, and adopting the notation therein, we should select $\mathbf{U}=\mathbf{U}_T'$ and $\mathbf{V}=\mathbf{V}_Q$. With these choices, the optimal power allocation problem becomes

$$\begin{array}{ll}\underset{\pi,\varpi,\kappa_i,i=1,2,\ldots,n_T}{\text{minimize}} & \sum_{i=1}^{n_T}\frac{1}{1+\frac{\pi\left(\left(\boldsymbol{\Lambda}_T'\right)_{i,i}\right)\kappa_i}{\varpi\left(\left(\boldsymbol{\Lambda}_Q\right)_{i,i}\right)}}\\ \text{s.t.} & \sum_{i=1}^{n_T}\kappa_i\leq\mathcal{P}\end{array} \tag{67}$$

    where $\left(\boldsymbol{\Lambda}_T'\right)_{i,i}, i=1,2,\ldots,n_T$ are the eigenvalues of $\mathbf{R}_T'$. Fixing the permutations $\pi(\cdot)$ and $\varpi(\cdot)$, we set $\gamma_i=\pi\left(\left(\boldsymbol{\Lambda}_T'\right)_{i,i}\right)/\varpi\left(\left(\boldsymbol{\Lambda}_Q\right)_{i,i}\right), i=1,2,\ldots,n_T$. With this notation, the problem of selecting the optimal κ_i's becomes

$$\begin{array}{ll}\underset{\kappa_i,i=1,2,\ldots,n_T}{\text{minimize}} & \sum_{i=1}^{n_T}\frac{1}{1+\gamma_i\kappa_i}\\ \text{s.t.} & \sum_{i=1}^{n_T}\kappa_i\leq\mathcal{P}.\end{array} \tag{68}$$

Following similar steps as in the proof of Proposition 2 in [8], we define the following parameter

$$m_* = \max\left\{m \in \{1,2,\ldots,n_T\} : \sqrt{\frac{1}{\gamma_k}} \cdot \sum_{i=1}^{m}\sqrt{\frac{1}{\gamma_i}} - \sum_{i=1}^{m}\frac{1}{\gamma_i} < \mathcal{P}, k = 1,2,\ldots,m\right\}. \quad (69)$$

Then, it can be easily seen that for $j = 1,2,\ldots,m_*$ the optimal $(\mathbf{D}_{P'})_{j,j}$ is given by the expression

$$\sqrt{\frac{\mathcal{P} + \sum_{i=1}^{m_*}\frac{1}{\gamma_i}}{\sum_{i=1}^{m_*}\sqrt{\frac{1}{\gamma_i}}}\sqrt{\frac{1}{\gamma_j}} - \frac{1}{\gamma_j}},$$

while $(\mathbf{D}_{P'})_{j,j} = 0$ for $j = m_* + 1,\ldots,n_T$.

With these expressions for the optimal power allocation, the objective of (67) equals

$$n_T - m_* + \frac{\left(\sum_{i=1}^{m_*}\frac{1}{\sqrt{\gamma_i}}\right)^2}{\mathcal{P} + \sum_{i=1}^{m_*}\frac{1}{\gamma_i}}$$

and therefore the problem of determining the optimal orderings $\pi(\cdot), \varpi(\cdot)$ becomes

$$\underset{\pi,\varpi}{\text{minimize}} \quad n_T - m_* + \frac{\left(\sum_{i=1}^{m_*}\frac{1}{\sqrt{\gamma_i}}\right)^2}{\mathcal{P} + \sum_{i=1}^{m_*}\frac{1}{\gamma_i}}. \quad (70)$$

The last problem seems to be difficult to solve analytically. Nevertheless, a simple numerical exhaustive search algorithm, namely Algorithm 1, can solve this problem[5]. Note that given the fact that n_T and B are small in practice, the complexity of the above algorithm and its necessary memory are not crucial. However, as n_T and B increase, complexity and memory become important. In this case, a good solution may be to order the eigenvalues of $\mathbf{R}'_T$ in decreasing order and those of $\mathbf{S}'_Q$ in increasing order. This can be analytically justified based on the fact that for a fixed m_*, the objective function of problem (70), say $\mathrm{MSE}(\gamma_1,\ldots,\gamma_{m_*})$, has negative partial derivatives with respect to $\gamma_i, i = 1,2,\ldots,m_*$ and it is also symmetric, since any permutation of its arguments does not change its value. This essentially shows that a good solution may maintain as active γ's the largest possible, through the selection of m_*. Additionally, the structure of $\mathrm{MSE}(\gamma_1,\ldots,\gamma_{m_*})$ reveals the fact that for every new active γ, something less than 1 is added to the MSE, while an inactive value corresponds to adding 1 to the MSE. This is intuitively appealing with the spatial diversity of MIMO systems and the usual properties that optimal training matrices possess in such systems (i.e., that they tend to fully exploit the available spatial diversity). The largest possible γ's can be achieved with a decreasing order of the eigenvalues of $\mathbf{R}'_T$ and an increasing order of the eigenvalues of $\mathbf{S}'_Q$. In this case, it can be checked that m_* can be found as follows

