HAL
open science

Optimal Training Sequences for Large-Scale MIMO-OFDM Systems

Zhichao Sheng, Hoang Duong Tuan, Ha H. Nguyen, Merouane Debbah

To cite this version:

Zhichao Sheng, Hoang Duong Tuan, Ha H. Nguyen, Merouane Debbah. Optimal Training Sequences for Large-Scale MIMO-OFDM Systems. IEEE Transactions on Signal Processing, 2017, 65 (13), pp.3329-3343. 10.1109/TSP.2017.2688978 . hal-02305669

HAL Id: hal-02305669

https://centralesupelec.hal.science/hal-02305669v1

Submitted on 18 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Optimal Training Sequences for Large-Scale MIMO-OFDM Systems

Zhichao Sheng, Hoang Duong Tuan, Ha H. Nguyen, *Senior Member, IEEE*, and Merouane Debbah, *Fellow, IEEE*

***Abstract*—This paper considers the optimal design of training sequences for channel estimation in *large-scale* multiple-input multiple-output orthogonal frequency-division multiplexing systems. The application scenario of interest is when the number of transmit antennas for the downlink (or the number of receive antennas for the uplink) is large, but not large enough to benefit the asymptotical optimality of using equipower training sequences (e.g., due to practical constraints on deployment costs, space, and antenna size). Under the criterion of minimizing the mean square error of the channel estimate, the optimal design of training sequences for such systems poses a truly large-scale optimization problem, to which existing optimization solvers are not applicable. We develop a fast convex programming (FCP) procedure to find its global optimal solution. In each iteration of the proposed FCP procedure, a solution is found in a scalable and closed form. The singularity and ill-conditionedness of the channel correlation matrices are also exploited to improve the computation efficiency. Furthermore, we also examine the design of reduced-length training sequences and develop a successive quadratic programming procedure to find the solutions. Intensive simulation results are provided to illustrate the performance of our methods.**

***Index Terms*—Large-scale MIMO, OFDM, spatial correlation, training sequences, reduced-length sequence, minimum mean square error (MMSE), large-scale optimization, fast convex programming, successive quadratic programming.**

I. Introduction

MULTI-INPUT MULTI-OUTPUT (MIMO)-orthogonal frequency division multiplexing (OFDM) is adopted in many wireless standards thanks to its implementation flexibility and robustness against multipath fading channels. Nevertheless, new technologies are still required to meet the exponentially-growing demand for wireless data traffic. Recently, large-scale (LS) MIMO configuration [1] has been proposed as one of the key candidate technologies for the continuing evolution toward beyond-4G and 5G cellular systems. In a nutshell, a large-scale MIMO transceiver deploys a large number of antennas that are placed in a two-dimensional (2D) planar array as opposed to a one-dimensional (1D) linear array as in a traditional MIMO setup. The chief benefit of 2D antenna arrangement is to reduce the form factor of the antennas, hence making it more practical [2]. For example, deploying 64 antennas at 5 GHz in an 8×8 planar array with a half-wavelength spacing would result in a form factor of 0.25 m $\times$ 0.25 m, which is much smaller than a linear array of 2 m wide would the antennas be placed in a linear array. In addition to the small form factor advantage, the large-scale MIMO setup allows the base station (BS) to transmit/receive spatially-multiplexed signals to/from a large number of user terminals. This is because with 2D antenna elements, dynamic and adaptive precoding/beamforming can be performed jointly across all antennas. As a result, the base station can realize more directed transmissions in the azimuth and elevation domains simultaneously for a larger number of user terminals.

This paper is concerned with training sequence design for channel state estimation (CSE) in a large-scale MIMO OFDM system. CSE is necessary to perform precoding and beamforming at the transmitter and coherent detection at the receiver. This allows practical systems to operate close to the theoretical capacity [3]. Compared to single-input single-output (SISO) systems, the problem of CSE in large-scale MIMO systems is much more difficult because there is a much larger number of correlated variables to be estimated (the number of variables is proportional to the number of transmit-receive antenna pairs and their associated channel delay spreads [4]). CSE for flat fading MIMO systems has been extensively studied in the literature either for small-scale systems (see e.g. [5]–[9] and references therein) or massive MIMO systems with a few hundred transmit/receive antennas (see e.g. [10]–[15]). For broadband wireless systems, which experience frequency-selective fading, it is common to use OFDM so that the frequency-selective fading channel is turned into parallel flat fading sub-channels. However, it is not efficient to estimate these equivalent large-scale MIMO flat fading sub-channels separately because (i) they are not independent but rather correlated in time, frequency and space, and (ii) the total number of channel variables is too large. In fact, even for small-scale MIMO-OFDM systems, the CSE problem and solutions are not a simple extension of the techniques developed for flat fading MIMO channels [16], [17].

Manuscript received December 6, 2016; revised March 12, 2017; accepted March 14, 2017. Date of publication March 29, 2017; date of current version April 27, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Itsik Bergel. This work was supported in part by the Australian Research Councils Discovery Projects under Project DP130104617, in part by the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering), and in part by an NSERC Discovery Grant. *(Corresponding author: Ha H. Nguyen.)*

Z. Sheng and H. D. Tuan are with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: kebon22@163.com; tuan.hoang@uts.edu.au).

H. H. Nguyen is the Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK S7N 5C5, Canada (e-mail: ha.nguyen@usask.ca).

M. Debbah is with Large Networks and Systems Group, CentraleSuplec, Universit Paris-Saclay, Paris 91190, France, and also with the Mathematical and Algorithmic Sciences Lab, Huawei France R&D, Paris 91190, France (e-mail: merouane.debbah@supelec.fr).

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.2017.2688978

Many channel state estimators proposed in the past for (small-scale) MIMO-OFDM simplify the problem by ignoring the spatial correlations among transceiver's antennas [16], [18]. The first estimator for spatially-correlated MIMO-OFDM channels was developed in [17] but the proposed training sequence and estimator are locally optimized at extreme SNR regimes only. The globally-optimized training sequences and estimator were thoroughly addressed in [19], [20] for all SNR regimes. The CSE problem faced in large-scale MIMO-OFDM systems is a highly-structured large-scale matrix problem, which cannot be easily factorized for tractably-computational solutions. Whenever the number of antennas is not more than 64 as in many current and future systems [1], the CSE solutions for massive MIMO systems (with several hundred antennas [10], [11], [13], [21]) may not apply. It should be pointed out that spatial correlations of MIMO-OFDM channels have been exploited in [22] for the estimation of path delays.

The present paper adopts the minimum mean square error (MMSE) criterion in optimizing the training sequences for CSE of large-scale MIMO-OFDM systems. In connection to a similar problem formulation in [19], [20] for small-scale MIMO-OFDM systems, the MMSE-based CSE problem for large-scale MIMO-OFDM systems at hand poses the following completely new issues:

- The channel correlation matrices are singular and ill-conditioned due to close antenna spacing [11], [17]. This prevents the application of the matrix inversion lemma, which is the main tool for formulating the training sequence design in tractable optimization.
- The semi-definite programming (SDP) based solution of [19], [20] for small-scale MIMO-OFDM systems becomes computationally prohibitive for large-scale MIMO-OFDM systems due to large dimensions of training sequences. Efficient procedures for solving large-scale SDP have not been developed.
- Optimization for training sequences of reduced length is practically relevant for large-scale MIMO-OFDM as it leads to higher data throughput, but is a highly nonconvex optimization problem, whose solution is not yet explored, even for small-scale MIMO-OFDM systems.

