KTH Electrical Engineering

A Framework for Training-Based Estimation in Arbitrarily Correlated Rician MIMO Channels with Rician Disturbance

IEEE TRANSACTIONS ON SIGNAL PROCESSING
Volume 58, Issue 3, Pages 1807-1820, March 2010.

EMIL BJÖRNSON AND BJÖRN OTTERSTEN

Stockholm 2010

KTH Royal Institute of Technology
ACCESS Linnaeus Center
Signal Processing Lab

DOI: 10.1109/TSP.2009.2037352
KTH Report: IR-EE-SB 2010:011

A Framework for Training-Based Estimation in Arbitrarily Correlated Rician MIMO Channels With Rician Disturbance

Emil Björnson, *Student Member, IEEE*, and Björn Ottersten, *Fellow, IEEE*

***Abstract*—In this paper, we create a framework for training-based channel estimation under different channel and interference statistics. The minimum mean square error (MMSE) estimator for channel matrix estimation in Rician fading multi-antenna systems is analyzed, and especially the design of mean square error (MSE) minimizing training sequences. By considering Kronecker-structured systems with a combination of noise and interference and arbitrary training sequence length, we collect and generalize several previous results in the framework. We clarify the conditions for achieving the optimal training sequence structure and show when the spatial training power allocation can be solved explicitly. We also prove that spatial correlation improves the estimation performance and establish how it determines the optimal training sequence length. The analytic results for Kronecker-structured systems are used to derive a heuristic training sequence under general unstructured statistics.**

The MMSE estimator of the squared Frobenius norm of the channel matrix is also derived and shown to provide far better gain estimates than other approaches. It is shown under which conditions training sequences that minimize the non-convex MSE can be derived explicitly or with low complexity. Numerical examples are used to evaluate the performance of the two estimators for different training sequences and system statistics. We also illustrate how the optimal length of the training sequence often can be shorter than the number of transmit antennas.

***Index Terms*—Arbitrary correlation, channel matrix estimation, majorization, MIMO systems, MMSE estimation, norm estimation, Rician fading, training sequence optimization.**

I. Introduction

Wireless communication systems with antenna arrays at both the transmitter and the receiver have gained much attention due to their potential of greatly improving the performance over single-antenna systems. In flat fading systems, the capacity and spectral efficiency have been shown to increase rapidly with the number of antennas [1], [2]. These results are based on the idealized assumption of full channel state information (CSI) and independent and identically distributed (i.i.d.) channel coefficients. In practice, field measurements have shown that the channel coefficients often are spatially correlated in outdoor scenarios [3], but correlation also frequently occurs in indoor environments [4], [5]. When it comes to acquiring CSI, the long-term statistics can usually be regarded as known, through reverse-link estimation or a negligible signaling overhead [6]. Instantaneous CSI needs however to be estimated with limited resources (time and power) due to the channel fading and interference.

In this paper, we consider training-based estimation of instantaneous CSI in multiple-input multiple-output (MIMO) systems. Thus, the estimation is conditioned on the received signal from a known training sequence, which potentially can be adapted to the long-term statistics. By nature, the channel is stochastic, which motivates Bayesian estimation—that is, modeling of the current channel state as a realization from a known multi-variate probability density function (PDF). There is also a large amount of literature on estimation of deterministic MIMO channels which are analytically tractable but in general provide less accurate channel estimates, as shown in [7], [8]. Herein, we concentrate on minimum mean square error (MMSE) estimation of the channel matrix and its squared Frobenius norm, given the first and second order system statistics.

Training-based MMSE estimation of MIMO channel matrices has previously been considered for Kronecker-structured Rayleigh fading systems that are either noise-limited [9]–[11] or interference-limited [12]. In these papers, optimization of the training sequence was considered under various limitations on the long-term statistics, and analogous structures of the optimal training sequence were derived. These results reduce the training optimization to a convex power allocation problem that can be solved explicitly in some special cases. When mentioning previous work, it is worth noting that simplified channel matrix estimators have been developed in [8] and [13] and claimed to be MMSE estimators, but we show herein that these estimators are in general restrictive.

In the present work, we collect previous results in a framework with general system properties and arbitrary length of the training sequence. The MMSE estimator is given for Kronecker-structured Rician fading channels that are corrupted by some Gaussian *disturbance*, where disturbance denotes a combination of noise and interference. The purpose of our frame-

Manuscript received September 21, 2009; accepted October 25, 2009. First published November 24, 2009; current version published February 10, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Amir Leshem. This work was supported in part by the ERC under FP7 Grant Agreement No. 228044 and the FP6 project Cooperative and Opportunistic Communications in Wireless Networks (COOPCOM), Project No. FP6-033533. This work was also partly performed in the framework of the CELTIC project CP5-026 WINNER+. Parts of this work were previously presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, Apr.19–24, 2009.

E. Björnson is with the Signal Processing Laboratory, ACCESS Linnaeus Center, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden (e-mail: emil.bjornson@ee.kth.se).

B. Ottersten is with the Signal Processing Laboratory, ACCESS Linnaeus Center, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden, and also with the securityandtrust.lu, University of Luxembourg, L-1359 Luxembourg-Kirchberg, Luxembourg (e-mail: bjorn.ottersten@ee.kth.se).

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

Digital Object Identifier 10.1109/TSP.2009.2037352

work is to enable joint analysis of different types of disturbance, including the noise-limited and interference-limited scenarios considered in [9]–[12] and certain combinations of both noise and interference. In this manner, we show that the MSE minimizing training sequence has the same structure and asymptotic properties under a wide range of different disturbance statistics. We give statistical conditions for finding the optimal training sequence explicitly, and propose a heuristic solution under general unstructured statistics. Finally, we prove analytically that the MSE decreases with increasing spatial correlation at both the transmitter and the receiver side. Based on this observation, we show that the optimal number of training symbols can be considerably fewer than the number of transmit antennas in correlated systems. This result is a generalization of [14], where completely uncorrelated systems were considered, and similar observations have been made in [15], [16].

Although estimation of the channel matrix is important for receive and transmit processing, knowledge of the squared Frobenius norm of the channel matrix provides instantaneous gain information and can be exploited for rate adaptation and scheduling [17], [18]. The squared norm can be determined indirectly from an estimated channel matrix, but as shown in [16] this approach gives poor estimation performance at most signal-to-interference-and-noise ratios (SINRs). The MMSE estimator of the squared channel norm was introduced in [16] for Kronecker-structured Rayleigh fading channels, assuming the same training structure as for channel matrix estimation. Herein, the estimator is proved and generalized to Rician fading channels, along with the design of MSE minimizing training sequences. Although the MSE is non-convex, we show that the optimal training sequence can be determined with limited complexity.

A. Outline

In Section II, the system model and the training-based estimation framework is introduced. The MMSE channel matrix estimator is given and discussed in Section III for arbitrary training sequences. In Section IV, MSE minimizing training sequence design is considered. The general structure and asymptotic properties are derived. It is also shown under which covariance conditions there exist explicit solutions, and how the estimation performance and the optimal length of the training sequence varies with the spatial correlation. Section V derives the MMSE estimator of the squared channel norm and analyzes training sequence design with respect to its MSE. The error performance of the different estimators are illustrated numerically in Section VI and conclusions are drawn in Section VII. Finally, proofs of the theorems are given in Appendix A.

B. Notations

Boldface (lower case) is used for column vectors, $\mathbf{x}$, and (upper case) for matrices, $\mathbf{X}$. Let $\mathbf{X}^T$, $\mathbf{X}^H$, and $\mathbf{X}^*$ denote the transpose, the conjugate transpose, and the conjugate of $\mathbf{X}$, respectively. The Kronecker product of two matrices $\mathbf{X}$ and $\mathbf{Y}$ is denoted $\mathbf{X} \otimes \mathbf{Y}$, $\mathrm{vec}(\mathbf{X})$ is the column vector obtained by stacking the columns of $\mathbf{X}$, $\mathrm{tr}\{\mathbf{X}\}$ is the matrix trace, and $\mathrm{diag}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is the N-by-N diagonal matrix with $\mathbf{x}_1, \ldots, \mathbf{x}_N$ at the main diagonal. The squared Frobenius norm of a matrix $\mathbf{X}$ is denoted $||\mathbf{X}||^2$ and is defined as the sum of the squared absolute values of all the elements. The functions $\max(x, y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $\min(x, y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ give the maximal and minimal value of the input parameters, respectively. $\mathcal{CN}(\bar{\mathbf{x}}, \mathbf{R})$ is used to denote circularly symmetric complex Gaussian random vectors, where $\bar{\mathbf{x}}$ is the mean and $\mathbf{R}$ the covariance matrix. The notation $\triangleq$ is used for definitions.

II. System Model

We consider flat and block-fading MIMO systems with a transmitter equipped with an array of n_T transmit antennas and a receiver with an array of n_R receive antennas. The symbol-sampled complex baseband equivalent of the flat fading channel when transmitting at channel use t is modeled as

$$\mathbf{y}(t) = \mathbf{H}\mathbf{x}(t) + \mathbf{n}(t) \tag{1}$$

where $\mathbf{x}(t) \in \mathbb{C}^{n_T}$ and $\mathbf{y}(t) \in \mathbb{C}^{n_R}$ are the transmitted and received signals, respectively, and $\mathbf{n}(t) \in \mathbb{C}^{n_R}$ represents arbitrarily correlated Gaussian disturbance. This disturbance models the sum of background noise and interference from adjacent communication links and is a stochastic process in t. The channel is represented by $\mathbf{H} \in \mathbb{C}^{n_R \times n_T}$ and is modeled as Rician fading with mean $\bar{\mathbf{H}} \in \mathbb{C}^{n_T \times n_R}$ and the positive definite covariance matrix $\mathbf{R} \in \mathbb{C}^{n_T n_R \times n_T n_R}$, which is defined on the column stacking of the channel matrix. Thus, $\mathrm{vec}(\mathbf{H}) \in \mathcal{CN}(\mathrm{vec}(\bar{\mathbf{H}}), \mathbf{R})$. In the estimation parts of this paper, the channel and disturbance statistics are known at the receiver. In the training sequence design, the statistics are also known to the transmitter.

Herein, estimation of the channel matrix $\mathbf{H}$ and its squared Frobenius norm $||\mathbf{H}||^2$ are considered. The receiver knows the long-term statistics, but in order to estimate the value of some function of the unknown realization of $\mathbf{H}$, the transmitter typically needs to send a sequence of known training vectors that spans $\mathbb{C}^{n_T}$. We consider training sequences of arbitrary length B under a total power constraint, and in Section IV-A the optimal value of B is studied.

Let the training matrix $\mathbf{P} \in \mathbb{C}^{n_T \times B}$ represent the training sequence. This matrix fulfills the total power constraint $\mathrm{tr}(\mathbf{P}^H\mathbf{P}) = \mathcal{P}$ and its maximal rank is $m \triangleq \min(n_T, B)$, which represents the maximal number of spatial channel directions that the training can excite. The columns of $\mathbf{P}$ are used as transmit signal in (1) for B channel uses (e.g., $t = 1, \ldots, B$). The combined received matrix $\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(B)] \in \mathbb{C}^{n_R \times B}$ of the training transmission is

$$\mathbf{Y} = \mathbf{H}\mathbf{P} + \mathbf{N} \tag{2}$$

where the combined disturbance matrix $\mathbf{N} = [\mathbf{n}(1), \ldots, \mathbf{n}(B)] \in \mathbb{C}^{n_R \times B}$ is uncorrelated with the channel $\mathbf{H}$. The disturbance is modeled as $\mathrm{vec}(\mathbf{N}) \in \mathcal{CN}(\mathrm{vec}(\bar{\mathbf{N}}), \mathbf{S})$, where $\mathbf{S} \in \mathbb{C}^{Bn_R \times Bn_R}$ is the positive definite covariance matrix and $\bar{\mathbf{N}} \in \mathbb{C}^{n_R \times B}$ is the mean disturbance.

The multipath propagation is modeled as quasi-static block fading; that is, the channel realization $\mathbf{H}$ is constant during

the whole training transmission and independent of previous channel estimates.

A. Preliminaries on Spatial Correlation and Majorization

A measure of the spatial channel correlation is the eigenvalue distribution of the channel covariance matrix; weak correlation is represented by almost identical eigenvalues, while strong correlation means that a few eigenvalues dominate. Thus, in a highly correlated system, the channel is approximately confined to a small eigensubspace, while all eigenvectors are equally important in an uncorrelated system. In urban cellular systems, base stations are typically elevated and exposed to little near-field scattering. Thus, their antennas are strongly spatially correlated, while the non-line-of-sight mobile users are exposed to rich scattering and have weak antenna correlation if the antenna spacing is sufficiently large [19].

The notion of majorization provides a useful measure of the spatial correlation [20]–[22] and will be used herein for various purposes. Let $\mathbf{x} = [x_1, \ldots, x_M]^T$ and $\mathbf{y} = [y_1, \ldots, y_M]^T$ be two non-negative real-valued vectors of arbitrary length M. We say that $\mathbf{x}$ majorizes $\mathbf{y}$ if

$$\sum_{k=1}^{m} x_{[k]} \geq \sum_{k=1}^{m} y_{[k]}, \text{ for } m = 1, \ldots, M-1, \quad \text{and} \sum_{k=1}^{M} x_k = \sum_{k=1}^{M} y_k \tag{3}$$

where $x_{[k]}$ and $y_{[k]}$ are the kth largest ordered elements of $\mathbf{x}$ and $\mathbf{y}$, respectively. This majorization property is denoted $\mathbf{x} \succeq \mathbf{y}$. If $\mathbf{x}$ and $\mathbf{y}$ contain eigenvalues of channel covariance matrices, then $\mathbf{x} \succeq \mathbf{y}$ corresponds to that $\mathbf{x}$ is more spatially correlated than $\mathbf{y}$. Majorization only provides a partial order of vectors, but is still very powerful due to its connection to certain order-preserving functions:

A function $f(\cdot) : \mathbb{R}^M \rightarrow \mathbb{R}$ is said to be Schur-convex if $f(\mathbf{x}) \geq f(\mathbf{y})$ for all $\mathbf{x}$ and $\mathbf{y}$, such that $\mathbf{x} \succeq \mathbf{y}$. Similarly, $f(\cdot)$ is said to be Schur-concave if $\mathbf{x} \succeq \mathbf{y}$ implies that $f(\mathbf{x}) \leq f(\mathbf{y})$.