$$m_* = \max\left\{m \in \{1,2,\ldots,n_T\} : \sqrt{\frac{1}{\gamma_m}} \cdot \sum_{i=1}^{m}\sqrt{\frac{1}{\gamma_i}} - \sum_{i=1}^{m}\frac{1}{\gamma_i} < \mathcal{P}\right\}.$$

[5]For easiness, we use the MATLAB notation in this table.

Algorithm 1 Optimal ordering for the eigenvalues of $\mathbf{R}'_T$ and $\mathbf{S}'_Q$, when $\mathbf{R}_R = \mathbf{S}_R$ and $\mathbf{R}_T^{-1} = \boldsymbol{\mathcal{I}}_T$.

Require: n_T, B such that $B \geq n_T$, $\mathcal{P}$, a row vector λ'_T containing all $(\boldsymbol{\Lambda}'_T)_{i,i}$'s for $i = 1,2,\ldots,n_T$ in any order and a row vector λ_Q containing all $(\boldsymbol{\Lambda}_Q)_{i,i}$'s for $i = 1,2,\ldots,B$ in any order.

Create two matrices $\boldsymbol{\Pi}_T$ and $\boldsymbol{\Pi}_Q$ containing as rows all possible permutations of λ'_T and λ_Q, respectively. Define also the matrix $\boldsymbol{\Gamma} = [\ ]$.
loop
  for $l = 1 : n_T!$
  loop
    for $t = 1 : B!$
    $\boldsymbol{\Gamma} = [\boldsymbol{\Gamma}; \boldsymbol{\Pi}_T(l,:)./\boldsymbol{\Pi}_Q(t, 1:n_T)]$.
loop
  For each row of $\boldsymbol{\Gamma}$ determine the corresponding m_* and place it in the corresponding row of a new vector $\mathbf{M}$.
loop
  for $l = 1 : n_T!B!$

$$J(l) = n_T - M(l) + \frac{\left(\sum_{i=1}^{M(l)}\frac{1}{\sqrt{\boldsymbol{\Gamma}(l,i)}}\right)^2}{\mathcal{P} + \sum_{i=1}^{M(l)}\frac{1}{\boldsymbol{\Gamma}(l,i)}}$$

$[\mathrm{val}, \mathrm{ind}] = \min J$
if $\mathrm{mod}(\mathrm{ind}, B!) == 0$ **then**
  $j = B!$
else
  $j = \mathrm{mod}(\mathrm{ind}, B!)$
$i = (\mathrm{ind} - j)/B! + 1$
The optimal $\pi(\cdot)$, say π_{opt}, corresponds to $\boldsymbol{\Pi}_T(i,:)$ and the optimal $\varpi(\cdot)$, say ϖ_{opt}, to $\boldsymbol{\Pi}_Q(j,:)$.

- If the modal matrices of $\mathbf{R}_R$ and $\mathbf{S}_R$ are the same, $\boldsymbol{\mathcal{I}}_T = \mathbf{I}$ and $\boldsymbol{\mathcal{I}}_R = \mathbf{I}$, then the optimal training is given by [9], as these assumptions correspond to the problem solved therein.
- In any other case (e.g., if $\mathbf{R}_R \neq \mathbf{S_R}$), the (optimal) training can be found using numerical methods like the semidefinite relaxation approach described in [28]. Note that this approach can handle also general $\boldsymbol{\mathcal{I}}_{adm}$, not necessarily Kronecker-structured.

REFERENCES

[1] V. Tarokh, A. Naguib, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: performance criteria in the presence of channel estimation errors, mobility, and multiple paths," *IEEE Trans. Commun.*, vol. 47, no. 2, pp. 199–207, Feb. 1999.

[2] P. Stoica and O. Besson, "Training sequence design for frequency offset and frequency-selective channel estimation," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1910–1917, Nov. 2003.

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

[4] J. Kotecha and A. Sayeed, "Transmit signal design for optimal estimation of correlated MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 546–557, Feb. 2004.

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

[6] M. Biguesh and A. 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.

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

[8] D. Katselis, E. Kofidis, and S. Theodoridis, "Training-based estimation of correlated MIMO fading channels in the presence of colored interference," *Signal Processing*, vol. 87, no. 9, pp. 2177–2187, Sep. 2007.

[9] 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, Mar. 2010.