The aim of the present paper is to resolve the aforementioned issues of CSE for large-scale MIMO-OFDM systems. The main contributions of the paper are elaborated in the following.

- Bypassing the need of matrix inverse (which does not exist) or pseudo-inverse (which is ill-conditioned) of the channel correlation matrices, we first obtain a convex optimization formulation for the problem of MSE minimization. Moreover, the singularity and ill-conditionedness of the channel correlation matrices are also exploited for this tractable formulation. It also reveals that the number of antennas to transmit the optimal training sequences does not excess the maximal rank of transmission correlation matrices. In fact this development is new even for small-scale MIMO-OFDM systems.
- Since the resulting convex optimization problem is large-scale, for which there is no computationally-feasible solution procedure to date, we develop a scalable fast convex programming (FCP) to find its global optimal solution. This development is new even from the optimization viewpoint.
- When seeking training sequences of a reduced length, the optimization problem is no longer convex but is still large-scale. Nevertheless, we develop an efficient successive quadratic programming (SQP) for its computational solution. Each iterative convex program admits a closed-form solution so the iterative procedure is very efficient. The full length of training sequences is LM_t where L is the delay spread and M_t is the number of antennas in downlink transmission, which is large for large M_t. We will show that a very good CSE can be obtained by optimizing the training sequences with the reduced length of $LM_t/2$.

The rest of the paper is organized as follows. Section II introduces the large-scale MIMO-OFDM system model and formulates the optimization problem of CSE under the MMSE criterion. Section III shows that the considered optimization problem can still be cast into a convex program, which is however large-scale. Fast convex programming for solving this large-scale convex program is developed in Section IV. Successive quadratic programming for efficiently solving the large-scale, highly nonconvex program of optimizing reduced-length training sequences is given in Section V. Section VI presents examples to illustrate the performance of our proposed methods. Concluding remarks are given in Section VII.

Notation: Bold capital and lower-case letters denote matrices and column vectors, respectively. The subscripts $(\cdot)^T$ and $(\cdot)^H$ denote transpose and Hermitian transpose operations, respectively. The symbol $\otimes$ is used for the Kronecker product of two matrices and $\mathsf{vec}(\mathbf{A})$ denotes the vectorization operation of matrix $\mathbf{A}$. $\mathsf{tr}(\cdot)$ and $|\cdot|$ stand for the trace and determinant of a matrix, respectively. Accordingly, $\langle \mathbf{A}, \mathbf{B} \rangle = \mathsf{tr}(\mathbf{A}^H \mathbf{B})$ and $\mathbf{A} \otimes \mathbf{B}$ are the dot product and Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$. $\mathbf{A}^+$ is pseudo-inverse of $\mathbf{A}$. $\mathbf{A} \succ 0$ ($\mathbf{A} \succeq 0$, resp.) means $\mathbf{A}$ is Hermitian positive definite (semi-definite, resp.), for which $\mathbf{A}^{1/2}$ is a matrix such that $\mathbf{A}^{1/2}(\mathbf{A}^{1/2})^H = \mathbf{A}$, while $\mathbf{A}_+^{1/2}$ is the unique Hermitian symmetric positive definite matrix such that $\mathbf{A}_+^{1/2}\mathbf{A}_+^{1/2} = \mathbf{A}$. $\mathbf{I}_n$ is the identity matrix of dimension $n \times n$. $x^+ = \max\{x, 0\}$ for a scalar x. The expectation operation is $\mathsf{E}\{\cdot\}$, while $\mathcal{CN}(0, \sigma^2)$ denotes a zero-mean circularly-symmetric complex Gaussian random variable with variance σ^2. $W_N = \mathrm{e}^{-\jmath 2\pi/N}$ for an integer N. Furthermore, $[\mathbf{A}_{ij}]_{i,j=0,1,\ldots,N}$ with matrices $\mathbf{A}_{ij}$ means the matrix with block entries $\mathbf{A}_{ij}$. Analogously, $\mathrm{diag}[\mathbf{A}_i]_{i=0,1,\ldots,N}$ means the matrix with diagonal blocks $\mathbf{A}_i$ and zero off-diagonal blocks.

II. Large-Scale MIMO-OFDM System Model and Training Design

Consider a large-scale MIMO-OFDM system with M_t transmit antennas, M_r receive antennas as depicted in Fig. 1, where either M_t is large (e.g., up to 64) for downlink transmission or M_r is large for an uplink receiver. The MIMO frequency-selective fading channel can be described by the following

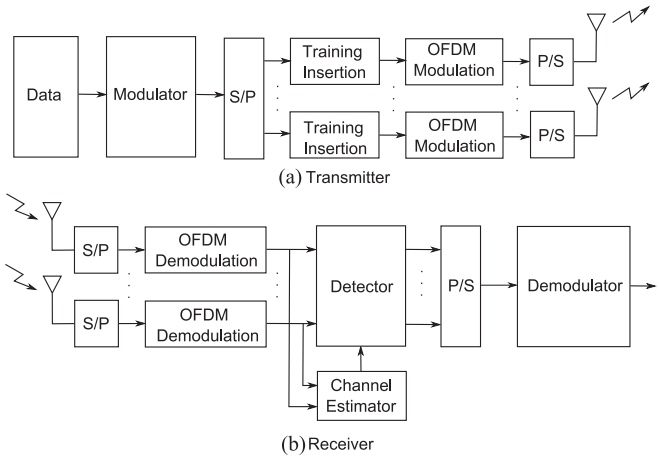

Fig. 1. A typical MIMO-OFDM system.

transfer matrix:

$$\mathbf{H}(z) = \sum_{\ell=0}^{L-1} \mathbf{H}_\ell z^{-\ell},$$

where each stationary process $\mathbf{H}_\ell \in \mathbb{C}^{M_r \times M_t}$ represents the gains of the ℓth MIMO path. It is noted that the matrix $\mathbf{H}_\ell$ is either very fat (when M_t is large) or very tall (when M_r is large). Under the assumption of Rayleigh block fading, the elements of $\mathbf{H}_\ell$ are circularly symmetric complex Gaussian random variables that remain unchanged over the period of one OFDM symbol.

Due to close antenna spacings in 2D planar array, large-scale MIMO channels exhibit spatial correlations. The spatial correlations of MIMO channels can be modeled as follows [23]–[25]:

$$\mathbf{H}_\ell = \mathbf{R}_{\mathrm{R},\ell}^{1/2}\, \mathbf{H}_{\mathrm{W},\ell}\, (\mathbf{R}_{\mathrm{T},\ell}^{1/2})^T, \quad \ell = 0, 1, \ldots, L-1, \quad (1)$$

where the deterministic Hermitian-symmetric matrix $0 \preceq \mathbf{R}_{\mathrm{T},\ell} \in \mathbb{C}^{M_t \times M_t}$ models the correlation among the transmit antenna elements, which is typically singular [11], [17] with

$$\mathrm{rank}(\mathbf{R}_{\mathrm{T},\ell}) = \tau_\ell, \quad (2)$$

which is less than M_t, while $0 \preceq \mathbf{R}_{\mathrm{R},\ell} = \mathbf{R}_{\mathrm{R},\ell}^{1/2}\mathbf{R}_{\mathrm{R},\ell}^{1/2} \in \mathbb{C}^{M_r \times M_r}$ captures the correlation among the antenna elements at the receiver. The matrix $\mathbf{H}_{\mathrm{W},\ell} \in \mathbb{C}^{M_r \times M_t}$ is a stationary process, whose elements are independent and identical distributed circularly symmetric complex Gaussian random variables with unit variance. Define $\mathbf{h} = (\mathrm{vec}^T(\mathbf{H}_0), \mathrm{vec}^T(\mathbf{H}_1), \ldots, \mathrm{vec}^T(\mathbf{H}_{L-1}))^T \in \mathbb{C}^{LM_tM_r}$. Then the correlation matrix of $\mathbf{h}$ is

$$\mathbf{R}_\mathbf{h} = E[\mathbf{h}\mathbf{h}^H] = \mathsf{diag}[\mathbf{R}_{\mathrm{T},\ell} \otimes \mathbf{R}_{\mathrm{R},\ell}]_{\ell=0,\ldots,L-1}. \quad (3)$$