III. MMSE Estimation of Channel Matrices

There are many reasons for estimating the channel matrix $\mathbf{H}$ at the receiver. Instantaneous CSI can, for example, be used for receive processing (improved interference suppression and simplified detection) and feedback (to employ beamforming and rate adaptation). In this section, we consider MMSE estimation of the channel matrix from the observation during training transmission. In general, the MMSE estimator of a vector $\mathbf{h}$ from an observation $\mathbf{y}$ is

$$\hat{\mathbf{h}}_{\mathrm{MMSE}} = \mathbb{E}\{\mathbf{h}|\mathbf{y}\} = \int \mathbf{h} f(\mathbf{h}|\mathbf{y}) d\mathbf{h} \tag{4}$$

where $\mathbb{E}\{\cdot\}$ denotes the expected value and $f(\mathbf{h}|\mathbf{y})$ is the conditional (posterior) PDF of $\mathbf{h}$ given $\mathbf{y}$ [23, Section 11.4]. The MMSE estimator minimizes the MSE, $\mathbb{E}\{\|\mathbf{h} - \hat{\mathbf{h}}_{\mathrm{MMSE}}\|^2\}$, and the optimal MSE can be calculated as the trace of the covariance matrix $\mathbf{C}_{\mathrm{MMSE}}$ of $f(\mathbf{h}|\mathbf{y})$ averaged over $\mathbf{y}$. The MMSE estimator is the Bayesian counterpart to the minimum variance unbiased (MVU) estimator developed for deterministic channels [23, Section 3.4].

By vectorizing the received signal in (2) and applying $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\mathrm{vec}(\mathbf{B})$, the received training signal of our system can be expressed as

$$\mathrm{vec}(\mathbf{Y}) = \widetilde{\mathbf{P}}\mathrm{vec}(\mathbf{H}) + \mathrm{vec}(\mathbf{N}) \tag{5}$$

where $\widetilde{\mathbf{P}} \triangleq (\mathbf{P}^T \otimes \mathbf{I})$. Then, by pre-subtracting the mean disturbance $\mathrm{vec}(\bar{\mathbf{N}})$ from $\mathrm{vec}(\mathbf{Y})$, it is straightforward to apply the results of [23, Chapter 15.8] to conclude that the MMSE estimator, $\hat{\mathbf{H}}_{\mathrm{MMSE}}$, of the Rician fading channel matrix $\mathbf{H}$ is

$$\begin{aligned} \mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{MMSE}}) &= \mathrm{vec}(\bar{\mathbf{H}}) + (\mathbf{R}^{-1} + \widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \widetilde{\mathbf{P}})^{-1} \widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \mathbf{d} \\ &= \mathrm{vec}(\bar{\mathbf{H}}) + \mathbf{R}\widetilde{\mathbf{P}}^H (\widetilde{\mathbf{P}}\mathbf{R}\widetilde{\mathbf{P}}^H + \mathbf{S})^{-1} \mathbf{d} \end{aligned} \tag{6}$$

where $\mathbf{d} = \mathrm{vec}(\mathbf{Y}) - \widetilde{\mathbf{P}}\mathrm{vec}(\bar{\mathbf{H}}) - \mathrm{vec}(\bar{\mathbf{N}})$. The error covariance $\mathbf{C}_{\mathrm{MMSE}} \triangleq \mathbb{E}\{(\mathrm{vec}(\mathbf{H}) - \mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{MMSE}}))(\mathrm{vec}(\mathbf{H}) - \mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{MMSE}}))^H\}$ becomes

$$\begin{aligned} \mathbf{C}_{\mathrm{MMSE}} &= (\mathbf{R}^{-1} + \widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \widetilde{\mathbf{P}})^{-1} \\ &= \mathbf{R} - \mathbf{R}\widetilde{\mathbf{P}}^H (\widetilde{\mathbf{P}}\mathbf{R}\widetilde{\mathbf{P}}^H + \mathbf{S})^{-1} \widetilde{\mathbf{P}}\mathbf{R} \end{aligned} \tag{7}$$

and the MSE $\triangleq \mathbb{E}\{\|\mathrm{vec}(\mathbf{H}) - \mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{MMSE}})\|^2\}$ is

$$\begin{aligned} \mathrm{MSE} &= \mathrm{tr}\left\{(\mathbf{R}^{-1} + \widetilde{\mathbf{P}}^H \mathbf{S}^{-1} \widetilde{\mathbf{P}})^{-1}\right\} \\ &= \mathrm{tr}\left\{\mathbf{R} - \mathbf{R}\widetilde{\mathbf{P}}^H (\widetilde{\mathbf{P}}\mathbf{R}\widetilde{\mathbf{P}}^H + \mathbf{S})^{-1} \widetilde{\mathbf{P}}\mathbf{R}\right\}. \end{aligned} \tag{8}$$

We stress that the *general* MMSE estimator in (6) is in fact linear (affine), but nonetheless it has repeatedly been referred to as the *linear* MMSE (LMMSE) estimator [10]–[12] which is correct but could lead to the incorrect conclusion that there may exist better non-linear estimators. The MMSE estimator in (6) is also the maximum *a posteriori* (MAP) estimator of $\mathbf{H}$ [23, Chapter 15.8] and the LMMSE estimator in the case of non-Gaussian fading and disturbance (with known first and second order statistics, independent fading and disturbance, and possibly unknown types of distributions [23, Chapter 12.3]).

Note that the computation of (6) only requires a multiplication of $\mathrm{vec}(\mathbf{Y})$ with a matrix and adding a vector, both of which depend only on the system statistics. Thus, the computational complexity of the estimator is limited.

Remark 1: For Rayleigh fading channels, the MMSE estimator in (6) has the general linear form $\mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{MMSE}}) = \mathbf{A}\mathrm{vec}(\mathbf{Y})$. A special kind of linear estimators with the alternative structure $\hat{\mathbf{H}} = \mathbf{Y}\mathbf{A}_o$ were studied in [8] and [13] and claimed to give rise to LMMSE estimators. In general, this claim is incorrect, which is seen by vectorizing the estimate; $\mathrm{vec}(\hat{\mathbf{H}}) = (\mathbf{A}_o^T \otimes \mathbf{I})\mathrm{vec}(\mathbf{Y})$ and thus the estimators in [8] and [13] belong to a subset of linear estimators with $\mathbf{A} = (\mathbf{A}_o^T \otimes \mathbf{I})$. The general MMSE estimator belongs to this subset when applied to Kronecker-structured systems with identical receive channel and disturbance covariance matrices,[1] while the difference between $\mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{MMSE}})$ and $\mathrm{vec}(\hat{\mathbf{H}})$ increases with the difference in receive-side correlation and how far from Kronecker-structured the statistics are.

[1]In this special case, the estimation of each row of $\mathbf{H}$ can be separated into independent problems with identical statistics.

IV. Training Sequence Optimization for Channel Matrix Estimation

Next, we consider the problem of designing the training sequence $\mathbf{P}$ to optimize the performance of the MMSE estimator in (6). The performance measure is the MSE and thus from (8) the optimization problem can be formulated as

$$\min_{\mathbf{P}} \operatorname{tr}\left\{\left(\mathbf{R}^{-1}+(\mathbf{P}^T\otimes\mathbf{I})^H\mathbf{S}^{-1}(\mathbf{P}^T\otimes\mathbf{I})\right)^{-1}\right\} \quad \text{subject to} \quad \operatorname{tr}(\mathbf{P}^H\mathbf{P})\leq\mathcal{P}. \tag{9}$$

Observe that the MSE depends on the training matrix $\mathbf{P}$ and on the covariance matrices of the channel and disturbance statistics, while it is unaffected by the mean values. Thus, the training matrix can potentially be designed to optimize the performance by adaptation to the second order statistics [9]–[12]. The intuition behind this training optimization is that more power should be allocated to estimate the channel in strong eigendirections (i.e., large eigenvalues). Observe that training optimization is useful in systems with dedicated training for each receiver, while multiuser systems with common training may require fixed or codebook-based training matrices (if users do not have the same channel statistics).

For general channel and disturbance statistics, the MSE minimizing training matrix will not have any special form that can be exploited when solving (9). However, if the covariance matrices $\mathbf{R}$ and $\mathbf{S}$ are structured, the optimal $\mathbf{P}$ may inherit this structure. Previous work in training optimization has showed that in Kronecker-structured systems with either noise-limited [9]–[11] or interference-limited [12] disturbance, the optimal training matrix has a certain structure based on the transmit-side channel covariance and temporal disturbance covariance. Herein, this result is generalized by showing that the same optimal structure appears in systems with *both* noise and interference. Then, we will show how the training matrix behaves asymptotically and under which conditions there exist explicit solutions to (9). Finally, we analyze how the statistics and total training power determines the smallest length of the training sequence necessary to achieve the minimal MSE.

Since the training matrix $\mathbf{P}$ only affects the channel matrix, $\mathbf{H}$, from the right hand (transmit) side in (2), we consider covariance matrices that also can be separated between the transmit and receive side. Thus, the covariance between the transmit antennas is identical irrespectively of where the receiver is located, and *vice versa* [24]. This model is known as the Kronecker-structure and is naturally applicable in uncorrelated systems. In practice, for example insufficient antenna spacing leads to antenna correlation, but field measurements have verified the Kronecker-structure for certain correlated channels [3], [4]. In general, certain weak scattering scenarios can be created and observed where the Kronecker-structure is not satisfied [25], and thus the Kronecker model should be seen as a good approximation that enables analysis. We will show numerically in Section VI that training sequences optimized based on this approximation perform well when applied for estimation under general conditions. In our context, we define Kronecker-structured systems in the following way.

Definition 1: In a *Kronecker-structured system*, the channel covariance, $\mathbf{R}$, and disturbance covariance matrix, $\mathbf{S}$, can be factorized as

$$\mathbf{R}=\mathbf{R}_T^T\otimes\mathbf{R}_R, \quad \mathbf{S}=\mathbf{S}_Q^T\otimes\mathbf{S}_R. \tag{10}$$

Here, $\mathbf{R}_T\in\mathbb{C}^{n_T\times n_T}$ and $\mathbf{R}_R\in\mathbb{C}^{n_R\times n_R}$ represent the spatial covariance matrices at the transmitter and receiver side, respectively, while $\mathbf{S}_Q\in\mathbb{C}^{B\times B}$ and $\mathbf{S}_R\in\mathbb{C}^{n_R\times n_R}$ represent the temporal covariance matrix and the received spatial covariance matrix.

We also assume that $\mathbf{R}_R$ and $\mathbf{S}_R$ have identical eigenvectors. This means that the disturbance is either spatially uncorrelated or shares the spatial structure of the channel (i.e., arriving from the same spatial direction). This assumption was first made in [12] for estimation of interference-limited systems. Under this assumption, we can jointly describe several types of disturbance, including the following examples:

- **Noise-limited**, $\mathbf{S}=\mu\mathbf{I}$ with some variance μ;
- **Interference-limited**, $\mathbf{S}=(\sum_{j\in\mathcal{S}}\mathbf{Q}_j)\otimes\mathbf{R}_R$ for a set of interferers with temporal covariance $\mathbf{Q}_j$;[2]
- **Noise and temporally uncorrelated interference**, $\mathbf{S}=\mu\mathbf{I}+(\sum_{j\in\mathcal{S}}\mathbf{I})\otimes\mathbf{R}_R=\mathbf{I}\otimes(\mu\mathbf{I}+|\mathcal{S}|\mathbf{R}_R)$;
- **Noise and spatially uncorrelated interference**, $\mathbf{S}=\mu\mathbf{I}+(\sum_{j\in\mathcal{S}}\mathbf{Q}_j)\otimes\mathbf{I}=(\mu\mathbf{I}+\sum_{j\in\mathcal{S}}\mathbf{Q}_j)\otimes\mathbf{I}$.

To simplify the notation, we will use the following eigenvalue decompositions:

$$\mathbf{R}_T=\mathbf{U}_T\mathbf{\Lambda}_T\mathbf{U}_T^H, \quad \mathbf{R}_R=\mathbf{U}_R\mathbf{\Lambda}_R\mathbf{U}_R^H, \tag{11}$$

$$\mathbf{S}_Q=\mathbf{V}_Q\mathbf{\Sigma}_Q\mathbf{V}_Q^H, \quad \mathbf{S}_R=\mathbf{U}_R\mathbf{\Sigma}_R\mathbf{U}_R^H, \tag{12}$$

where the eigenvalues of $\mathbf{\Lambda}_T=\operatorname{diag}(\lambda_1^{(T)},\ldots,\lambda_{n_T}^{(T)})$ and $\mathbf{\Sigma}_Q=\operatorname{diag}(\sigma_1^{(Q)},\ldots,\sigma_{n_T}^{(Q)})$ are ordered in *decreasingly* and *increasingly*, respectively. The diagonal eigenvalue matrices $\mathbf{\Lambda}_R=\operatorname{diag}(\lambda_1^{(R)},\ldots,\lambda_{n_R}^{(R)})$, and $\mathbf{\Sigma}_R=\operatorname{diag}(\sigma_1^{(R)},\ldots,\sigma_{n_R}^{(R)})$ are arbitrarily ordered.

Next, we provide a theorem that derives the general structure of the MSE minimizing training sequence, along with its asymptotic properties.