[10] M. Biguesh, S. Gazor, and M. Shariat, "Optimal training sequence for MIMO wireless systems in colored environments," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3144–3153, Aug. 2009.

[11] H. Vikalo, B. Hassibi, B. Hochwald, and T. Kailath, "On the capacity of frequency- selective channels in training-based transmission schemes," *IEEE Trans. Signal Process.*, vol. 52, no. 9, pp. 2572–2583, Sep. 2004.

[12] K. Ahmed, C. Tepedelenlioglu, and A. Spanias, "Pep-based optimal training for MIMO systems in wireless channels," in *Proc. IEEE ICASSP*, vol. 3, Mar. 2005, pp. 793–796.

[13] H. Jansson and H. Hjalmarsson, "Input design via LMIs admitting frequency-wise model specifications in confidence regions," *IEEE Trans. Autom. Control*, vol. 50, no. 10, pp. 1534–1549, 2005.

[14] X. Bombois, G. Scorletti, M. Gevers, P. M. J. Van den Hof, and R. Hildebrand, "Least costly identification experiment for control," *Automatica*, vol. 42, no. 10, pp. 1651–1662, 2006.

[15] H. Hjalmarsson, "System identification of complex and structured systems," *Plenary address European Control Conference / European Journal of Control*, vol. 15, no. 4, pp. 275–310, 2009.

[16] C. R. Rojas, J. C. Agüero, J. S. Welsh, and G. C. Goodwin, "On the equivalence of least costly and traditional experiment design for control," *Automatica*, vol. 44, no. 11, pp. 2706–2715, 2008.

[17] J. Kiefer, "General equivalence theory for optimum designs (approximate theory)," *Annals of Statistics*, vol. 2(5), pp. 849–879, 1974.

[18] P. Ciblat, P. Bianchi, and M. Ghogho, "Training sequence optimization for joint channel and frequency offset estimation," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3424–3436, Aug. 2008.

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

[20] K. Yu, M. Bengtsson, B. Ottersten, D. McNamara, P. Karlsson, and M. Beach, "Modeling of wideband MIMO radio channels based on NLOS indoor measurements," *IEEE Trans. Veh. Technol.*, vol. 53, no. 3, pp. 655–665, May 2004.

[21] S. Gazor and H. Rad, "Space-time frequency characterization of MIMO wireless channels," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2369–2376, Sep. 2006.

[22] H. Rad and S. Gazor, "The impact of non-isotropic scattering and directional antennas on MIMO multicarrier mobile communication channels," *IEEE Trans. Commun.*, vol. 56, no. 4, pp. 642–652, Apr. 2008.

[23] K. Werner and M. Jansson, "Estimating MIMO channel covariances from training data under the Kronecker model," *Signal Processing*, vol. 89, no. 1, pp. 1–13, Jan. 2009.

[24] S. Kay, *Fundamentals of Statistical Signal Processing. Estimation Theory*. Englewood Cliffs, New Jersey: Prentice-Hall, 1993.

[25] M. Barenthin and H. Hjalmarsson, "Identication and control: Joint input design and H_∞ state feedback with ellipsoidal parametric uncertainty via LMIs," *Automatica*, vol. 44, no. 2, pp. 543–551, 2008.

[26] X. Bombois and H. Hjalmarsson, "Optimal input design for robust H_2 deconvolution filtering," in *15th IFAC Symposium on System Identification*, Saint-Malo, France, July 2009.

[27] C. R. Rojas, D. Katselis, H. Hjalmarsson, R. Hildebrand, and M. Bengtsson, "Chance constrained input design," in *Proceedings of CDC-ECC*, Orlando, Florida, USA, December 2011.

[28] D. Katselis, C. R. Rojas, H. Hjalmarsson, and M. Bengtsson, "Application-oriented finite sample experiment design: A semidefinite relaxation approach," in *SYSID 2012*, Brussels, Belgium, July 2012, invited.

[29] L. Gerencsér, H. Hjalmarsson, and J. Mårtensson, "Adaptive input design for ARX systems," in *European Control Conference*, Kos, Greece, July 2-5 2007.

[30] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*. Cambridge, United Kingdom: Cambridge University Press, 2003.

[31] S. Verdú, *Multiuser Detection*. Cambridge, UK: Cambridge University Press, 1998.

[32] S. Haykin, *Adaptive Filter Theory*, 4th ed. Prentice-Hall, 2001.

[33] B. Hochwald, C. B. Peel, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication — part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

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

[35] A. Ostrowski, "A quantitative formulation of Sylvester's law of inertia, II," *Proc. National Academy of Sciences of the United States of America*, vol. 46, no. 6, pp. 859–862, Mar. 1960.