Suppose the MIMO-OFDM system operates with $M = 2^{\mathcal{M}}$ sub-carriers. Each information block of length M goes through an OFDM modulator to form an OFDM block and is transmitted via one transmit antenna. The OFDM cyclic prefix (CP) length is chosen to be longer than the channel order, $L-1$, to avoid inter-block interference (IBI). Specifically, sequences $\mathbf{x}(j) \in \mathbb{C}^{M_t},\ j = 0, 1, \ldots, M-1$ are transmitted from M_t antennas on the jth sub-carrier. Inside just one OFDM block, $N = 2^{\mathcal{N}} << M$ training symbols are inserted on the t_0th, t_1th,..., t_{N-1}th sub-carriers for channel estimation. Mathematically, the training sequences are inserted as

$$\mathbf{s}(t_k) = \mathbf{x}(t_k) \in \mathbb{C}^{M_t},\ k = 0, 1, \ldots, N-1.$$

The channel transfer function corresponding to the t_kth sub-channel is

$$\mathbf{H}_f(t_k) := \sum_{\ell=0}^{L-1} \mathbf{H}_\ell W_M^{\ell t_k},\ k = 0, 1, \ldots, N-1. \quad (4)$$

Thus, the normalized input-output equation for each pilot sub-carrier is

$$\mathbf{r}(t_k) = \sqrt{\frac{\rho}{M_t}}\mathbf{H}_f(t_k)\mathbf{s}(t_k) + \mathbf{n}(t_k),\ k = 0, 1, \ldots, N-1, \quad (5)$$

where ρ is the average training signal-to-noise-ratio (SNR) at the receiver, $\mathbf{r}(t_k) = (r_0(t_k), \ldots, r_{M_r-1}(t_k))^T \in \mathbb{C}^{M_r}$ is the t_kth received signal vector, $\mathbf{s}(t_k) = (s_0(t_k), \ldots, s_{M_t-1}(t_k))^T \in \mathbb{C}^{M_t}$ is the training vector, and $\mathbf{n}(t_k) = (n_0(t_k), \ldots, n_{M_r-1}(t_k))^T$ represents additive white Gaussian noise (AGWN), whose elements are i.i.d $\mathcal{CN}(0,1)$ random variables. It is pointed out that, in practice, ρ depends on the effects of path loss and shadowing (or shadow fading). The path loss is a function of distance between the transmitter and receiver, while shadowing represents the variation of received signal power due to blockages from objects in the signal path [26]. The effect of shadow fading differs from multipath fading captured by (4) in an important way: the former occurs at a much slower time-scale compared to multipath fading (the duration of a shadow fade could last for multiple seconds or minutes) [27]. As such, it is common in the design of estimation and detection algorithms for wireless communication systems to treat ρ as a constant. Furthermore, most modern communication systems exercise power control schemes that adapt to the slow variation of shadowing so that the quantity ρ is maintained at a certain level at the receiver [26], [27].

From Equation (5), we can write the received signal in the training phase as

$$\mathbf{r}(t_k) = \sqrt{\frac{\rho}{M_t}}\mathbf{M}(\mathbf{s}(t_k))\mathbf{h} + \mathbf{n}(t_k), \quad k = 0, 1, \ldots, N-1, \quad (6)$$

where $\mathbf{M}(\mathbf{s}(t_k))$ is defined as

$$\mathbf{M}(\mathbf{s}(t_k)) = [W_M^{t_k \cdot 0}\mathbf{s}^T(t_k) \quad W_M^{t_k \cdot 1}\mathbf{s}^T(t_k) \quad \cdots \quad W_M^{t_k(L-1)}\mathbf{s}^T(t_k)] \otimes \mathbf{I}_{M_r}. \quad (7)$$

It can be seen from (5) that there are M_rN measurements for the estimation of LM_tM_r unknown parameters, which are the entries of matrices $\mathbf{H}_\ell \in \mathbb{C}^{M_r \times M_t},\ \ell = 0, 1, \ldots, L-1$. When all the entries of $\mathbf{H}_\ell$ are *independent*, to make the estimation problem meaningful, it is widely known that the number of measurements is not less than the number of unknowns [6], i.e.,

$$N \geq LM_t, \quad (8)$$

which is large whenever M_t is large. The above condition on the number of sub-carriers necessary for OFDM channel estimation is also stated in [16], [17], [28]. We shall see in the next section that $N = LM_t$ is sufficient. Moreover, as the entries of $\mathbf{H}_\ell$ are correlated, we will also see that $N = LM_t/2$ when M_t is large, or $N = 2L$ when M_r is large, can still provide good CSE quality, hence improving the efficiency of the training phase.

Define the following training symbol matrix

$$\mathbf{S} = [\mathbf{s}(t_0) \quad \mathbf{s}(t_1) \quad \dots \quad \mathbf{s}(t_{N-1})]^T \in \mathbb{C}^{N\times M_t}. \tag{9}$$

Then (6) can be compactly represented as

$$\mathbf{r} = \sqrt{\frac{\rho}{M_t}}\mathbf{M}(\mathbf{S})\mathbf{h} + \mathbf{n}, \tag{10}$$

where

$$\mathbf{r} = \begin{pmatrix} \mathbf{r}(t_0) \\ \mathbf{r}(t_1) \\ \dots \\ \mathbf{r}(t_{N-1}) \end{pmatrix} \in \mathbb{C}^{NM_r},\ \mathbf{n} = \begin{pmatrix} \mathbf{n}(t_0) \\ \mathbf{n}(t_1) \\ \dots \\ \mathbf{n}(t_{N-1}) \end{pmatrix} \in \mathbb{C}^{NM_r},$$

$$\begin{aligned}\mathbf{M}(\mathbf{S}) &= \begin{bmatrix} \mathbf{M}(\mathbf{s}(t_0)) \\ \mathbf{M}(\mathbf{s}(t_1)) \\ \dots \\ \mathbf{M}(\mathbf{s}(t_{N-1})) \end{bmatrix} \\ &= \mathcal{F}(\mathbf{S}) \otimes \mathbf{I}_{M_r} \in \mathbb{C}^{NM_r \times LM_t M_r}, \\ \mathcal{F}(\mathbf{S}) &= \begin{bmatrix} \mathbf{F}_0\mathbf{S} & \mathbf{F}_1\mathbf{S} & \dots & \mathbf{F}_{L-1}\mathbf{S} \end{bmatrix} \in \mathbb{C}^{N\times LM_t}\end{aligned}$$

and

$$\begin{aligned}\mathbf{F}_\ell &= \mathrm{diag}\{W_M^{t_k \ell}\}_{k=0,1,\dots,N-1}, \\ \ell &= 0, 1, \dots, L-1, \\ \mathbf{M}^H(\mathbf{S})\mathbf{M}(\mathbf{S}) &= [(\mathbf{S}^H\mathbf{F}_\ell^H\mathbf{F}_m\mathbf{S}) \otimes \mathbf{I}_{M_r}]_{\ell,m=0,1,\dots,L-1} \\ &= [(\mathbf{S}^H\mathbf{F}_{m-\ell}\mathbf{S}) \otimes \mathbf{I}_{M_r}]_{\ell,m=0,1,\dots,L-1}.\end{aligned} \tag{11}$$