Theorem 1: Under the Kronecker-structured assumptions, the solution to (9) has the singular value decomposition $\mathbf{P}=\mathbf{U}_T\mathbf{D}\mathbf{V}_Q^H$, where $\mathbf{D}\in\mathbb{C}^{n_T\times B}$ has $\sqrt{p_1},\ldots,\sqrt{p_m}$ on its main diagonal. The MSE with such a training matrix is convex with respect to the positive training powers $p_1,\ldots,p_m$, and the training powers should be ordered such that $p_j/\sigma_j^{(Q)}$ decreases with j (i.e., in the same order as $\lambda_j^{(T)}$). The MSE minimizing power allocation, $p_1,\ldots,p_m$, is achieved from the following system of equations:

$$\alpha=\sum_{l=1}^{n_R}\frac{\left(\lambda_j^{(T)}\lambda_l^{(R)}\right)^2\sigma_j^{(Q)}\sigma_l^{(R)}}{\left(\sigma_j^{(Q)}\sigma_l^{(R)}+p_j\lambda_j^{(T)}\lambda_l^{(R)}\right)^2} \tag{13}$$

[2]It worth noting that since a flat and block fading channel model was assumed in (1), the potential temporal covariance in $\mathbf{Q}_j$ primarily originate from the interfering signals and not from their channels. Also observe that if $\mathbf{R}_R\neq\mathbf{I}$, the interference will be received from the same spatial direction as the training signal.

for all j such that $\alpha < \sum_{j=1}^{m} \left(\lambda_j^{(T)}\lambda_l^{(R)}\right)^2/(\sigma_j^{(Q)}\sigma_l^{(R)})$ and $p_j = 0$ otherwise. The Lagrange multiplier $\alpha > 0$ is chosen to fulfill the constraint $\sum_{j=1}^{m} p_j = \mathcal{P}$.

The limiting training matrix at high power $\mathcal{P}$ is given by $p_j = \mathcal{P}\sqrt{\sigma_j^{(Q)}}/C$ for all j, where $C = \sum_{i=1}^{m}\sqrt{\sigma_i^{(Q)}}$. At low power $\mathcal{P}$, let $\tilde{m}$ be the minimum of the multiplicities of the largest $\lambda_j^{(T)}$ and the smallest $\sigma_j^{(Q)}$. Then, the limiting training matrix is given by allocating all power in an arbitrary manner among $p_1, \ldots, p_{\tilde{m}}$, while $p_j = 0$ for $j > \tilde{m}$.

Proof: The proof is given in Appendix A. ∎

The theorem showed that the MSE minimizing training matrix $\mathbf{P}$ in Kronecker-structured systems has a special structure based on the eigenvectors of the channel at the transmitter side and the temporal disturbance; the jth strongest channel eigendirection is assigned to the jth weakest disturbance eigendirection (i.e., in opposite order of magnitude). In other words, the strongest channel direction is estimated when the disturbance is as weak as possible (and vice versa). This was proved in [12] for interference-limited systems, and Theorem 1 generalizes it to cover various combinations of noise and interference.

At high training power, the power should be allocated to the m statistically strongest eigendirections of the channel, and proportionally to the square root of the m weakest eigendirections of the disturbance. At low training power, all power should be allocated in a single direction where a certain combination of strong channel gain and weak disturbance is maximized. These asymptotic results unify previous results, including the special cases of uncorrelated noise [9], [11] and single-antenna receivers [26].

Although the structure of the MSE minimizing training sequence is given in Theorem 1, the solution to the remaining power allocation problem is in general unknown. Since the problem is convex, the solution can however be derived with limited computational effort. The following corollary summarize results on when the power allocation can be solved explicitly.

Corollary 1: If $\mathbf{\Lambda}_T = \mathbf{I}$ and $\mathbf{\Sigma}_T = \mathbf{I}$, then equal power allocation ($p_j = \mathcal{P}/m$ for all j) minimizes the MSE.

If $\mathbf{R}_R = \mathbf{S}_R$, then MSE minimizing power allocation is given by

$$p_j = \max\left(\sqrt{\frac{\sigma_j^{(Q)}}{\alpha}} - \frac{\sigma_j^{(Q)}}{\lambda_j^{(T)}}, 0\right) \tag{14}$$

where the Lagrange multiplier α is chosen to fulfill the power constraint $\sum_{j=1}^{m} p_j = \mathcal{P}$.

Proof: In the first case, the conditions in (13) are identical for all j and thus the solutions are identical. In the second case, an explicit expression for each p_j can be achieved from (13) since each term of the sum is identical. See [12, Theorem 5.3] for details. ∎

The first part of the corollary represents the case of uncorrelated transmit antennas and temporal disturbance, and has previously been shown in [9] for noise-limited systems. The waterfilling solution in the second part of the corollary was derived in [12] for interference-limited disturbance, but is also valid in noise-limited systems with uncorrelated receive antennas as was shown in [9]–[11].

Next, we give a theorem that shows how the MSE with an optimal training sequence depends on the spatial correlation at the transmitter and receiver side.

Theorem 2: The MSE with the MSE minimizing training matrix is Schur-concave with respect to the eigenvalues of $\mathbf{\Lambda}_T$ (for fixed $\mathbf{\Lambda}_R$). If $\mathbf{\Sigma}_R = \mathbf{I}$, then the MSE is also Schur-concave with respect to the eigenvalues of $\mathbf{\Lambda}_R$ (for fixed $\mathbf{\Lambda}_T$).

Proof: The proof is given in Appendix A. ∎

The interpretation of the theorem is that the MSE with an optimal training matrix will decrease with increasing spatial correlation. This result is intuitive if we consider the extreme: it is easier to estimate the channel in one eigendirection with full training power, than in two eigendirections where each receive half the training power. This analytical behavior provides insight to the selection of parameters like the length of training sequence, B, and the total training power $\mathcal{P}$; as the spatial correlation increases, less power is required to achieve a given MSE and this power will be concentrated in the most important eigendirections of the channel. This will be further analyzed in Section IV-A.

To summarize the results of this section, we have showed the structure of the MSE minimizing training matrix in Kronecker-structured systems and analyzed the allocation of power between the eigendirections. Based on these results, we propose a heuristic training matrix that can be applied under general system conditions. Observe that even when Kronecker-structured approximations are used in the training sequence design, the general MMSE estimator in (6) should always be applied without these approximations.

Heuristic 1: Let $\widetilde{\mathbf{R}}_T \triangleq \mathbb{E}\{\mathbf{H}^H\mathbf{H}\}$ and $\widetilde{\mathbf{S}}_Q \triangleq \mathbb{E}\{\mathbf{N}^H\mathbf{N}\}$. Let their eigenvalue decompositions be $\widetilde{\mathbf{R}}_T = \widetilde{\mathbf{U}}_T\widetilde{\mathbf{\Lambda}}_T\widetilde{\mathbf{U}}_T^H$ and $\widetilde{\mathbf{S}}_Q = \widetilde{\mathbf{V}}_Q\widetilde{\mathbf{\Sigma}}_Q\widetilde{\mathbf{V}}_Q^H$, where the eigenvalues are ordered decreasingly and increasingly, respectively. Then, the training matrix $\mathbf{P} = \widetilde{\mathbf{U}}_T\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}_Q^H$, with diagonal elements $\sqrt{p_1}, \ldots, \sqrt{p_m}$ in $\widetilde{\mathbf{D}}$ that are calculated by inserting the eigenvalues in $\widetilde{\mathbf{\Lambda}}_T$ and $\widetilde{\mathbf{\Sigma}}_Q$ into (14), should provide good performance and minimize the MSE under the Kronecker-structured conditions given in Corollary 1.

It will be illustrated numerically in Section VI that this heuristic training matrix yields good performance, even when the covariance matrices are far from being Kronecker-structured.

A. Optimal Length of Training Sequences

The results of this paper are derived for an arbitrary training sequence length B. Next, we will provide some guidance on how to select this variable under different system statistics and based on the rank of $\mathbf{P}$. Recall from Theorem 1 that all power is allocated in a single eigendirection for low $\mathcal{P}$ (i.e., $\mathrm{rank}(\mathbf{P}) = 1$). Corollary 1 gave a waterfilling solution to the power allocation, and thus strong eigendirections receive more power than weak and only a subset of $p_1, \ldots, p_m$ with cardinality $\tilde{m} \leq m$ will receive any power. Under these conditions, the rank of $\mathbf{P}$ is equal to $\tilde{m}$, which in principle means that the training power is spread in the temporal dimension at the $\tilde{m}$ best channel uses

out of the B allocated for training. Unless the disturbance varies heavily over time, it is not worth wasting $B-\tilde{m}$ channel uses just waiting for better disturbance conditions. Thus, we should select $B=\tilde{m}$. This observation is formalized by the following general theorem.

Theorem 3: Let $\mathbf{P}=\mathbf{U}_P\mathbf{D}_P\mathbf{V}_P^H$ denote the singular value decomposition of the training matrix for $B\geq m$ and suppose that $\mathrm{rank}(\mathbf{P})=\tilde{m}$. If $\mathbf{S}=\mathbf{I}$, then identical MSE is achieved by the $\tilde{m}$-dimensional training sequence $\mathbf{P}'=\mathbf{U}_P[\mathbf{D}_P]_{1:\tilde{m}}$. Here, $[\cdot]_{k_1:k_2}$ denotes the minor matrix that contains column k_1 to k_2 of the given matrix ($k_1\leq k_2$).

Proof: The proof is given in Appendix A. ∎

The interpretation of Theorem 3 is that the optimal training sequence length in noise-limited systems is equal to the rank of $\mathbf{P}$. In this case, optimal means that it is the smallest length B that can achieve the minimal MSE. In general, the rank of $\mathbf{P}$ can only be determined numerically. In certain Kronecker-structured systems, the rank can however be derived explicitly. This is shown by the following corollary, which also relaxes the requirement of uncorrelated disturbance.

Corollary 2: In a Kronecker-structured system with $\mathbf{R}_R=\mathbf{S}_R$, the MSE minimizing training matrix $\mathbf{P}\in\mathbb{C}^{n_T\times B}$ will have rank $m\triangleq\min(n_T,B)$ if

$$\mathcal{P}>\sum_{j=1}^{m-1}\frac{\sqrt{\sigma_j^{(Q)}\sigma_m^{(Q)}}}{\lambda_m^{(T)}}-\frac{\sigma_j^{(Q)}}{\lambda_j^{(T)}} \tag{15}$$

and otherwise have $\mathrm{rank}(\mathbf{P})=\tilde{m}<m$ if $\tilde{m}$ where the positive integer that fulfills

$$\sum_{j=1}^{\tilde{m}-1}\frac{\sqrt{\sigma_j^{(Q)}\sigma_{\tilde{m}}^{(Q)}}}{\lambda_{\tilde{m}}^{(T)}}-\frac{\sigma_j^{(Q)}}{\lambda_j^{(T)}}<\mathcal{P}\leq\sum_{j=1}^{\tilde{m}}\frac{\sqrt{\sigma_j^{(Q)}\sigma_{\tilde{m}+1}^{(Q)}}}{\lambda_{\tilde{m}+1}^{(T)}}-\frac{\sigma_j^{(Q)}}{\lambda_j^{(T)}}. \tag{16}$$

In addition, if $\mathrm{rank}(\mathbf{P})=\tilde{m}<m$ and there exist an integer B'' in $\tilde{m}\leq B''<B$ that factorizes $\mathbf{V}_Q$ as $\mathbf{V}_Q=\begin{bmatrix}\mathbf{V}_Q^{(1)} & \mathbf{0}\\ \mathbf{0} & \mathbf{V}_Q^{(2)}\end{bmatrix}$, for some $\mathbf{V}_Q^{(1)}\in\mathbb{C}^{B''\times B''}$ and $\mathbf{V}_Q^{(2)}\in\mathbb{C}^{B-B''\times B-B''}$. Then, identical MSE is achieved by the B'-dimensional training sequence $\mathbf{P}''=[\mathbf{P}]_{1:B''}=\mathbf{U}_T[\mathbf{D}]_{1:B''}(\mathbf{V}_Q^{(1)})^H\in\mathbb{C}^{n_T\times B''}$.

Proof: The proof is given in Appendix A. ∎

According to the corollary, $\mathbf{P}$ is rank deficient in systems with pronounced spatial correlation and/or limited total training power $\mathcal{P}$. Corollary 2 relaxed the conditions in Theorem 3 by proving that the optimal training sequence length also depend on $\mathrm{rank}(\mathbf{P})$ under certain correlated disturbances. The conditions for this are for example satisfied when $\mathbf{S}_Q=\mathbf{I}$.

Theorem 3 and Corollary 2 constitute a generalization of [14], where it was shown that the optimal training sequence length in spatially uncorrelated and noise-limited systems is exactly equal to n_T. Observe that the generalized results in Corollary 2 stands in contrast to the belief that the training sequence length needs to be at least of length n_T in correlated systems [27].

Under general system statistics, one can expect that $\mathbf{P}$ is rank deficient when the training power is limited and there is a strong eigenvalue spread in either $\mathbf{R}$ or $\mathbf{S}$ (i.e., strong spatial or temporal correlation). Even if the disturbance is correlated so that Theorem 3 cannot be applied, the training sequence length can sometimes be reduced towards $\mathrm{rank}(\mathbf{P})$ with only a slight degradation in MSE and with an improved overall data throughput. The optimal training sequence length under non-Kronecker conditions will be illustrated numerically in Section VI.

V. MMSE Estimation of Squared Channel Norms

In many applications, it is of great interest to estimate the squared Frobenius norm $\|\mathbf{H}\|^2$ of the channel matrix. This norm corresponds directly to the SINR in space-time block coded (STBC) systems and has a large impact on the SINR in many other types of systems [17], [28]. The channel norm can be estimated indirectly from an estimated channel matrix, for example using the estimator in (6). This will however lead to suboptimal performance and gives poor estimates at low training power (see Section VI). Thus, we consider training-based MMSE estimation of $\|\mathbf{H}\|^2$ in this section.

Analysis of the squared channel norm is considerably more involved than for the channel matrix. The next theorem gives a general expression for the MMSE estimator and its MSE, and special expressions for Kronecker-structured systems. In order to derive these expressions, we limit the analysis to training matrices with the structure $\mathbf{P}=\mathbf{U}_T\mathbf{D}\mathbf{V}_Q^H$. It is our conjecture that the MSE minimizing training matrix has this form,[3] as was proved in Theorem 1 for channel matrix estimation. This training matrix structure is also of most practical importance, since the same training signalling will be used to estimate both $\mathbf{H}$ and $\|\mathbf{H}\|^2$.

Theorem 4: The MMSE estimator of $\rho\triangleq\|\mathbf{H}\|^2$, with the observation $\mathbf{Y}=\mathbf{H}\mathbf{P}+\mathbf{N}$ and training sequence $\mathbf{P}$, is

$$\begin{aligned}\hat{\rho}_{\mathrm{MMSE}}=&\int_{\mathbb{C}^{n_Tn_R}}\|\mathbf{H}\|^2\frac{e^{-(\mathrm{vec}(\mathbf{H})-\mathbf{r})^H\mathbf{C}_{\mathrm{MMSE}}^{-1}(\mathrm{vec}(\mathbf{H})-\mathbf{r})}}{\pi^{n_Tn_R}\det(\mathbf{Q})}d\mathrm{vec}(\mathbf{H})\\ =&\mathbf{r}^H\mathbf{r}+\mathrm{tr}\{\mathbf{C}_{\mathrm{MMSE}}\}\end{aligned} \tag{17}$$

where $\mathbf{r}\triangleq\mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{MMSE}})$ and $\mathbf{C}_{\mathrm{MMSE}}$ are defined in (6) and (7), respectively. The corresponding MSE is $\mathrm{MSE}=\mathrm{tr}\{\mathbf{C}_{\mathrm{MMSE}}(2\mathbf{R}-\mathbf{C}_{\mathrm{MMSE}})\}+2\mathrm{vec}(\bar{\mathbf{H}})^H\mathbf{C}_{\mathrm{MMSE}}\mathrm{vec}(\bar{\mathbf{H}})$.

In Kronecker-structured systems with the eigenvalue decompositions in (11) and a training matrix with the structure $\mathbf{P}=\mathbf{U}_T\mathbf{D}\mathbf{V}_Q^H$, the estimator in (17) can be evaluated as

$$\begin{aligned}\hat{\rho}_{\mathrm{MMSE}}=&\sum_{l=1}^{n_R}\sum_{j=1}^{m}\frac{\lambda_j^{(T)}\lambda_l^{(R)}}{\frac{p_j\lambda_j^{(T)}\lambda_l^{(R)}}{\sigma_j^{(Q)}\sigma_l^{(R)}}+1}\left(1+\frac{\left|\frac{g_{lj}}{\lambda_j^{(T)}\lambda_l^{(R)}}+\frac{\sqrt{p_j}\tilde{y}_{lj}}{\sigma_j^{(Q)}\sigma_l^{(R)}}\right|^2}{\frac{p_j}{\sigma_j^{(Q)}\sigma_l^{(R)}}+\frac{1}{\lambda_j^{(T)}\lambda_l^{(R)}}}\right)\\ &+\sum_{l=1}^{n_R}\sum_{j=m+1}^{n_T}\lambda_j^{(T)}\lambda_l^{(R)}+|g_{lj}|^2,\end{aligned} \tag{18}$$

[3]If the mean channel component is strong and has different directivity than the strongest eigenvectors, it might be necessary to permute the eigenvectors in $\mathbf{U}_T$ when constructing the MSE minimizing training matrix $\mathbf{P}$. To simplify the notation, this has been ignored herein, but it is only a matter of reordering the eigenvalues in (11).

where $\tilde{y}_{lj}$ and g_{lj} are the ljth elements of $\widetilde{\mathbf{Y}} \triangleq \mathbf{U}_R^H(\mathbf{Y}-\bar{\mathbf{N}})\mathbf{V}_Q$ and $\mathbf{G} \triangleq \mathbf{U}_R^H\bar{\mathbf{H}}\mathbf{U}_T$, respectively. The corresponding MSE is

$$\mathrm{MSE} = \sum_{l=1}^{n_R}\sum_{j=1}^{m}\left(\frac{\left(\lambda_j^{(T)}\lambda_l^{(R)}\right)^2 + \frac{2p_j\left(\lambda_j^{(T)}\lambda_l^{(R)}\right)^3}{\sigma_j^{(Q)}\sigma_l^{(R)}}}{\left(\frac{p_j\lambda_j^{(T)}\lambda_l^{(R)}}{\sigma_j^{(Q)}\sigma_l^{(R)}}+1\right)^2} + \frac{2|g_{lj}|^2\lambda_j^{(T)}\lambda_l^{(R)}}{\frac{p_j\lambda_j^{(T)}\lambda_l^{(R)}}{\sigma_j^{(Q)}\sigma_l^{(R)}}+1}\right) + \sum_{l=1}^{n_R}\sum_{j=m+1}^{n_T}\left(\lambda_j^{(T)}\lambda_l^{(R)}\right)^2 + 2|g_{lj}|^2\lambda_j^{(T)}\lambda_l^{(R)}. \quad (19)$$

Proof: The proof is given in Appendix A. ∎

The explicit estimator in (18), and its MSE, can also be expressed as matrix multiplications for simplified implementation, see [16] for examples.

A. Training Sequence Design for Channel Norm Estimation

Next, we consider minimization of the MSE of the explicit estimator in (18) by training sequence optimization, which means that we seek the training power allocation $p_1,\ldots,p_m$ in $\mathbf{D}$ that minimizes the MSE. The optimization principles in this section will be similar to those for training matrix estimation, but the MSE of squared norm estimation is not always convex in the training powers, which makes it difficult to derive explicit solutions. The following theorem will however give necessary conditions on the convexity, and provide equations that can be used to determine the solution. We will also analyze the asymptotic behaviors of the power allocation.

Theorem 5: The MSE in (19) is convex in the training power $p_j \geq 0$ if $2|g_{lj}|^2 \geq \lambda_j^{(T)}\lambda_l^{(R)}$ for all l. In general, the MSE can however be non-convex in training powers, but the set of p_j that minimizes the MSE is always given as one of the solutions to the following system of equations:

$$\alpha = \sum_{l=1}^{n_R}\frac{2\frac{\left(\lambda_j^{(T)}\lambda_l^{(R)}\right)^4}{\left(\sigma_j^{(Q)}\sigma_l^{(R)}\right)^2}\left(p_j+|g_{lj}|^2\left(\frac{p_j}{\lambda_j^{(T)}\lambda_l^{(R)}}+\frac{\sigma_j^{(Q)}\sigma_l^{(R)}}{\left(\lambda_j^{(T)}\lambda_l^{(R)}\right)^2}\right)\right)}{\left(p_j\frac{\lambda_j^{(T)}\lambda_l^{(R)}}{\sigma_j^{(Q)}\sigma_l^{(R)}}+1\right)^3} \quad (20)$$

for all active p_j (among $j=1,\ldots,n_T$) and $p_j=0$ otherwise. The Lagrange multiplier $\alpha>0$ is chosen to fulfill the power constraint $\sum_{j=1}^{m}p_j=\mathcal{P}$.

The limiting training matrix at high power $\mathcal{P}$ is given by $p_j = \mathcal{P}\sqrt{c_j}/\sum_{i=1}^{m}\sqrt{c_i}$, where $c_i = \sum_{l=1}^{n_R}\sigma_i^{(Q)}\sigma_l^{(R)}(\lambda_i^{(T)}\lambda_l^{(R)} + |g_{li}|^2)$.

At low power $\mathcal{P}$, the limiting solution is given by $p_{j^*}=\mathcal{P}$ for $j^* = \arg\max_j \sum_{l=1}^{n_R}(\lambda_j^{(T)}\lambda_l^{(R)}/\sigma_j^{(Q)}\sigma_l^{(R)})(\lambda_j^{(T)}\lambda_l^{(R)}+2|g_{lj}|^2)$ and $p_j=0$ for all $j\neq j^*$. If the solution has multiplicity, the power can be distributed arbitrarily among the different p_{j^*}.

Proof: The proof is given in Appendix A. ∎

Although the MSE cannot be guaranteed to be convex, Theorem 5 showed that the limiting training sequences at high and low training power can be derived explicitly. Observe that the MSE in (19) depends on the mean value of the channel, while the MSE for channel matrix estimation is independent of the mean. The limiting solutions are however similar in the sense that all power is allocated in a single eigendirection at low power and are spread in all m spatial direction at high power. The definition of the strongest direction at low training power and the proportional power distribution at large power are however different, which means that the MSE minimizing training matrices usually are different for matrix and squared norm estimation.

The next theorem shows that under certain conditions, the training power allocation can be solved with low complexity, and a unique solution exists if all eigendirections are required to carry a minimal amount of training power.

Corollary 3: If $\mathbf{R}_R = \mathbf{S}_R$, then MSE minimizing power allocation is given by either $p_j=0$ or

$$p_j = \sqrt{\frac{8\sigma_j^{(Q)}(\gamma_j+\nu_j)}{3\lambda_j^{(T)}\alpha}}\cos\left(\frac{\pi(-1)^k-\phi_j}{3}\right) - \frac{\sigma_j^{(Q)}}{\lambda_j^{(T)}} \quad (21)$$

for $k=0,1$, where

$$\gamma_j \triangleq \sum_{l=1}^{n_R}\left(\lambda_j^{(T)}\lambda_l^{(R)}\right)^2,$$

$$\nu_j \triangleq \sum_{l=1}^{n_R}\lambda_j^{(T)}\lambda_l^{(R)}|g_{lj}|^2,$$

and

$$\phi_j \triangleq \arctan\sqrt{(8\lambda_j^{(T)}(\gamma_j+\nu_j)^3/27\sigma_j^{(Q)}\gamma_j^2\alpha)-1}.$$

The Lagrange multiplier $\alpha>0$ is chosen to fulfill the power constraint and the solutions in (21) are only feasible if $\alpha \leq 8\lambda_j^{(T)}(\gamma_j+\nu_j)^3/27\sigma_j^{(Q)}\gamma_j^2$ and when they are positive. Depending on α, solutions in the interval $\sigma_j^{(Q)}(\gamma_j-2\nu_j)/2\lambda_j^{(T)}(\gamma_j+\nu_j) \leq p_j < \infty$ are given by $k=0$, while the interval $-(\sigma_j^{(Q)}/\lambda_j^{(T)}) < p_j \leq \sigma_j^{(Q)}(\gamma_j-2\nu_j)/2\lambda_j^{(T)}(\gamma_j+\nu_j)$ can be achieved by $k=1$ in (21). Thus, if $\gamma_j-2\nu_j<0$ for some j, then $k=1$ will never give a feasible solution for p_j.

If training sequence optimization is combined with the additional constraints $p_j \geq \max(\sigma_j^{(Q)}\sigma_l^{(R)}(\lambda_j^{(T)}\lambda_l^{(R)} - 2|g_{lj}|^2)/2\lambda_j^{(T)}\lambda_l^{(R)}(\lambda_j^{(T)}\lambda_l^{(R)}+|g_{lj}|^2),0)$ for all l and j, the resulting MSE is guaranteed to be convex in the training powers p_j. Then, the system of equations in (20) has a unique solution. In the special case $\mathbf{R}_R=\mathbf{S}_R$, the constraint can be relaxed to $p_j \geq \max(\sigma_j^{(Q)}(\gamma_j-2\nu_j)/2\lambda_j^{(T)}(\gamma_j+\nu_j),0)$ and the optimal power allocation is given by $k=0$ in (21) for all active p_j (i.e., those larger than the new lower bound).

Proof: The proof is given in Appendix A. ∎

The corollary has two important implications. Firstly, in an interference-limited system or in the case of uncorrelated receive antennas, the worst case complexity of finding the solution to the potentially non-convex problem scales with the number of transmit antennas as 3^{n_T}. Secondly, if we impose the additional constraint that all eigendirections are allocated a minimum amount of training power, the power allocation is assured

to be convex and has a unique solution. Observe that in some cases (e.g., for channels with strong mean components), the suggested additional constraint in Corollary 3 can be identical to $p_j \geq 0$ for some j and then the MSE is convex with respect to this p_j without the need of imposing any constraints.

To summarize the results of this section, we have derived an explicit MMSE estimator of the squared channel norm based on the type of training matrices derived in Theorem 1. The power allocation in the training sequence has been analyzed and solved in certain cases. Based on these results, we conclude this section with a heuristic training matrix that can be applied in general Kronecker-structured systems.

Heuristic 2: The training matrix $\mathbf{P} = \mathbf{U}_T \widetilde{\mathbf{D}} \mathbf{V}_Q^H$, with diagonal elements $\sqrt{p_1}, \ldots, \sqrt{p_m}$ in $\widetilde{\mathbf{D}}$ from

$$p_j = \max\left(\sqrt{\frac{8(\gamma_j + \nu_j)}{3 b_j \alpha}} \cos\left(\frac{\pi - \phi_j}{3} \right) - \frac{1}{b_j}, 0 \right) \quad (22)$$

where the Lagrange multiplier α is chosen to fulfill the power constraint $\sum_{j=1}^{m} p_j = \mathcal{P}$, should provide good performance in Kronecker-structured systems. Here, $b_j = \lambda_j^{(T)} \mathrm{tr}(\mathbf{R}_R)/\sigma_j^{(Q)} \mathrm{tr}(\mathbf{S}_R)$, $\gamma_j \triangleq \sum_{l=1}^{n_R} (\lambda_j^{(T)} \lambda_l^{(R)})^2$, $\nu_j \triangleq \sum_{l=1}^{n_R} \lambda_j^{(T)} \lambda_l^{(R)} |g_{lj}|^2$, and $\phi_j \triangleq \arctan \sqrt{(8 b_j (\gamma_j + \nu_j)^3 / 27 \gamma_j^2 \alpha) - 1}$ for all j. If $\mathbf{R}_R = \mathbf{S}_R$ and $p_j \geq \sigma_j^{(Q)} (\gamma_j - 2\nu_j)/2\lambda_j^{(T)} (\gamma_j + \nu_j)$, then the power allocation in (22) will minimize the MSE.

VI. Numerical Examples

In this section, the performance of the MMSE estimators and the training sequence design will be illustrated numerically. The MSE performance of the channel matrix estimator was thoroughly evaluated in [12] for interference-limited Kronecker-structured systems. Thus, we consider the opposite setting of a noise-limited non-Kronecker-structured system, and we will compare the MMSE estimation performance with other recently proposed estimators. This section will also illustrate the advantage of direct MMSE estimation of the squared channel norm over indirect calculation from an estimated channel matrix. Finally, we will illustrate how the smallest necessary length of the training sequence depends on the spatial correlation and available training power.