The CSE problem is to obtain an estimator $\hat{\mathbf{h}}$ for $\mathbf{h}$ from (10) with a known (deterministic) training signal $\mathbf{S}$. Since all the random vectors $\mathbf{r}, \mathbf{h}$ and $\mathbf{n}$ in Equation (10) are Gaussian with zero mean, $\hat{\mathbf{h}}$ is a Wiener filter (or MMSE estimator), which is the mean of a conditional distribution of $\mathbf{h}$ [29]:

$$\hat{\mathbf{h}} = \sqrt{\frac{\rho}{M_t}}\mathbf{R}_\mathbf{h}\mathbf{M}^H(\mathbf{S})\left(\frac{\rho}{M_t}\mathbf{M}(\mathbf{S})\mathbf{R}_\mathbf{h}\mathbf{M}^H(\mathbf{S}) + \mathbf{I}_{NM_r}\right)^{-1}\mathbf{r}. \tag{12}$$

Obviously, the error $\mathbf{e} = \mathbf{h} - \hat{\mathbf{h}}$ is a Gaussian random variable with zero mean and covariance

$$\mathbf{R}_\mathbf{e} := \mathbf{R}_\mathbf{h} - \frac{\rho}{M_t}\mathbf{R}_\mathbf{h}\mathbf{M}^H(\mathbf{S})\left(\frac{\rho}{M_t}\mathbf{M}(\mathbf{S})\mathbf{R}_\mathbf{h}\mathbf{M}^H(\mathbf{S}) + \mathbf{I}_{NM_r}\right)^{-1}\mathbf{M}(\mathbf{S})\mathbf{R}_\mathbf{h}. \tag{13}$$

In general, training design is to find the training matrix $\mathbf{S} \in \mathbb{C}^{N\times M_t}$ to obtain the conditional mean $\hat{\mathbf{h}}$ of the channel state $\mathbf{h}$ (which is considered as time-varying stationary process) under some criterion and subject to the following training power constraint:

$$\mathsf{tr}\{\mathbf{S}^H\mathbf{S}\} = P_t. \tag{14}$$

The commonly-used estimation criteria are minimum least square error [18], minimum mean squared error $\mathsf{E}\{||\mathbf{e}||^2\} = \mathsf{tr}(\mathbf{R}_\mathrm{e})$ [17], [19], [20] or minimum error-entropy $H(\mathbf{e})$ [5]–[7], [30]. The crucial assumption in all these works is the non-singularity of the covariance matrix $\mathbf{R}_\mathbf{h}$, under which the Matrix Inversion Lemma can be applied to simplify $\mathbf{R}_\mathbf{e}$ in (13) to

$$\mathbf{R}_\mathbf{e} = \left(\mathbf{R}_\mathbf{h}^{-1} + \frac{\rho}{M_t}\mathbf{M}^H(\mathbf{S})\mathbf{M}(\mathbf{S})\right)^{-1}. \tag{15}$$

For singular $\mathbf{R}_\mathbf{h}$ as the case in this paper, as pointed out in [20], [30] the Matrix Inverse Lemma is no longer applicable, i.e., one cannot simply replace $\mathbf{R}_\mathbf{h}^{-1}$ in (15) by its pseudo-inverse $\mathbf{R}_\mathbf{h}^{+}$ as done in [7] for MIMO channel estimation and [17] for MIMO-OFDM channel estimation. There are also other positive but small eigenvalues of $\mathbf{R}_\mathbf{h}$ that make (15) ill-conditioned.

To overcome these issues, we first derive $\mathbf{R}_\mathbf{e}$ in a tractable form that does not involve either $\mathbf{R}_\mathbf{h}^{-1}$ (which does not exist) or $\mathbf{R}_\mathbf{h}^{+}$ (which is often ill-conditioned) and is always well-posed. This is done by using the matrix inverse lemma [31] as follows:

$$\begin{aligned}\mathbf{R}_\mathbf{e} =\ & \mathbf{R}_\mathbf{h}^{1/2}\left(\mathbf{I}_{\tau_\mathrm{T} M_r} - \frac{\rho}{M_t}\left(\mathbf{R}_\mathbf{h}^{1/2}\right)^H\mathbf{M}^H(\mathbf{S})\right. \\ & \times\left(\frac{\rho}{M_t}\mathbf{M}(\mathbf{S})\mathbf{R}_\mathbf{h}^{1/2}\left(\mathbf{R}_\mathbf{h}^{1/2}\right)^H\mathbf{M}^H(\mathbf{S}) + \mathbf{I}_{NM_r}\right)^{-1} \\ & \left.\times\ \mathbf{M}(\mathbf{S})\mathbf{R}_\mathbf{h}^{1/2}\right)\left(\mathbf{R}_\mathbf{h}^{1/2}\right)^H \\ =\ & \mathbf{R}_\mathbf{h}^{1/2}\left(\mathbf{I}_{\tau_\mathrm{T} M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_\mathbf{h}^{1/2}\right)^H\mathbf{M}^H(\mathbf{S})\mathbf{M}(\mathbf{S})\mathbf{R}_\mathbf{h}^{1/2}\right)^{-1} \\ & \times\left(\mathbf{R}_\mathbf{h}^{1/2}\right)^H,\end{aligned} \tag{16}$$

where $\tau_\mathrm{T} = \sum_{\ell=0}^{L-1}\tau_\ell$ and $\mathbf{R}_\mathbf{h}^{1/2} = \mathsf{diag}[\mathbf{R}_{\mathrm{T},\ell}^{1/2} \otimes \mathbf{R}_{\mathrm{R},\ell+}^{1/2}]_{\ell=0,\dots,L-1}$ with

$$\mathbf{R}_{\mathrm{T},\ell}^{1/2} \in \mathbb{C}^{M_t\times\tau_\ell}. \tag{17}$$

From now on, we consider the following optimization problem:

$$\begin{aligned}\min_{\mathbf{S}\in\mathbf{C}^{N\times M_t}}\ & \mathsf{tr}\left(\mathbf{R}_\mathbf{h}^{1/2}\left(\mathbf{I}_{\tau_\mathrm{T} M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_\mathbf{h}^{1/2}\right)^H\mathbf{M}^H(\mathbf{S})\mathbf{M}(\mathbf{S})\right.\right. \\ & \left.\left.\times\mathbf{R}_\mathbf{h}^{1/2}\right)^{-1}\left(\mathbf{R}_\mathbf{h}^{1/2}\right)^H\right) \\ \text{s.t.}\ & \mathsf{tr}\{\mathbf{S}^H\mathbf{S}\} = P_t.\end{aligned} \tag{18}$$

One can see that the singularity of the correlation matrices $\mathbf{R}_{\mathrm{T},\ell}$ has been exploited in dimension reduction for the optimization formulation. In general, the optimization problem (18) is a difficult nonconvex problem because of the Kronecker product structures of matrices $\mathbf{R}_\mathbf{h}$ and $\mathbf{M}(\mathbf{S})$. In the next section we

show that (18) can still be recast as a convex optimization problem. However, unlike small-scale convex optimization problems considered in [19], [20], this convex optimization problem for large-scale MIMO-OFDM systems is computationally prohibitive for existing solvers. Therefore, Section IV develops a very fast computation procedure for its solution.

III. Optimized Training Sequence by Large-Scale Convex Optimization

Inside an OFDM block, $N = 2^{\mathcal{N}} << M = 2^{\mathcal{M}}$ training symbols are inserted on the 0th, M/Nth,..., $(N-1)M/N$th subcarriers for channel estimation. Mathematically, the training symbols are inserted as

$$\mathbf{s}(k) = \mathbf{x}(kM/N) \in \mathbb{C}^{M_t},\ k = 0, 1, \ldots, N-1.$$

The kM/Nth sub-channel in (4) is

$$\begin{aligned}\mathbf{H}_f(kM/N) :=\ & \mathbf{H}\left(W_M^{kM/N}\right)\\ =\ & \sum_{\ell=0}^{L-1} \mathbf{H}_\ell W_M^{\ell kM/N}\\ =\ & \sum_{\ell=0}^{L-1} \mathbf{H}_\ell W_N^{\ell k},\\ k =\ & 0, 1, \ldots, N-1. \end{aligned} \tag{19}$$

Then $\mathbf{M}(\mathbf{s}(kM/N))$ in (6) is

$$\mathbf{M}(\mathbf{s}(kM/N)) = \Big[W_N^{k\cdot 0}\mathbf{s}^T(k) \quad W_N^{k\cdot 1}\mathbf{s}^T(k) \quad \cdots \quad W_N^{k(L-1)}\mathbf{s}^T(k)\Big] \otimes \mathbf{I}_{M_r}. \tag{20}$$

The matrices in (11) are simplified to

$$\begin{aligned}\mathcal{F}(\mathbf{S}) =\ & [\mathbf{F}_0\mathbf{S} \quad \mathbf{F}_1\mathbf{S} \quad \ldots \quad \mathbf{F}_{L-1}\mathbf{S}],\\ \mathbf{F}_\ell =\ & \mathrm{diag}\{W_N^{k\ell}\}_{k=0,1,\ldots,N-1},\\ \ell =\ & 0, 1, \ldots, L-1,\\ \mathbf{M}^H(\mathbf{S})\mathbf{M}(\mathbf{S}) =\ & [(\mathbf{S}^H\mathbf{F}_\ell^H\mathbf{F}_m\mathbf{S}) \otimes \mathbf{I}_{M_r}]_{\ell,m=0,1,\ldots,L-1}. \end{aligned} \tag{21}$$

As pointed out before, the Kronecker product structures of matrices $\mathbf{R_h}$ and $\mathbf{M}(\mathbf{S})$, which are due to the spatially correlated channels and OFDM modulation, render the difficulty in solving the optimization problem in (18). To overcome this challenge, we shall make use of the following lemmas on the inequality of a positive definite matrix and orthogonality of the roots of unity.

Lemma 1: [31] For a positive definite block matrix $\mathbf{X} = [\mathbf{A}_{ij}]_{i,j=1,2,\ldots,L}$ with matrix entries $\mathbf{A}_{ij}$ and its inverse $\mathbf{X}^{-1}$ with matrix entries $[\mathbf{X}^{-1}]_{i,j=1,2,\ldots,L}$, the following matrix inequality holds true

$$[\mathbf{X}^{-1}]_{i,i} \succeq \mathbf{A}_{ii}^{-1}, i = 1, \ldots, L, \tag{22}$$

with the equality if and only if the off-diagonal entries $\mathbf{A}_{ij}, i \neq j$ are zeros, i.e., $\mathbf{X} = \mathrm{diag}[\mathbf{A}_{ii}]_{i=1,2,\ldots,L}$.

Lemma 2: [32] It is true that $W_N^{kN+i} = W_N^i$ and

$$\frac{1}{N}\sum_{k=0}^{N-1} W_N^{k\ell} = \begin{cases} 1, & \text{for } \ell \equiv 0 \pmod{N},\\ 0, & \text{otherwise.}\end{cases}$$

Using Lemma 1, the following theorem shows that $N = LM_t$ is the optimal number of training symbols since the estimation quality under the same power constraint is identical to that with using training sequences of longer lengths.

Theorem 1: Suppose that $\mathbf{Q} \in \mathbb{C}^{N\times M_t}$ is composed of M_t orthonormal columns $\mathbf{q}_i \in \mathbb{C}^N,\ i = 1, \ldots, M_t$ such that

$$\mathbf{Q}^H\mathbf{F}_\alpha\mathbf{Q} = [\mathbf{q}_i^H\mathbf{F}_\alpha\mathbf{q}_j]_{i,j=1,2,\ldots,M_t} = \mathbf{0}_{M_t} \quad \text{for } 1 \le \alpha \le L-1. \tag{23}$$

Then the optimization problem in (18) in $\mathbf{S} \in \mathbb{C}^{N\times M_t}$ is equivalent to the following convex optimization problem in $\mathbf{X} \in \mathbb{C}^{M_t\times M_t}$:

$$\mathcal{L}_\star = \min_{0\preceq\mathbf{X}\in\mathbf{C}^{M_t\times M_t}} \sum_{\ell=0}^{L-1} \mathsf{tr}\left[\mathbf{R}_\ell^{1/2}\left(\mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\mathbf{X} \times \mathbf{R}_{\mathrm{T},\ell}^{1/2}\right) \otimes \mathbf{R}_{\mathrm{R},\ell}\right)^{-1}\mathbf{R}_\ell^{1/2H}\right] \text{ s.t. } \mathsf{tr}\{\mathbf{X}\} = P_t \tag{24}$$

for $\mathbf{R}_\ell^{1/2} = \mathbf{R}_{\mathrm{T},\ell}^{1/2} \otimes \mathbf{R}_{\mathrm{R},\ell+}^{1/2}$, $\ell = 0, 1, \ldots, L-1$. The optimal solution $\mathbf{S}_{\mathrm{opt}}$ of (18) is obtained from the optimal solution $\mathbf{X}_{\mathrm{opt}}$ of (24) as

$$\mathbf{S}_{\mathrm{opt}} = \left[\mathbf{Q}\mathbf{X}_{\mathrm{opt}}^{1/2} \quad \mathbf{0}_{N\times(M_t-\tau_{\mathrm{opt}})}\right] \tag{25}$$

where $\tau_{\mathrm{opt}} = \mathrm{rank}(\mathbf{X}_{\mathrm{opt}})$ and $\mathbf{X}_{\mathrm{opt}}^{1/2} \in \mathbb{C}^{M_t\times\tau_{\mathrm{opt}}}$.

Proof: For notational simplicity, we denote the objective function of (18) by $\mathcal{I}(\mathbf{S})$. By Lemma 1, the following optimization problem gives a lower bound for (18):

$$\min_{\mathbf{S}\in\mathbf{C}^{N\times M_t}} \sum_{\ell=0}^{L-1} \mathsf{tr}\left[\mathbf{R}_\ell^{1/2}\left(\mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\mathbf{S}^H\mathbf{S}\mathbf{R}_{\mathrm{T},\ell}^{1/2}\right) \otimes \mathbf{R}_{\mathrm{R},\ell}\right)^{-1}\mathbf{R}_\ell^{1/2H}\right] \text{ s.t. } \mathsf{tr}\{\mathbf{S}^H\mathbf{S}\} = P_t. \tag{26}$$

Hence

$$\begin{aligned}\min \text{ of Equation (18)} &\ge \text{min of Equation (26)}\\ &= \text{min of Equation (24)},\end{aligned} \tag{27}$$

where the equality in (27) follows from the variable change $0 \preceq \mathbf{X} = \mathbf{S}^H\mathbf{S} \in \mathbf{C}^{M_t\times M_t}$ in (26), which is possible thanks to $N \ge M_t$.