To illustrate the performance of the training sequence design for channel matrix estimation in Section IV under general channel conditions, we consider the Weichselberger model [25]. This model has recently attracted much attention for its accurate representation of measurement data. According to this model, the channel matrix can be expressed as $\mathbf{H} = \mathbf{U}_A \widetilde{\mathbf{H}} \mathbf{U}_B^H$, where $\mathbf{U}_A, \mathbf{U}_B$ are unitary matrices and $\widetilde{\mathbf{H}} \in \mathbb{C}^{n_R \times n_T}$ has independent elements with variances given by the corresponding elements of the coupling matrix $\mathbf{\Omega}$. The unitary matrices will not affect the performance when MSE minimizing precoding design is employed, and can therefore be selected as identity matrices. Without loss of generality, we always scale the coupling matrices as $\mathrm{tr}(\mathbf{\Omega}) = n_T n_R$ to make sure that the training SINR can be described by the training power constraint: $\mathrm{SINR}_{\mathrm{training}} = \mathcal{P}\mathrm{tr}(\mathbf{R})/\mathrm{tr}(\mathbf{S}) = \mathcal{P}$. To enable comparison with other estimators, the channel is zero-mean, but recall from the MSE expression in (8) that the performance is unaffected by non-zero mean components.

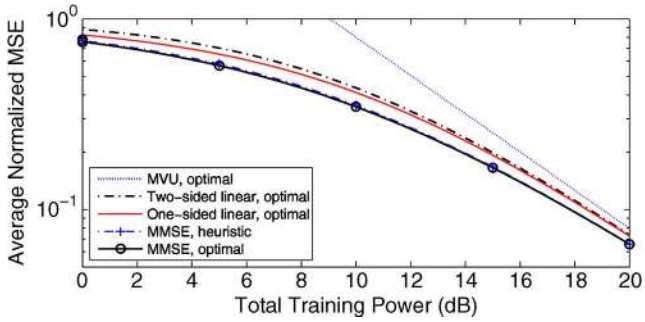

Fig. 1. The average normalized MSEs of channel matrix estimation as a function of the total training power in a system with the Weichselberger model and χ_2^2-distributed coupling matrices. The performance of four different estimators with MSE minimizing training matrices is compared. The performance with the training matrix design in Heuristic 1 is also given.

We define the normalized MSE as $\mathbb{E}\{\|\mathbf{H} - \widehat{\mathbf{H}}_{\mathrm{MMSE}}\|^2\}/\mathrm{tr}(\mathbf{R})$. In Fig. 1, we give the normalized MSEs averaged over 5,000 scenarios with different coupling matrices with $n_T = 8$, $n_R = 4$, and independent χ_2^2-distributed elements. The performance of four different estimators with MSE minimizing training matrices are compared: the MVU/ML channel estimator $\widehat{\mathbf{H}} = \mathbf{Y}\mathbf{P}^H(\mathbf{P}\mathbf{P}^H)^{-1}$ [8], the one-sided linear estimator in [8], [13] that was incorrectly claimed to be the linear MMSE estimator, the two-sided Bayesian linear estimator proposed in [27], and the MMSE estimator in (6). The MVU/ML estimator[4] is unaware of the channel statistics (i.e., non-Bayesian), and it is clear from Fig. 1 that this leads to poor estimation performance. The two-sided linear estimator also performs poorly under the given premises, but can provide good performance in special cases [27]. The performance gap between the one-sided linear estimator and the MMSE estimator (which is also linear) is noticeable, while the difference between employing the optimal training matrix and the one proposed in Heuristic 1 is small. It should be pointed out that the use of independent χ_2^2-distributed elements in the coupling matrix induces a spatially correlated environment with a few dominating paths. In less correlated scenarios, the difference between the estimators decreases, but the order of quality is usually the same.

In Fig. 2, the performance of the MMSE estimator is shown for a uniform training matrix ($\mathbf{P} = \sqrt{\mathcal{P}/n_T}\mathbf{I}$), MSE minimizing training matrix (achieved numerically), and the simple explicit training matrix proposed in Heuristic 1. The one-sided linear estimator is given as a reference. In this simulation, we used the coupling matrix that was proposed in [29, Eq. 28] to describe an environment with two small scatterers, two big scatterers, and one large cluster. It is clear that the gain of employing an MSE minimizing training sequence is substantial, and the heuristic approach captures most of this gain although uniform training is asymptotically optimal at high training power.

Next, we illustrate the optimal length of the training sequence for varying spatial correlation and training power. Recall from Theorem 3 that the optimal length in noise-limited systems is

[4]For this problem, the maximum likelihood (ML) estimator is equivalent to the MVU [23, Theorem 7.5].

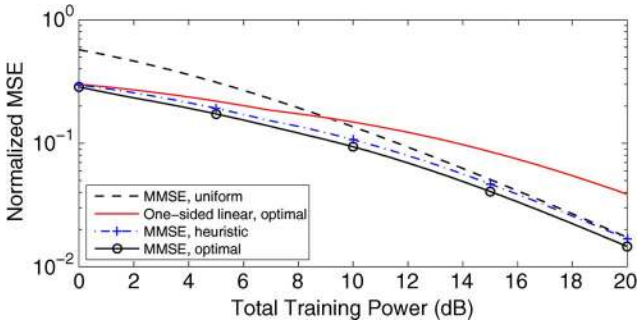

Fig. 2. The normalized MSEs of channel matrix estimation as a function of the total training power in a system with the Weichselberger model and the coupling matrix proposed in [29, Eq. 28]. The MMSE estimator with three different training matrices is compared with the one-sided linear estimator.

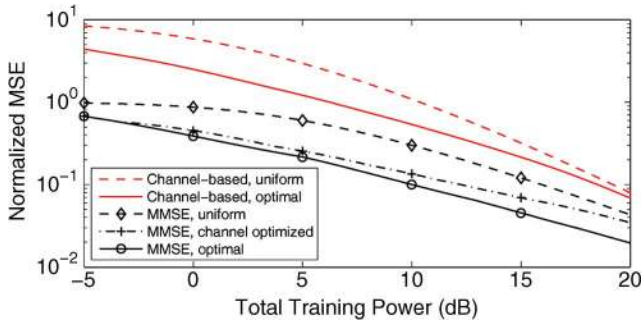

Fig. 4. The normalized MSEs of channel squared norm estimation as a function of the total training power in a system with uncorrelated receive antennas and a transmit antenna correlation of 0.8. The MMSE estimator is compared with indirect estimation from an MMSE estimated channel matrix for different training matrices.

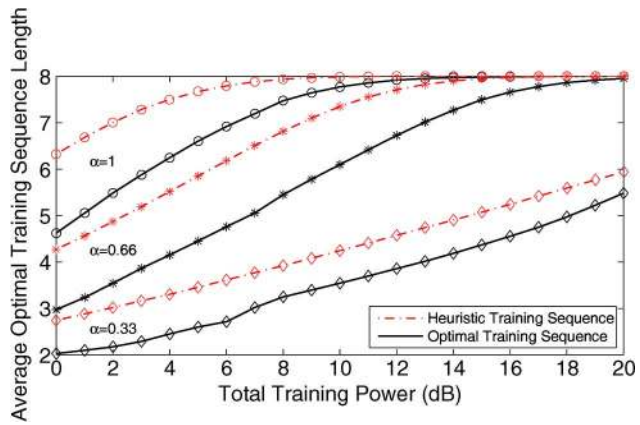

Fig. 3. The average optimal training sequence length (smallest length that minimizes the MSE) as a function of the total training power $\mathcal{P}$. The system follows the Weichselberger model where the jth column of the coupling matrix has independent χ^2_2-distributed elements scaled by α^{j-1}, for different α. Decreasing α means increasing spatial correlation.

equal to the rank of the training matrix. We consider coupling matrices with $n_T = 8, n_R = 4$, and independent χ^2_2-distributed elements, and we induce random transmit-side correlation by scaling the jth column by α^{j-1} for different values on α. The average optimal training sequence length (i.e., average rank of $\mathbf{P}$) is shown in Fig. 3 for both an MSE minimizing training matrix and the training matrix proposed in Heuristic 1. The average length is given as a function of the total training power and for the spatial correlation induced by $\alpha \in \{0.33, 0.66, 1\}$. In the case of identically distributed elements of the coupling matrix ($\alpha = 1$), there is sufficient spatial correlation to have $\mathrm{rank}(\mathbf{P}) < n_T$ at low training power. As the spatial correlation increases (i.e., α decreases), the optimal training length decreases and the convergence towards full rank becomes slower. The heuristic training approach is clearly overestimating the training length, which explains the performance difference in Fig. 1. An important observation is that the conclusion of [14] that the optimal length in an uncorrelated system is equal to the number transmit antennas does not hold in general. Careful system analysis is always required to determine the optimal length under general statistics, and the loss in performance by employing an even shorter training sequence may be minor compared with the gain of having more data symbols.

Finally, we illustrate the performance of squared norm estimation. The normalized MSEs for channel squared norm estimation, defined as $\mathbb{E}\{|\|\mathbf{H}\|^2 - \|\widehat{\mathbf{H}}_{\mathrm{MMSE}}\|^2|^2\}/\mathrm{tr}(\mathbf{R}\mathbf{R}^H)$, are given in Fig. 4 as a function of the total training power. In this case, we limit the simulation to Kronecker-structured systems (i.e., rank-one coupling matrices), since the explicit estimator in Theorem 5 is based on this assumption. We consider uncorrelated receive antennas and a correlation between adjacent transmit antennas of 0.8, using the exponential model [30]. The performance of the MMSE estimator in Theorem 5 is compared with indirect calculation of the squared norm from a channel matrix that is estimated using (6). In both approaches, uniform and optimal training sequences are considered. For the MMSE estimator, the performance with a channel matrix optimized training sequence is also shown for comparison. This is probably the most important case in practice; the training sequence will be used to optimize estimation of the channel matrix (or some receive filter), but the received training signal can simultaneously be used to calculate an MMSE estimate of the squared norm (e.g., for the purpose of feedback). The first observation from Fig. 4 is that the indirect approach yields poor performance at low SINR (even worse than the purely statistical estimator $\hat{\rho}_{\mathrm{stat}} = \mathrm{tr}(\mathbf{R})$ which would give unit normalized MSE) and is not even asymptotically optimal at high SINR. The performance of the MMSE estimator can be considerably improved by proper training sequence design. A training sequence designed for channel matrix estimation will improve the performance over uniform training at low SINR, but they both share the same suboptimal asymptotic behavior.

VII. Conclusion

A framework for training-based estimation of Rician fading MIMO channel matrices has been introduced, for the purpose of joint analysis under different noise and interference conditions. The MMSE estimator was analyzed in terms of the MSE minimizing training sequence and the optimal training structure was derived in Kronecker-structured systems. The limiting solutions at high and low training power were given, along with sufficient conditions for when the training optimization can be solved explicitly. Based on these results, a heuristic training sequence was proposed for arbitrary system statistics.

In addition, we proved analytically that the MSE improves with the spatial correlation at both the transmitter and the receiver side. This result was used to clarify how the optimal length of the training sequence depends on the system statistics

and the total training power. An interesting result was that the optimal training sequence length can be considerably smaller than the number of transmit antennas in systems with strong spatial correlation. This was proved analytically for certain Kronecker-structured systems.

Finally, the framework was extended to MMSE estimation of the squared Frobenius norm of the channel, using the same type of training sequences as for channel matrix estimation. Although the MSE of this estimator can be non-convex, the limiting solutions at high and low training power were derived and it was shown under which conditions the solution can be derived explicitly or with low complexity.

APPENDIX A
COLLECTION OF LEMMAS AND PROOFS

In the appendix, we will first state two lemmas and then apply them when proving the theorems of this paper. The first lemma provides the necessary structure of the training matrix when the weighted sum of MSEs is minimized, and is essentially a generalization of [12, Corollary 5.1] where a single MSE was minimized (i.e., $L = 1$).

Lemma 1: Let $a_1, \ldots, a_L$ and $b_1, \ldots, b_L$ be positive coefficients, and let $\mathbf{\Lambda} \in \mathbb{R}^{N \times N}$ and $\mathbf{\Sigma} \in \mathbb{R}^{M \times M}$ be diagonal matrices with strictly positive elements ordered decreasingly and increasingly, respectively. Then, the optimization problem

$$\min_{\bar{\mathbf{P}} \in \mathbb{C}^{N \times M}} \sum_{j=1}^{L} \mathrm{tr}\left\{ (a_j \mathbf{\Lambda}^{-1} + b_j \bar{\mathbf{P}} \mathbf{\Sigma}^{-1} \bar{\mathbf{P}}^H)^{-1} \right\} \quad \text{subject to} \quad \mathrm{tr}(\bar{\mathbf{P}}^H \bar{\mathbf{P}}) \leq \mathcal{P} \tag{23}$$

is solved by $\bar{\mathbf{P}} \in \mathbb{C}^{N \times M}$ being a rectangular diagonal matrix that satisfies $\mathrm{tr}(\bar{\mathbf{P}}^H \bar{\mathbf{P}}) = \mathcal{P}$ and gives decreasingly ordered diagonal elements of $\bar{\mathbf{P}} \mathbf{\Sigma}^{-1} \bar{\mathbf{P}}^H$ (i.e., the same order as for $\mathbf{\Sigma}^{-1}$).

Proof: We will derive the structure of the optimal $\bar{\mathbf{P}}$ by contradiction; that is, for every $\bar{\mathbf{P}}$ that fulfill the constraint we can find a solution that satisfy the given structure and achieves a smaller or identical function value. Observe that the function $\mathrm{tr}\{(\cdot)^{-1}\}$ is strictly convex in each eigenvalue of its argument matrix. Therefore, if the constraint is not fulfilled with equality for a given $\bar{\mathbf{P}}$, we can always achieve a smaller function value by replacing it by $\alpha \bar{\mathbf{P}}$ for some $\alpha > 1$ and still satisfy the constraint.

Suppose that $\bar{\mathbf{P}}$ fulfills the constraint with equality, and let its singular value decomposition be denoted $\bar{\mathbf{P}} = \mathbf{U}_P \mathbf{D}_P \mathbf{V}_P^H$. We will first show that $\mathbf{V}_P$ can be removed if the diagonal elements of $\mathbf{D}_P$ are reordered. For this purpose we introduce $\mathbf{W} \triangleq \bar{\mathbf{P}} \mathbf{\Sigma}^{-1/2}$ and let its singular value decomposition be denoted $\mathbf{W} = \mathbf{U}_W \mathbf{D}_W \mathbf{V}_W^H$, where the singular values in $\mathbf{D}_W$ are ordered decreasingly. Now, observe that $\bar{\mathbf{P}}$ only appears in the cost function as $\bar{\mathbf{P}} \mathbf{\Sigma}^{-1} \bar{\mathbf{P}}^H = \mathbf{U}_W \mathbf{D}_W \mathbf{D}_W^H \mathbf{U}_W^H$ and thus we can modify $\mathbf{V}_W$ without affecting the function value. Using the new notation, the power constraint can be expressed as

$$\begin{aligned} \mathcal{P} &= \mathrm{tr}(\bar{\mathbf{P}}^H \bar{\mathbf{P}}) = \mathrm{tr}(\mathbf{W}^H \mathbf{W} \mathbf{\Sigma}) \\ &\geq \sum_{i=1}^{M} \lambda_i(\mathbf{W}^H \mathbf{W}) \lambda_{M-i+1}(\mathbf{\Sigma}) \end{aligned} \tag{24}$$

where $\lambda_i(\cdot)$ denotes the ith largest eigenvalue. The last inequality is given in [22, Theorem 20.A.4] and is fulfilled with equality if and only if $\mathbf{W}^H \mathbf{W}$ is diagonal with elements in the opposite order of $\mathbf{\Sigma}$, which means that $\mathbf{V}_W = \mathbf{I}$ would minimize the constraint. For this $\mathbf{V}_W$, we have the relationship

$$\bar{\mathbf{P}} \mathbf{\Sigma}^{-1/2} = \mathbf{U}_P \mathbf{D}_P \mathbf{V}_P^H \mathbf{\Sigma}^{-1/2} = \mathbf{U}_W \mathbf{D}_W \tag{25}$$

which is satisfied if $\mathbf{V}_P = \mathbf{I}$ and the diagonal values of $\mathbf{D}_P$ is ordered such that $\mathbf{D}_P \mathbf{\Sigma}^{-1/2}$ is in decreasing order. If this is not fulfilled for the given $\bar{\mathbf{P}}$, we can always find a better solution that fulfills them by first reordering the elements of $\mathbf{D}_P$ and removing $\mathbf{V}_P$ which will give strict inequality in the constraint. Then, a smaller function value is achieved by scaling the new solution to achieve equality in the constraint. Thus, the optimal solution has the structure $\bar{\mathbf{P}} = \mathbf{U}_P \mathbf{D}_P$, where $\mathbf{D}_P$ is ordered as described.

Finally, for a solution of the type $\bar{\mathbf{P}} = \mathbf{U}_P \mathbf{D}_P$, we will show that we always can reduce the function value by selecting $\mathbf{U}_P = \mathbf{I}$. Let $\mathbf{A}_j \triangleq a_j \mathbf{\Lambda}^{-1} + b_j \bar{\mathbf{P}} \mathbf{\Sigma}^{-1} \bar{\mathbf{P}}^H$, and observe that

$$\mathrm{tr}\left\{ \mathbf{A}_j^{-1} \right\} = \sum_{l=1}^{N} \frac{1}{\lambda_l(\mathbf{A}_j)}. \tag{26}$$

As mentioned in the beginning of the proof, each component of the sum is strictly convex in its eigenvalue. Thus, (26) is a Schur-convex function for all j [20, Proposition 2.7]. Recall that $\mathbf{A}_j$ is a linear combination of $\mathbf{\Lambda}^{-1}$ and $\bar{\mathbf{P}} \mathbf{\Sigma}^{-1} \bar{\mathbf{P}}^H$ with positive coefficients for each j. Then, we have from [20, Theorem 2.11] that each $\mathrm{tr}\{\mathbf{A}_j^{-1}\}$ is minimized when the eigenvalues of $\mathbf{\Lambda}^{-1}$ and $\bar{\mathbf{P}} \mathbf{\Sigma}^{-1} \bar{\mathbf{P}}^H$ are added together in opposite order. If $\mathbf{U}_P \neq \mathbf{I}$, we can therefore decrease the function value by replacing it by an identity matrix, without affecting the power constraint.

To summarize, we have showed that for every given $\bar{\mathbf{P}}$, we can reduce the cost function by removing the unitary matrices of its singular value decomposition, reordering the diagonal elements, and scaling the remaining matrix to satisfy the constraint with equality. ■

The next lemma provides a simple condition to determine if a function that originates from an optimal power allocation is Schur-convex or Schur-concave.

Lemma 2: Consider a continuous and twice continuously differentiable function $f(\boldsymbol{\lambda}, \mathbf{p})$ of two non-negative vectors $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_N]^T$ and $\mathbf{p} = [p_1, \ldots, p_M]^T$. For every $\boldsymbol{\lambda}$ that $f(\boldsymbol{\lambda}, \mathbf{p})$ is convex and the Hessian and all its square minors are non-singular with respect to $\mathbf{p}$, the solution to the optimization

$$g(\boldsymbol{\lambda}) = \min_{\mathbf{p}} f(\boldsymbol{\lambda}, \mathbf{p}) \quad \text{subject to} \quad \sum_{l=1}^{M} p_l = \mathcal{P} \quad \text{and} \quad p_l \geq 0 \tag{27}$$

is differentiable. The partial derivatives of the solution at optimal power allocation $\mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})$ are

$$\frac{\partial g(\boldsymbol{\lambda})}{\partial \lambda_j} = f'_{\lambda_j}(\boldsymbol{\lambda}, \mathbf{p})|_{\mathbf{p} = \mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})} \quad \text{for all } j. \tag{28}$$

Then, the function $g(\boldsymbol{\lambda})$ is Schur-convex with respect to $\boldsymbol{\lambda}$ if and only if $\partial g(\boldsymbol{\lambda}) / \partial \lambda_i \geq \partial g(\boldsymbol{\lambda}) / \partial \lambda_j$ for all $\lambda_i \geq \lambda_j$, and Schur-concave if and only if $\partial g(\boldsymbol{\lambda}) / \partial \lambda_i \leq \partial g(\boldsymbol{\lambda}) / \partial \lambda_j$.

Proof: Since the cost function is convex with respect to $\mathbf{p}$ for every given $\boldsymbol{\lambda}$ and the domain of $\mathbf{p}$ is closed, the Karush-Kuhn-Tucker (KKT) conditions guarantee the existence of one or several solutions to (27) and these are given by the following system of stationarity equations

$$0 = f'_{\lambda_j}(\boldsymbol{\lambda}, \mathbf{p}) - \alpha \tag{29}$$

for all $p_l > 0$ (otherwise $p_l = 0$), where the Lagrangian multiplier α makes sure that $\sum_{l=1}^{M} p_l = \mathcal{P}$ [31]. Let $\mathcal{S}$ denote the index set of all non-zero p_l and those $p_l = 0$ for which the corresponding equation in (29) also would be satisfied with equality (i.e., on those that are on the boundary of becoming active). Observe that the Jacobian of the equation system in (29) for these $\mathcal{S}$ will be identical to a minor of the Hessian of $f(\boldsymbol{\lambda}, \mathbf{p})$ with respect to $\mathbf{p}$, and thus non-singular by assumption. If we denote the power allocation solution in (27) as a function $\mathbf{p}_{\mathrm{opt}} = \mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})$, we can then apply the Implicit function theorem to conclude all elements in $\mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})$ with indexes in $\mathcal{S}$ are differentiable with respect to $\boldsymbol{\lambda}$ [32, Theorem 9.28]. For those $p_l = 0$ with $l \notin \mathcal{S}$, this variable can be replaced with a zero in the optimization problem without affecting the solution, and thus its derivative can be defined as being zero.

We can now use that $\mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})$ is differentiable with respect to $\boldsymbol{\lambda}$ to calculate the partial derivative of $g(\boldsymbol{\lambda})$ with respect to λ_j:

$$\begin{aligned}\frac{\partial g(\boldsymbol{\lambda})}{\partial \lambda_j} &= \frac{\partial}{\partial \lambda_j} f\left(\boldsymbol{\lambda}, \mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})\right) \\ &= f'_{\lambda_j}(\boldsymbol{\lambda}, \mathbf{p})|_{\mathbf{p}=\mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})} \\ &\quad + \sum_{l=0}^{M} f'_{p_l}(\boldsymbol{\lambda}, \mathbf{p})|_{\mathbf{p}=\mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})} \frac{\partial p_l}{\partial \lambda_j}|_{\mathbf{p}=\mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})}. \end{aligned} \tag{30}$$

Since $f'_{p_l}(\boldsymbol{\lambda}, \mathbf{p})|_{\mathbf{p}=\mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})} = \alpha$ for $l \in \mathcal{S}$ and $\partial p_l / \partial \lambda_j|_{\mathbf{p}=\mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})} = 0$ for $l \notin \mathcal{S}$, we have that

$$\begin{aligned}&\sum_{l=0}^{M} f'_{p_l}(\boldsymbol{\lambda}, \mathbf{p})|_{\mathbf{p}=\mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})} \frac{\partial p_l}{\partial \lambda_j}|_{\mathbf{p}=\mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})} \\ &\qquad = \alpha \sum_{l \in \mathcal{S}} \frac{\partial p_l}{\partial \lambda_j}|_{\mathbf{p}=\mathbf{p}_{\mathrm{opt}}(\boldsymbol{\lambda})} = 0 \end{aligned} \tag{31}$$

where the last equality follows from that $\sum_{l \in \mathcal{S}} p_l = P$ implies that $\sum_{l \in \mathcal{S}} (\partial p_l / \partial \lambda_j) = 0$. Thus, we have proved (28). The last sentence of the lemma follows directly from Schur's condition in [22, Theorem 3.A.4], which states that $g(\boldsymbol{\lambda})$ is Schur-convex if and only if

$$(\lambda_i - \lambda_j)\left(\frac{\partial g(\boldsymbol{\lambda})}{\partial \lambda_i} - \frac{\partial g(\boldsymbol{\lambda})}{\partial \lambda_j}\right) \geq 0 \tag{32}$$

for all i and j, and Schur-concave if the conditions are fulfilled with inverted inequalities. ∎

Finally, we give the proofs of Theorems 1–5 and Corollary 3.

Proof of Theorem 1: First, we derive the structure of the MSE minimizing training matrix. For Kronecker-structured systems, the MSE can be expressed as $\mathrm{MSE} = \mathrm{tr}\{(\mathbf{R}_T^{-T} \otimes \mathbf{R}_R^{-1} + (\mathbf{P}^* \mathbf{S}_Q^{-T} \mathbf{P}^T) \otimes \mathbf{S}_R^{-1})^{-1}\}$. By taking the conjugate transpose of the training transmission model in (2) and then applying the results of [23, Chapter 15.8] in the same manner as in Section III, we achieve an alternative expression of the MSE:

$$\begin{aligned}&\mathrm{MSE} \\ &= \mathrm{tr}\left\{\left(\mathbf{R}_R^{-T} \otimes \mathbf{R}_T^{-1} + \mathbf{S}_R^{-T} \otimes \left(\mathbf{P}\mathbf{S}_Q^{-1}\mathbf{P}^H\right)\right)^{-1}\right\} \\ &= \sum_{l=1}^{n_R} \mathrm{tr}\left\{\left(\frac{1}{\lambda_l^{(R)}}\boldsymbol{\Lambda}_T^{-1} + \frac{1}{\sigma_l^{(R)}}\mathbf{U}_T^H \mathbf{P}\mathbf{V}_Q \boldsymbol{\Sigma}_Q^{-1} \mathbf{V}_Q^H \mathbf{P}^H \mathbf{U}_T\right)^{-1}\right\} \end{aligned} \tag{33}$$

where the second equality follows from that the identical eigenvectors of $\mathbf{R}_R$ and $\mathbf{S}_R$ are not affecting the trace and that the trace of block matrices is equal to the sum of the traces for each block.

Using the notation $a_j = 1/\lambda_l^{(R)}$, $b_j = 1/\sigma_l^{(R)}$, and $\bar{\mathbf{P}} = \mathbf{U}_T^H \mathbf{P} \mathbf{V}_Q$, we can apply Lemma 1 to conclude that the MSE minimizing $\bar{\mathbf{P}}$ should be rectangularly diagonal, fulfill the element ordering given in the theorem, and satisfy the power constraint with equality. With a training matrix of this type, the argument in (33) will be diagonal, and the MSE can be expressed as

$$\mathrm{MSE} = \sum_{j=1}^{m} \sum_{l=1}^{n_R} \frac{\lambda_j^{(T)} \lambda_l^{(R)}}{1 + p_j \frac{\lambda_j^{(T)} \lambda_l^{(R)}}{\sigma_j^{(Q)} \sigma_l^{(R)}}} + \mathrm{tr}(\mathbf{R}_R) \sum_{j=m+1}^{n_T} \lambda_j^{(T)} \tag{34}$$

which is a convex function with respect to each p_j (since $a/(1+abp_j)$ is a convex function for all $a, b > 0$). Thus, the KKT conditions give the necessary and sufficient condition for the optimal power allocation [31, Ch. 5.5] and these are summarized in (13).