On the other hand, for $\mathbf{S}_{\rm opt}$ defined by (25), it can be seen that

$$\begin{aligned}
&\text{min of Equation (18)} \le \mathcal{I}(\mathbf{S}_{\rm opt}) \\
&= \sum_{\ell=0}^{L-1} \mathsf{tr}\left[\mathbf{R}_\ell^{1/2}\left(\mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\left[\mathbf{R}_{{\rm T},\ell}^{1/2H}\mathbf{X}_{\rm opt}\right.\right.\right. \\
&\times \left.\left.\left.\mathbf{R}_{{\rm T},\ell}^{1/2}\right] \otimes \mathbf{R}_{{\rm R},\ell}\right)^{-1}\mathbf{R}_\ell^{1/2H}\right] \\
&= \text{min of Equation (24)} \qquad (28)
\end{aligned}$$

which together with (27) yield

$$\text{min of Equation (18)} = \text{min of Equation(24)}.$$

Thus, the proof of Theorem 1 is complete. ■

It is interesting to observe from (25) that only $\tau_{\rm opt}$ out of M_t antennas need to send the training pilots for channel estimation. Furthermore, the number $\tau_{\rm opt}$ can be effectively taken as the number of the eigenvalues of $\mathbf{X}_{\rm opt}$ that are large than a threshold.

Using Lemma 2, we now provide a construction of matrix $\mathbf{Q}$ that is composed of M_t orthornormal columns $\mathbf{q}_i,\ i = 1, 2, \ldots, M_t$ satisfying (23), namely

$$\mathbf{q}_i^H \mathbf{F}_\alpha \mathbf{q}_j = 0, \quad \text{for } i, j = 1, 2, \ldots, M_t,\ 1 \le \alpha \le L-1. \qquad (29)$$

Define

$$K = \lceil N/M_t \rceil, \qquad (30)$$

i.e., K is the maximum integer not exceeding N/M_t. Then take

$$\begin{aligned}
\mathbf{q}_1 &= (q_1(0), q_1(2), \ldots, q_1(N-1))^T, \\
|q_1(i)| &= 1/\sqrt{N},\ i = 0, 1, \ldots, N-1, \\
\mathbf{q}_i(k) &= \mathbf{q}_1(k) W_N^{K(i-1)k}, \\
&k = 0, 1, \ldots, N-1; i = 2, \ldots, M_t \qquad (31)
\end{aligned}$$

so that $||\mathbf{q}_i|| = 1$, $i = 1, \ldots, M_t$. According to Lemma 2

$$\begin{aligned}
\mathbf{q}_i^H \mathbf{F}_\alpha \mathbf{q}_j &= \sum_{k=0}^{N-1} |\mathbf{q}_1(k)|^2 W_N^{-[K(i-j)-\alpha]k} \\
&= \frac{1}{N}\sum_{k=0}^{N-1} W_N^{-[K(i-j)-\alpha]k} \\
&= \begin{cases} 1, & \text{for } K(i-j) \equiv \alpha \pmod N, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

For N satisfying (8), $K(i-j) \ne \alpha$ for all $1 \le i, j \le M_t, 1 \le \alpha \le L-1$ and $|K(i-j)| < N$. Therefore

$$K(i-j) \ne \alpha \ (\mod N) \quad \forall\, 1 \le i,\ j \le M_t; \quad \alpha = 1, 2, \ldots, L-1 \qquad (32)$$

and (29) is verified. Moreover, $|K(i-j)| < N$ particularly implies that $K(i-j) \ne 0 \ (\mod N)$. Thus from Lemma 2 one has

$$\mathbf{q}_i^H \mathbf{q}_j = \sum_{k=0}^{N-1} W_N^{K(j-i)k} = 0 \qquad (33)$$

whenever $i \ne j$. That is, the matrix $\mathbf{Q}$ consisting of columns $\mathbf{q}_i,\ i = 1, 2, \ldots, M_t$, satisfies the orthogonality condition:

$$\mathbf{Q}^H\mathbf{Q} = [\mathbf{q}_i^H \mathbf{q}_j]_{i,j=1,2,\ldots,M_t} = \mathbf{I}_{M_t}. \qquad (34)$$

In summary, we have shown that for N satisfying (8), a matrix $\mathbf{Q}$ satisfying the condition of Theorem 1 exists, making the nonconvex optimization problem (18) equivalent to the convex optimization problem (24).

One can recognize that (24) is convex since it is in fact the following SDP

$$\begin{aligned}
&\min_{\substack{0 \preceq \mathbf{X} \in \mathbf{C}^{M_t \times M_t}, \\ \mathbf{Z}_\ell \in \mathbf{C}^{(M_t M_r)\times(M_t M_r)}}} \sum_{\ell=0}^{L-1} \mathsf{tr}(\mathbf{Z}_\ell) \quad \text{s.t.} \quad \mathsf{tr}\{\mathbf{X}\} = P_t, \\
&\begin{bmatrix} \mathbf{Z}_\ell & \mathbf{R}_\ell^{1/2} \\ \mathbf{R}_\ell^{1/2H} & \mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_{{\rm T},\ell}^{1/2H}\mathbf{X}\mathbf{R}_{{\rm T},\ell}^{1/2}\right) \otimes \mathbf{R}_{{\rm R},\ell} \end{bmatrix} \succeq 0, \\
&\ell = 0, \ldots, L-1, \qquad (35)
\end{aligned}$$

where $\mathbf{Z}_\ell$ are slack variables.

It is pointed out that the SDP in (35) is equivalent to the SDP in [20, (32)/(34)] when $\mathbf{R}_{{\rm T},\ell}$ is nonsingular. However, unlike [20], solving the SDP (35) with existing solvers is not computationally feasible because of the large dimensions of the matrix variables $\mathbf{Z}_\ell \in \mathbb{C}^{(M_t M_r)\times(M_r M_r)}$ and $\mathbf{X} \in \mathbb{C}^{M_t \times M_t}$. The next section develops a novel fast computational procedure to find the solution of (24) for the case of "full length" training sequence, i.e., when N satisfies (8).