Finally, we consider the two asymptotic cases. At high power, we approximate the MSE in (34) as

$$\begin{aligned}\mathrm{MSE} &\approx \sum_{j=1}^{m} \sum_{l=1}^{n_R} \frac{\lambda_j^{(T)} \lambda_l^{(R)}}{p_j \frac{\lambda_j^{(T)} \lambda_l^{(R)}}{\sigma_j^{(Q)} \sigma_l^{(R)}}} + \mathrm{tr}(\mathbf{R}_R) \sum_{j=m+1}^{n_T} \lambda_j^{(T)} \\ &= \mathrm{tr}(\mathbf{S}_R) \sum_{j=1}^{m} \frac{\sigma_j^{(Q)}}{p_j} + \mathrm{tr}(\mathbf{R}_R) \sum_{j=m+1}^{n_T} \lambda_j^{(T)} \end{aligned} \tag{35}$$

which is minimized by $p_j = \mathcal{P}\sqrt{\sigma_j^{(Q)}}/C$ for all j (using straightforward Lagrangian methods). At low power, we approximate (34) as

$$\begin{aligned}\mathrm{MSE} &\approx \sum_{j=1}^{m} \sum_{l=1}^{n_R} \lambda_j^{(T)} \lambda_l^{(R)} \left(1 - \frac{p_j \lambda_j^{(T)} \lambda_l^{(R)}}{\sigma_j^{(Q)} \sigma_l^{(R)}}\right) + \mathrm{tr}(\mathbf{R}_R) \sum_{j=m+1}^{n_T} \lambda_j^{(T)} \\ &= \mathrm{tr}(\mathbf{R}_T)\mathrm{tr}(\mathbf{R}_R) - \sum_{l=1}^{n_R} \frac{\left(\lambda_l^{(R)}\right)^2}{\sigma_l^{(R)}} \sum_{j=1}^{m} \frac{p_j \left(\lambda_j^{(T)}\right)^2}{\sigma_j^{(Q)}} \end{aligned} \tag{36}$$

using a first order Taylor polynomial. This expression is minimized by assigning all power in an arbitrary manner among the strongest term/terms of the second sum.

Proof of Theorem 2: First, we will prove that the MSE in (34) is Schur-concave with respect to the eigenvalues $\lambda_1^{(T)}, \ldots, \lambda_{n_T}^{(T)}$. It is straightforward to show that the MSE is convex in the power allocation, differentiable with respect to $\lambda_j^{(T)}$ and p_j for

all j, and that the determinant of the Hessian is non-zero if the eigenvalues of $\mathbf{\Lambda}_T$ and $\mathbf{\Lambda}_R$ are distinct. Thus, we can apply Lemma 2. According to the lemma, it is sufficient to show that $\partial\mathrm{MSE}/\partial\lambda_i^{(T)} \leq \partial\mathrm{MSE}/\partial\lambda_j^{(T)}$ for all i,j such that $\lambda_i^{(T)} \geq \lambda_j^{(T)}$, where MSE denotes the pre-optimization MSE in (34) evaluated at the optimal solution. Thus, we can calculate the partial derivatives of (34) as

$$\frac{\mathrm{MSE}}{\partial\lambda_j^{(T)}} = \sum_{l=1}^{n_R} \frac{\lambda_l^{(R)}}{\left(1+p_j\frac{\lambda_j^{(T)}\lambda_l^{(R)}}{\sigma_j^{(Q)}\sigma_l^{(R)}}\right)^2} \quad \text{for } j=1,\ldots,m \quad (37)$$

and $\partial\mathrm{MSE}/\partial\lambda_j^{(T)} = \mathrm{tr}(\mathbf{R}_R)$ for $j=m+1,\ldots,n_T$. Observe that the derivatives are positive and that $\lambda_j^{(T)}$ and $p_j/\sigma_j^{(Q)}$ only appear in the denominator of (37). From Theorem 1, we have that $p_i/\sigma_i^{(Q)} \geq p_j/\sigma_j^{(Q)}$ whenever $\lambda_i^{(T)} \geq \lambda_j^{(T)}$. Hence, it follows that $\partial\mathrm{MSE}/\partial\lambda_i^{(T)} \leq \partial\mathrm{MSE}/\partial\lambda_j^{(T)}$ and that the MSE is Schur-concave.

Next, we have the case when $\mathbf{\Sigma}_R = \mathbf{I}$, and then the MSE in (34) can be expressed as

$$\mathrm{MSE} = \sum_{l=1}^{n_R}\left(\underbrace{\sum_{j=1}^{m}\frac{\lambda_j^{(T)}}{\frac{1}{\lambda_l^{(R)}}+p_j\frac{\lambda_j^{(T)}}{\sigma_j^{(Q)}}}}_{(a)} + \underbrace{\lambda_l^{(R)}\sum_{j=m+1}^{n_T}\lambda_j^{(T)}}_{(b)}\right) \quad (38)$$

which is a concave function in $\lambda_l^{(R)}$ for all l. We apply [22, Proposition 3.C.1] to conclude that parts (a) and (b) are both Schur-concave with respect to $\lambda_1^{(R)},\ldots,\lambda_{n_R}^{(R)}$, and thus the MSE is Schur-concave.

Proof of Theorem 3: For $\mathbf{S}=\mathbf{I}$, the MSE in (9) becomes

$$\mathrm{tr}\left\{\left(\mathbf{R}^{-1}+(\mathbf{P}^T\otimes\mathbf{I})^H\mathbf{S}^{-1}(\mathbf{P}^T\otimes\mathbf{I})\right)^{-1}\right\} = \mathrm{tr}\left\{\left(\mathbf{R}^{-1}+\left(\mathbf{U}_P\mathbf{D}_P\mathbf{D}_P\mathbf{U}_P^H\otimes\mathbf{I}\right)^T\right)^{-1}\right\}. \quad (39)$$

The theorem follows from that (39) is independent of $\mathbf{V}_P$ and that $\mathbf{U}_P\mathbf{D}_P\mathbf{D}_P\mathbf{U}_P^H = \mathbf{P}'(\mathbf{P}')^H$.

Proof of Corollary 2: The rank of $\mathbf{P}$ is equal to the number of active training powers p_j. From Theorem 1, we have that the $\tilde{m}$th training power is active if and only if $\alpha < \left(\lambda_{\tilde{m}}^{(T)}\right)^2/\sigma_{\tilde{m}}^{(Q)}$. Suppose we only have $\tilde{m}-1$ active training powers, then $\alpha \geq \left(\lambda_{\tilde{m}}^{(T)}\right)^2/\sigma_{\tilde{m}}^{(Q)}$. Substitution into the power constraint gives

$$\mathcal{P} = \sum_{j=1}^{\tilde{m}-1}\sqrt{\frac{\sigma_j^{(Q)}}{\alpha}} - \frac{\sigma_j^{(Q)}}{\lambda_j^{(T)}} \leq \sum_{j=1}^{\tilde{m}-1}\frac{\sqrt{\sigma_j^{(Q)}\sigma_{\tilde{m}}^{(Q)}}}{\lambda_{\tilde{m}}^{(T)}} - \frac{\sigma_j^{(Q)}}{\lambda_j^{(T)}} \quad (40)$$

for $1 \leq \tilde{m} < m$. All p_j will be active if and only if $\mathcal{P}$ is larger than the constraint for $\tilde{m} = m-1$.

Finally, if there exist a B'' that fulfills the requirements, then $\mathbf{Y} = \mathbf{HP}+\mathbf{N}$ can be factorized as $[[\mathbf{Y}]_{1:B''}[\mathbf{Y}]_{B''+1:B}] = [[\mathbf{HP}]_{1:B''}\mathbf{0}] + [[\mathbf{N}]_{1:B''}[\mathbf{N}]_{B''+1:B}]$, where $[\mathbf{N}]_{1:B''}$ and $[\mathbf{N}]_{B''+1:B}$ are independent. Thus, $[\mathbf{Y}]_{B''+1:B}$ neither contain information of the channel matrix nor is correlated with previous disturbance in $[\mathbf{N}]_{1:B''}$, and will therefore not affect the estimation. We can therefore use the shorter training sequence $\mathbf{P}''$ without any loss in performance.

Proof of Theorem 4: In the general case, the integral expression of the MMSE estimator in (17) follows directly from the definition of $\mathbb{E}\{||\mathbf{H}||^2|\mathbf{Y}\}$ by exploiting that the posterior distribution, $f_{\mathrm{vec}(\mathbf{H})|\mathrm{vec}(\mathbf{Y})}(\mathrm{vec}(\mathbf{H})|\mathrm{vec}(\mathbf{Y}))$, is complex Gaussian distributed with the mean and covariance matrix derived in [23, Chapter 15.8].

To derive the explicit expressions, we begin with the one-dimensional case ($n_T = n_R = 1$) with the received signal $y = qh+n$, where q is the training signal, $h \in \mathcal{CN}(a,\lambda)$, and $n \in \mathcal{CN}(b,\sigma)$. Using Bayes' formula or [23, Chapter 15.8], the posterior distribution $f_{h|y}(h|y)$ can be expressed as

$$\begin{aligned} f_{h|y}(h|y) &= \frac{f(y|h)f(h)}{f(y)} = \frac{\frac{e^{-|y-qh-b|^2/\sigma}}{\pi\sigma}\frac{e^{-|h-a|^2/\lambda}}{\pi\lambda}}{\frac{e^{-|y-qa-b|^2/(|q|^2\lambda+\sigma)}}{\pi(|q|^2\lambda+\sigma)}} \\ &= \frac{1}{\pi}\left(\frac{1}{\lambda}+\frac{|q|^2}{\sigma}\right) \\ &\quad\times e^{-\left|h-a-\frac{q^*\lambda(y-qa-b)}{|q|^2\lambda+\sigma}\right|^2\left(\frac{1}{\lambda}+\frac{|q|^2}{\sigma}\right)}. \end{aligned} \quad (41)$$

We want to estimate $\varrho_h \triangleq |h|^2$, while the phase $\varphi \triangleq \arg(h)$ is not of interest. To achieve the conditional distribution $f(\varrho_h|y)$, we change variables in $f(h|y)$ to ϱ_h,φ (with the Jacobian 1/2) and marginalize the distribution by integrating over the phase φ:

$$\begin{aligned} f_{\varrho_h|y}(\varrho_h|y) &= \int_{-\pi}^{\pi} f_{h|y}\left(\sqrt{\varrho_h}(\cos\varphi+i\sin\varphi)|y\right)\frac{d\varphi}{2} \\ &= \frac{d}{2\pi}e^{-\varrho_h d}e^{-c\frac{\lambda\sigma}{|q|^2\lambda+\sigma}}\int_{-\pi}^{\pi}e^{2\sqrt{\varrho_h c}\cos(\varphi-\beta)}d\varphi \\ &= de^{-\varrho_h d}e^{-c\frac{\lambda\sigma}{|q|^2\lambda+\sigma}}I_0(2\sqrt{\varrho_h c}) \end{aligned} \quad (42)$$

where $c = |(a/\lambda)+(q^*(y-b)/\sigma)|^2$, $d = ((1/\lambda)+(|q|^2/\sigma))$, $\beta = \arg((a/\lambda)+((q^*(y-b))/\sigma))$, and $I_\nu(\cdot)$ is the modified Bessel function of the first kind. The last equality in (42) follows by applying the formula $\pi I_0(x) = \int_0^\pi e^{x\cos(\varphi)}d\varphi$ [33, Eq. 8.431.3]. The first and second order central moments of $f_{\varrho_h|y}(\varrho_h|y)$ are

$$\begin{aligned} \mathbb{E}\{\varrho_h|y\} &= \frac{\lambda\sigma}{|q|^2\lambda+\sigma}\left(1+\frac{c\sigma\lambda}{|q|^2\lambda+\sigma}\right), \\ V\{\varrho_h|y\} &= \left(\frac{\lambda\sigma}{|q|^2\lambda+\sigma}\right)^2\left(1+\frac{2c\sigma\lambda}{|q|^2\lambda+\sigma}\right), \end{aligned} \quad (43)$$

respectively. These moments follows from straightforward integration, by noting that $I_0(x) = \sum_{j=0}^{\infty}((x^2/4)^j/(j!)^2)$, $\int_0^\infty x^m e^{-Ax} = m!/A^{m+1}$ (for $m\in\mathbb{Z}$, $A>0$), and by identifying the Maclaurin expansion of e^x. The MSE is achieved by replacing c in the expression of $V\{\varrho_h|y\}$ with its average $(|q|^2/\sigma^2)(|q|^2\lambda+\sigma)+(|a|^2/\lambda^2\sigma^2)(|q|^2\lambda+\sigma)^2$.

In the MIMO case, observe that the elements of $\mathbf{U}_R^H\mathbf{Y}\mathbf{V}_Q$ are independent. Since the Frobenius norm is the sum of the squared magnitude of each element, we will have the sum of n_Tn_R independent variables that can be estimated separately. Thus, the MMSE estimate and MSE in (18) and (19) follows

from a MIMO transformation of (43), with c replaced with its average.

Proof of Theorem 5: A function is convex if and only if its second derivative is non-negative. The second derivative of the MSE in (19) with respect to p_j is

$$\frac{\partial^2 \mathrm{MSE}}{\partial p_j^2} = \sum_{l=1}^{n_R} \frac{4p_j \frac{\left(\lambda_j^{(T)}\lambda_l^{(R)}\right)^4}{\left(\sigma_j^{(Q)}\sigma_l^{(R)}\right)^3}\left(\lambda_j^{(T)}\lambda_l^{(R)} + |g_{lj}|^2\right)}{\left(p_j \frac{\lambda_j^{(T)}\lambda_l^{(R)}}{\sigma_j^{(Q)}\sigma_l^{(R)}} + 1\right)^4} + \sum_{l=1}^{n_R} \frac{2\frac{\left(\lambda_j^{(T)}\lambda_l^{(R)}\right)^3}{\left(\sigma_j^{(Q)}\sigma_l^{(R)}\right)^2}\left(2|g_{lj}|^2 - \lambda_j^{(T)}\lambda_l^{(R)}\right)}{\left(p_j \frac{\lambda_j^{(T)}\lambda_l^{(R)}}{\sigma_j^{(Q)}\sigma_l^{(R)}} + 1\right)^4} \quad (44)$$

which in general is negative in the neighborhood of $p_j = 0$ and thus the MSE is non-convex (for small values of p_j). If the condition for convexity in the theorem is fulfilled, all terms in the sum will however be positive at $p_j = 0$. Even if the MSE is non-convex, the KKT conditions give necessary condition for the optimal power allocation [31, Chapter 5.5]. By a straightforward Lagrangian approach, the power allocation that minimizes (19) needs to fulfill the stationarity conditions in (20).

At high training power, the necessary condition in (20) can be approximated and simplified as

$$\alpha \approx \sum_{l=1}^{n_R} \frac{2\sigma_j^{(Q)}\sigma_l^{(R)}\left(\lambda_j^{(T)}\lambda_l^{(R)} + |g_{lj}|^2\right)}{p_j^2} \quad (45)$$

which has the unique solution $p_j = \mathcal{P}\sqrt{c_j}/\sum_i \sqrt{c_i}$ for all j.

At low training power, the MSE in (19) can be approximated as

$$\mathrm{MSE} \approx \sum_{l=1}^{n_R}\sum_{j=1}^{n_T}\left(\lambda_j^{(T)}\lambda_l^{(R)}\right)^2 + 2|g_{lj}|^2\lambda_j^{(T)}\lambda_l^{(R)} - \sum_{l=1}^{n_R}\sum_{j=1}^{m} \frac{p_j\lambda_j^{(T)}\lambda_l^{(R)}}{\sigma_j^{(Q)}\sigma_l^{(R)}}\left(\lambda_j^{(T)}\lambda_l^{(R)} + 2|g_{lj}|^2\right) \quad (46)$$

using a first order Taylor expansions of the denominators and disregarding terms with p_j in the numerator. Hence, the MSE is minimized by allocating all the power to the p_j associated with the largest $\sum_{l=1}^{n_R}(\lambda_j^{(T)}\lambda_l^{(R)}/\sigma_j^{(Q)}\sigma_l^{(R)})(\lambda_j^{(T)}\lambda_l^{(R)} + 2|g_{lj}|^2)$. If there is multiplicity in the largest value of the sum, the power can be allocated freely among these eigendirections.

Proof of Corollary 3: The condition $\mathbf{R}_R = \mathbf{S}_R$ means that $\lambda_l^{(R)} = \sigma_l^{(R)}$ for all l, and therefore we can remove the dependence of l in the denominator of (20). For all active training powers ($p_j > 0$), the remaining expression in (20) can be formulated as a third degree polynomial equation in p_j: $2p_jb_j^2\gamma_j + 2b_j\nu_j(p_jb_j + 1) = \alpha(p_jb_j + 1)^3$, using the notation $b_j = \lambda_j^{(T)}/\sigma_j^{(Q)}$. Its three solutions ($k = -1, 0, 1$) are

$$p_j = \left(\frac{b_j\gamma_j}{\alpha}\right)^{1/3}\frac{e^{i\frac{2\pi}{3}k}}{b_j}A_j + \frac{2(\gamma_j+\nu_j)b_j^{2/3}}{3\alpha^{2/3}\gamma_j^{1/3}}\frac{e^{-i\frac{2\pi}{3}k}}{b_j}\frac{1}{A_j} - \frac{1}{b_j} \quad (47)$$

where $A_j = \left(-1 + \sqrt{1 - 8b_j(\gamma_j+\nu_j)^3/27\gamma_j^2\alpha}\right)^{1/3}$. Observe that this expression has the form $C_1|A_j|e^{i(\arg(A_j)+(2\pi/3)k)} + (C_2/|A_j|)e^{-i(\arg(A_j)+(2\pi/3)k)} - C_3$, where C_i are positive real-valued constants. Thus, in order for any of the solutions to be real-valued we need $C_1|A_j| = C_2/|A_j|$. If $\alpha > 8b_j(\gamma_j + \nu_j)^3/27\gamma_j^2$, this condition can be expressed as

$$\left(-1 + \sqrt{1 - \frac{8b_j(\gamma_j+\nu_j)^3}{27\gamma_j^2\alpha}}\right) = \frac{8b_j(\gamma_j+\nu_j)^3}{27\gamma_j^2\alpha} \quad (48)$$

which has no solutions in the interval. For all $0 < \alpha \leq 8b_j(\gamma_j + \nu_j)^3/27\gamma_j^2$, we observe that $|A_j| = (8b_j(\gamma_j+\nu_j)^3/27\gamma_j^2\alpha)^{1/6}$ which satisfies the condition $C_1|A_j| = C_2/|A_j|$. Thus, for these α we can rewrite (47) as

$$p_j = 2\Re\left\{\left(\frac{b_j\gamma_j}{\alpha}\right)^{1/3}\frac{e^{i\frac{2\pi}{3}k}}{b_j}A_j\right\} - \frac{1}{b_j} = 2\sqrt{\frac{2(\gamma_j+\nu_j)}{3b_j\alpha}}\cos\left(\frac{\pi}{3}(1+2k) - \frac{\phi_j}{3}\right) - \frac{1}{b_j} \quad (49)$$

where we used that $\arg(A_j) = (\pi - \phi_j)/3$ with ϕ_j defined as in the corollary. Since $\phi_j \in [0, \pi/2]$, $k = 1$ will only give negative solutions. For $k = -1, 0$, we see that the interval boundary $\alpha = 8b_j(\gamma_j+\nu_j)^3/27\gamma_j^2$ gives the coinciding solution $p_j = (\gamma_j - 2\nu_j)/2b_j(\gamma_j+\nu_j)$, while the limit $\alpha \to 0$ gives $p \to -(1/b_j)$ and $p \to \infty$, respectively. Thus, in order to show the intervals for the solutions, it remains to show that p_j is monotonically decreasing in α for $k = 0$ and increasing in α for $k = -1$. The derivative of p_j with respect to α can be expressed as

$$\frac{dp_j}{d\alpha} = -\sqrt{\frac{2(\gamma_j+\nu_j)}{3^3b_j\alpha^3}}\frac{\cos\left(\frac{\pi}{3}(1+2k) - \frac{\phi}{3}\right)}{\sqrt{\frac{8b_j(\gamma_j+\nu_j)^3}{27\gamma_j^2\alpha} - 1}} \times\left(\tan\left(\frac{\pi}{3}(1+2k) - \frac{\phi}{3}\right) + 3\sqrt{\frac{8b_j(\gamma_j+\nu_j)^3}{27\gamma_j^2\alpha} - 1}\right) \quad (50)$$

where the multiplicative term outside the brackets is positive for all α and k. The bracketed term can be expressed as $\tan(\pi(1+2k)/3 - \arctan(x)/3 + \tan(\arctan(3x))$ for $x = \sqrt{(8b_j(\gamma_j+\nu_j)^3/27\gamma_j^2\alpha) - 1}$. Then, the intervals follows from the observation that $\arctan(3x) - (\arctan(x) - \pi)/3 > 0$ and $\arctan(3x) - (\arctan(x) + \pi)/3 < 0$.

Finally, we see that the second derivative of the MSE in (44) is positive if we limit ourself to $p_j \geq \sigma_j^{(Q)}\sigma_l^{(R)}(\lambda_j^{(T)}\lambda_l^{(R)} - 2|g_{lj}|^2)/2\lambda_j^{(T)}\lambda_l^{(R)}(\lambda_j^{(T)}\lambda_l^{(R)} + |g_{lj}|^2)$, since then each term in the sum is positive. Thus, the MSE will be convex with respect to these p_j and the KKT conditions in (20) becomes necessary and sufficient. In the special case $\mathbf{R}_R = \mathbf{S}_R$, we can strengthen the condition since we know that the necessary KKT conditions only give a single feasible solution if $p_j \geq \max(\sigma_j^{(Q)}(\gamma_j - 2\nu_j)/2\lambda_j^{(T)}(\gamma_j+\nu_j), 0)$. In both the general and special case, these conditions need be combined with the original constraint $p_j \geq 0$.

ACKNOWLEDGMENT

The authors would like to thank the reviewers and the associate editor for their insightful comments and suggestions, which led to a more precise and readable paper.

REFERENCES

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Commun.*, vol. 6, pp. 311–335, 1998.
[2] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, pp. 585–595, 1999.
[3] D. Chizhik, J. Ling, P. Wolniansky, R. Valenzuela, N. Costa, and K. Huber, "Multiple-input-multiple-output measurements and modeling in Manhattan," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 3, pp. 321–331, 2003.
[4] 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, 2004.
[5] J. Wallace and M. Jensen, "Measured characteristics of the MIMO wireless channel," in *Proc. IEEE VTC'01-Fall*, 2001, vol. 4, pp. 2038–2042.
[6] K. Werner and M. Jansson, "Estimating MIMO channel covariances from training data under the Kronecker model," *Signal Process.*, vol. 89, pp. 1–13, 2009.
[7] F. Dietrich and W. Utschick, "Pilot-assisted channel estimation based on second-order statistics," *IEEE Trans. Signal Process.*, vol. 53, no. 3, 2005.
[8] 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, 2006.
[9] P. Jiyong, L. Jiandong, L. Zhuo, Z. Linjing, and C. Liang, "Optimal training sequences for MIMO systems under correlated fading," *J. Syst. Eng. Electron.*, vol. 19, pp. 33–38, 2008.
[10] X. Ma, L. Yang, and G. Giannakis, "Optimal training for MIMO frequency-selective fading channels," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 453–466, 2005.
[11] 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, 2004.
[12] 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, 2007.
[13] D. Katselis, E. Kofidis, and S. Theodoridis, "Training-based estimation of correlated MIMO fading channels in the presence of colored interference," *Signal Process.*, vol. 87, pp. 2177–2187, 2007.
[14] 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, 2003.
[15] J. Pang, J. Li, L. Zhao, and Z. Lü, "Optimal training sequences for MIMO channel estimation with spatial correlation," in *Proc. IEEE VTC'07-Fall*, 2007, pp. 651–655.
[16] E. Björnson and B. Ottersten, "Training-based Bayesian MIMO channel and channel norm estimation," in *Proc. IEEE ICASSP'09*, 2009, pp. 2701–2704.
[17] E. Björnson, D. Hammarwall, and B. Ottersten, "Exploiting quantized channel norm feedback through conditional statistics in arbitrarily correlated MIMO systems," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 4027–4041, 2009.
[18] X. Zhang, E. Jorswieck, and B. Ottersten, "User selection schemes in multiple antenna broadcast channels with guaranteed performance," presented at the IEEE SPAWC'07, Helsinki, Finland, Jun. 17–20, 2007.
[19] R. Ertel, P. Cardieri, K. Sowerby, T. Rappaport, and J. Reed, "Overview of spatial channel models for antenna array communication systems," *IEEE Personal Commun. Mag.*, vol. 5, pp. 10–22, 1998.
[20] E. Jorswieck and H. Boche, "Majorization and matrix-monotone functions in wireless communications," *Foundations and Trends in Communication and Information Theory*, vol. 3, pp. 553–701, 2007.
[21] E. Björnson, E. Jorswieck, and B. Ottersten, "Impact of spatial correlation and precoding design in OSTBC MIMO systems," *IEEE Trans. Wireless Commun.*, submitted for publication.
[22] A. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. Boston, MA: Academic Press, 1979.
[23] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.
[24] A. Tulino, A. Lozano, and S. Verdú, "Impact of antenna correlation on the capacity of multiantenna channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2491–2509, 2005.
[25] W. Weichselberger, M. Herdin, H. Özcelik, and E. Bonek, "A stochastic MIMO channel model with joint correlation of both link ends," *IEEE Trans. Wireless Commun.*, vol. 5, no. 1, pp. 90–100, 2006.
[26] W. Hager, Y. Liu, and T. Wong, "Optimization of generalized mean square error in signal processing and communication," *Linear Algebra and Its Applications*, vol. 416, pp. 815–834, 2006.
[27] D. Katselis, E. Kofidis, and S. Theodoridis, "On training optimization for estimation of correlated MIMO channels in the presence of multiuser interference," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4892–4904, 2008.
[28] E. Björnson and B. Ottersten, "Post-user-selection quantization and estimation of correlated Frobenius and spectral channel norms," presented at the IEEE PIMRC'08, Cannes, France, Sep. 15–18, 2008.
[29] V. Veeravalli, Y. Liang, and A. Sayeed, "Correlated MIMO wireless channels: Capacity, optimal signaling, and asymptotics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2058–2072, 2005.
[30] S. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Commun. Lett.*, vol. 5, pp. 369–371, 2001.
[31] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[32] W. Rudin, *Principles of Mathematical Analysis*. New York: McGraw-Hill, 1976.
[33] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*. Boston, MA: Academic Press, 1980.

Emil Björnson (S'07) was born in Malmö, Sweden, in 1983. He received the M.S. degree in engineering mathematics from Lund University, Lund, Sweden, in 2007. He is currently working towards the Ph.D. degree in telecommunications at the Signal Processing Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden.

His research interests include wireless communications, resource allocation, estimation theory, stochastic signal processing, and mathematical optimization.

For his work on MIMO communications, he received a Best Paper Award at the 2009 International Conference on Wireless Communications and Signal Processing (WCSP 2009).

Björn Ottersten (S'87-M'89-SM'99-F'04) was born in Stockholm, Sweden, in 1961. He received the M.S. degree in electrical engineering and applied physics from Linköping University, Linköping, Sweden, in 1986 and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1989.

He has held research positions at the Department of Electrical Engineering, Linköping University; the Information Systems Laboratory, Stanford University; and the Katholieke Universiteit Leuven, Leuven, Belgium. During 1996–1997, he was Director of Research at ArrayComm Inc., San Jose, CA, a start-up company based on Ottersten's patented technology. In 1991, he was appointed Professor of Signal Processing at the Royal Institute of Technology (KTH), Stockholm, Sweden. From 2004 to 2008, he was Dean of the School of Electrical Engineering at KTH, and from 1992 to 2004 he was head of the Department for Signals, Sensors, and Systems at KTH. He is also Director of security and trust at the University of Luxembourg. His research interests include wireless communications, stochastic signal processing, sensor array processing, and time-series analysis.

Dr. Ottersten has coauthored papers that received an IEEE Signal Processing Society Best Paper Award in 1993, 2001, and 2006. He has served as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and on the Editorial Board of the IEEE *Signal Processing Magazine*. He is currently Editor-in-Chief of the EURASIP *Signal Processing Journal* and a member of the Editorial Board of the EURASIP *Journal of Advances in Signal Processing*. He is a Fellow of EURASIP. He is one of the first recipients of the European Research Council advanced research grant.