IV. Fast Convex Programming for Optimizing Full-Length Training Sequences

First, make the following singular value decompositions (SVDs)

$$\mathbf{R}_{{\rm R},\ell} = \mathbf{V}_\ell \boldsymbol{\Upsilon}_\ell \mathbf{V}_\ell^H, \qquad (36)$$

where $\boldsymbol{\Upsilon}_\ell = \mathrm{diag}(\gamma_{\ell 1}, \ldots, \gamma_{\ell M_r}),\ \ell = 0, 1, \ldots, L-1$. It can be easily verified that problem (24) is equivalently simplified to the following semi-definite program:

$$\begin{aligned}
&\min_{0 \preceq \mathbf{X} \in \mathbf{C}^{M_t \times M_t}} \sum_{\ell=0}^{L-1} \mathsf{tr}\left[\tilde{\mathbf{R}}_\ell^{1/2}\left(\mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_{{\rm T},\ell}^{1/2H}\mathbf{X}\mathbf{R}_{{\rm T},\ell}^{1/2}\right)\right.\right. \\
&\left.\left. \otimes\ \boldsymbol{\Upsilon}_\ell\right)^{-1}\tilde{\mathbf{R}}_\ell^{1/2H}\right] \quad \text{s.t.} \quad \mathsf{tr}\{\mathbf{X}\} = P_t, \qquad (37)
\end{aligned}$$

for $\tilde{\mathbf{R}}_\ell = \mathbf{R}_{\mathrm{T},\ell} \otimes \boldsymbol{\Upsilon}_\ell$. We will approximate this semi-definite program by

$$\min_{0 \preceq \mathbf{X} \in \mathbf{C}^{M_t \times M_t}} \sum_{\ell=0}^{L-1} \mathsf{tr}\left[\tilde{\mathbf{R}}_\ell^{1/2}\left(\mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\mathbf{X}(\epsilon)\mathbf{R}_{\mathrm{T},\ell}^{1/2}\right) \otimes \boldsymbol{\Upsilon}_\ell\right)^{-1}\tilde{\mathbf{R}}_\ell^{1/2H}\right] \quad \text{s.t.} \quad \mathsf{tr}\{\mathbf{X}\} = P_t, \tag{38}$$

where $\mathbf{X}(\epsilon) = \mathbf{X} + \epsilon \mathbf{I}_{M_t} \succ 0$ whenever $\mathbf{X} \succeq 0$ and $\epsilon > 0$. It is obvious that the optimal solution of (38) will tend to that of (37). Moreover, the objective function in (38) monotonically increases to that in (37) as $\epsilon \to 0+$.

Now, define

$$f_\ell(\mathbf{X}) = \mathsf{tr}\left[\tilde{\mathbf{R}}_\ell^{1/2}\left(\mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\mathbf{X}(\epsilon)\mathbf{R}_{\mathrm{T},\ell}^{1/2}\right) \otimes \boldsymbol{\Upsilon}_\ell\right)^{-1} \times \tilde{\mathbf{R}}_\ell^{1/2H}\right].$$

Suppose $\mathbf{X}^{(\kappa)}$ is feasible for (38). It follows from Lemma 3 in the Appendix that

$$f_\ell(\mathbf{X}) \le f_\ell^{(\kappa)}(\mathbf{X}), \tag{39}$$

for

$$\begin{aligned} f_\ell^{(\kappa)}(\mathbf{X}) := {} & f_\ell(\mathbf{X}^{(\kappa)}) + \frac{\rho}{M_t}\mathsf{tr}\Bigg\{\Bigg[\mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\Big(\mathbf{R}_{\mathrm{T},\ell}^{1/2H} \\ & \times \mathbf{X}^{(\kappa)}(\epsilon)\mathbf{R}_{\mathrm{T},\ell}^{1/2}\Big) \otimes \boldsymbol{\Upsilon}_\ell\Bigg]^{-1}\tilde{\mathbf{R}}_\ell^{1/2H}\tilde{\mathbf{R}}_\ell^{1/2}\Bigg[\mathbf{I}_{\tau_\ell M_r} \\ & + \frac{\rho}{M_t}\left(\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\mathbf{X}^{(\kappa)}(\epsilon)\mathbf{R}_{\mathrm{T},\ell}^{1/2}\right) \otimes \boldsymbol{\Upsilon}_\ell\Bigg]^{-1}\Bigg(\Big(\mathbf{R}_{\mathrm{T},\ell}^{1/2H}(\mathbf{X}^{(\kappa)}(\epsilon) \\ & \times \mathbf{X}^{-1}(\epsilon)\mathbf{X}^{(\kappa)}(\epsilon) - \mathbf{X}^{(\kappa)}(\epsilon))\mathbf{R}_{\mathrm{T},\ell}^{1/2}\Big) \otimes \boldsymbol{\Upsilon}_\ell\Bigg)\Bigg\}. \end{aligned} \tag{40}$$

By partitioning

$$\begin{aligned} & \left[\mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\mathbf{X}^{(\kappa)}(\epsilon)\mathbf{R}_{\mathrm{T},\ell}^{1/2}\right) \otimes \boldsymbol{\Upsilon}_\ell\right]^{-1}\tilde{\mathbf{R}}_\ell^{1/2H}\tilde{\mathbf{R}}_\ell^{1/2} \\ & \quad \times \left[\mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\mathbf{X}^{(\kappa)}(\epsilon)\mathbf{R}_{\mathrm{T},\ell}^{1/2}\right) \otimes \boldsymbol{\Upsilon}_\ell\right]^{-1} \\ & \quad = [\mathcal{A}_{mn}^{(\ell,\kappa)}]_{m=1,\ldots,\tau_\ell;n=1,\ldots,\tau_\ell}, \quad \mathcal{A}_{mn}^{(\ell,\kappa)} \in \mathbb{C}^{M_r \times M_r} \end{aligned}$$

one has

$$\begin{aligned} & \mathsf{tr}\Bigg\{\left[\mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\mathbf{X}^{(\kappa)}(\epsilon)\mathbf{R}_{\mathrm{T},\ell}^{1/2}\right) \otimes \boldsymbol{\Upsilon}_\ell\right]^{-1}\tilde{\mathbf{R}}_\ell^{1/2H} \\ & \times \tilde{\mathbf{R}}_\ell^{1/2}\left[\mathbf{I}_{\tau_\ell M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\mathbf{X}^{(\kappa)}(\epsilon)\mathbf{R}_{\mathrm{T},\ell}^{1/2}\right) \otimes \boldsymbol{\Upsilon}_\ell\right]^{-1} \\ & \times \left(\left(\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\mathbf{X}^{(\kappa)}(\epsilon)\mathbf{X}^{-1}(\epsilon)\mathbf{X}^{(\kappa)}(\epsilon)\mathbf{R}_{\mathrm{T},\ell}^{1/2}\right) \otimes \boldsymbol{\Upsilon}_\ell\right)\Bigg\} \\ & = \mathsf{tr}\{\mathbf{X}^{(\kappa)}(\epsilon)\mathbf{R}_{\mathrm{T},\ell}^{1/2}A^{(\ell,\kappa)}\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\mathbf{X}^{(\kappa)}(\epsilon)\mathbf{X}^{-1}(\epsilon)\} \end{aligned}$$

for $\mathcal{A}^{(\ell,\kappa)} = [\mathsf{tr}\{A_{mn}^{(\ell,\kappa)}\boldsymbol{\Upsilon}_\ell\}]_{m=1,\ldots,\tau_\ell;n=1,\ldots,\tau_\ell}$.

Consider

$$\min_{0 \prec \mathbf{X} \in \mathbf{C}^{M_t \times M_t}} \sum_{\ell=0}^{L-1} f_\ell^{(\kappa)}(\mathbf{X}) \quad \text{s.t.} \quad \mathsf{tr}\{\mathbf{X}\} = P_t, \tag{41}$$

which is the same as

$$\begin{aligned} \min_{0 \preceq \mathbf{X} \in \mathbf{C}^{M_t \times M_t}} \quad & \mathsf{tr}\{\mathbf{X}^{(\kappa)}(\epsilon)\left(\sum_{\ell=0}^{L-1}\mathbf{R}_{\mathrm{T},\ell}^{1/2}\mathcal{A}^{(\ell,\kappa)}\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\right) \\ & \times \mathbf{X}^{(\kappa)}(\epsilon)\mathbf{X}^{-1}(\epsilon)\} \\ \text{s.t.} \quad & \mathsf{tr}\{\mathbf{X}\} = P_t. \end{aligned} \tag{42}$$

Under the following SVD

$$\mathbf{X}^{(\kappa)}(\epsilon)\left(\sum_{\ell=0}^{L-1}\mathbf{R}_{\mathrm{T},\ell}^{1/2}\mathcal{A}^{(\ell,\kappa)}\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\right)\mathbf{X}^{(\kappa)}(\epsilon) = \mathbf{U}^{(\kappa)}\mathrm{diag}(\alpha_1,\ldots,\alpha_{M_t})(\mathbf{U}^{(\kappa)})^H, \tag{43}$$

the optimal solution $\mathbf{X}^{(\kappa+1)}$ of (41)/(42) is given in a closed form as follows:

$$\mathbf{X}^{(\kappa+1)} = \mathbf{U}^{(\kappa)}\mathrm{diag}\left(\left(\frac{(P_t + k\epsilon)\sqrt{\alpha_1}}{\sum_{i=1}^{M_t}\sqrt{\alpha_i}} - \epsilon\right)^+, \ldots, \left(\frac{(P_t + k\epsilon)\sqrt{\alpha_{M_t}}}{\sum_{i=1}^{M_t}\sqrt{\alpha_i}} - \epsilon\right)^+\right)(\mathbf{U}^{(\kappa)})^H, \tag{44}$$

where k is the rank of

$$\mathbf{X}^{(\kappa)}(\epsilon)\left(\sum_{\ell=0}^{L-1}\mathbf{R}_{\mathrm{T},\ell}^{1/2}\mathcal{A}^{(\ell,\kappa)}\mathbf{R}_{\mathrm{T},\ell}^{1/2H}\right)\mathbf{X}^{(\kappa)}(\epsilon).$$

As $\mathbf{X}^{(\kappa)}$ is also feasible for (41) and $\mathbf{X}^{(\kappa+1)}$ is the optimal solution of (41), it is true that

$$\sum_{\ell=0}^{L-1} f_\ell^{(\kappa)}(\mathbf{X}^{(\kappa+1)}) \le \sum_{\ell=0}^{L-1} f_\ell^{(\kappa)}(\mathbf{X}^{(\kappa)})$$

which together with (39) yield

$$\begin{aligned} \sum_{\ell=0}^{L-1} f_\ell(\mathbf{X}^{(\kappa+1)}) & \le \sum_{\ell=0}^{L-1} f_\ell^{(\kappa)}(\mathbf{X}^{(\kappa+1)}) \\ & \le \sum_{\ell=0}^{L-1} f_\ell^{(\kappa)}(\mathbf{X}^{(\kappa)}) \\ & = \sum_{\ell=0}^{L-1} f_\ell(\mathbf{X}^{(\kappa)}), \end{aligned}$$

i.e., $\mathbf{X}^{(\kappa+1)}$ is a better feasible point than $\mathbf{X}^{(\kappa)}$ for (38) as far as $\mathbf{X}^{(\kappa+1)} \neq \mathbf{X}^{(\kappa)}$. If $\mathbf{X}^{(\kappa+1)} = \mathbf{X}^{(\kappa)}$ then it is obvious that $\mathbf{X}^{(\kappa)}$ is the optimal solution of (41) and thus is the KKT point for (41). It can be easily checked that any KKT point for (41) is also a KKT point for (38).

The above analysis shows that $\{\mathbf{X}^{(\kappa)}\}$ is a sequence of improved solutions of (38), which converges to the KKT point [33]. Since (38) is convex, such KKT point is its optimal solution. Thus we have shown that the sequence $\{\mathbf{X}^{(\kappa)}\}$ converges

Algorithm 1: Fast Convex Programming (FCP) for Optimizing Full-Length Training Sequences.

1: Construct $\mathbf{q}_i$, $i = 1, \ldots, M_t$ by (31) and form $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_{M_t})$;
2: Initialize $\kappa := 0$ and $\mathbf{X}^{(0)}$ feasible to (38).
3: **repeat**
4: Compute $\mathbf{X}^{(\kappa+1)}$ by (43), (44).
5: Set $\kappa := \kappa + 1$.
6: **until** $(\sum_{\ell=0}^{L-1} f_\ell(\mathbf{X}^{(\kappa)}) - \sum_{\ell=0}^{L-1} f_\ell(\mathbf{X}^{(\kappa+1)}))/\sum_{\ell=0}^{L-1} f_\ell(\mathbf{X}^{(\kappa)}) \leq \epsilon_{\text{tol}}$.
7: Accept $\mathbf{X} = \mathbf{X}^{(\kappa)}$ and $\mathbf{S} = \mathbf{Q}\mathbf{X}^{1/2H}$ and calculate the performance metric by $\sum_{\ell=0}^{L-1} f_\ell(\mathbf{X})$ in (24).

to the optimal solution of the convex optimization problem (38). The pseudo-code for this iterative process is provided by Algorithm 1. The following Proposition summarizes the main result.

Proposition 1: Under any computational tolerance $\epsilon_{\text{tol}} > 0$, Algorithm 1 will terminate after finitely many iterations, yielding the global optimal solution of the large-scale optimization problem (38).

Although the global convergence of Algorithm 1 is guaranteed for any initial point $\mathbf{X}^{(0)}$ that is feasible to (24), it is also important to initialize Algorithm 1 from a good point to further reduce the computation effort. To this end, we now develop a suboptimal strategy for solution of (24), which is faster than Algorithm 1 but performs optimally only at the following special case of the transmit correlation matrix[1]:

$$\mathbf{R}_{\text{T},\ell} = \mathbf{U}_0 \mathbf{\Lambda}_\ell \mathbf{U}_0^H, \ \ell = 0, 1, \ldots, L-1, \tag{45}$$

where $\mathbf{\Lambda}_\ell = \mathsf{diag}[\lambda_{\ell i}]_{i=1,\ldots,M_t}$.

Under SVD (36), instead of (38), we consider another equivalent semi-definite formulation

$$\min_{0 \preceq \mathbf{X} \in \mathbf{C}^{M_t \times M_t}} \sum_{\ell=0}^{L-1} \mathsf{tr}\left[\tilde{\mathbf{R}}_{\ell+}^{1/2}\left(\mathbf{I}_{M_t M_r} + \frac{\rho}{M_t}\left(\mathbf{R}_{\text{T},\ell+}^{1/2}\mathbf{X}\mathbf{R}_{\text{T},\ell+}^{1/2}\right) \otimes \mathbf{\Upsilon}_\ell\right)^{-1}\tilde{\mathbf{R}}_{\ell+}^{1/2}\right] \quad \text{s.t.} \quad \mathsf{tr}\{\mathbf{X}\} = P_t, \tag{46}$$

with $\tilde{\mathbf{R}}_{\ell+}^{1/2} = \mathbf{R}_{\text{T},\ell+}^{1/2} \otimes \mathbf{\Upsilon}_\ell^{1/2}$, which can be written as

$$\min_{0 \preceq \tilde{\mathbf{X}} \in \mathbf{C}^{M_t \times M_t}} \sum_{\ell=0}^{L-1} (\mathbf{I}_{M_t} \otimes \mathbf{\Lambda}_\ell^{1/2})\left(\mathbf{I}_{M_t M_r} + \frac{\rho}{M_t}(\mathbf{\Lambda}_\ell^{1/2}\tilde{\mathbf{X}}\mathbf{\Lambda}_\ell^{1/2}) \otimes \mathbf{\Upsilon}_\ell\right)^{-1}(\mathbf{I}_{M_t} \otimes \mathbf{\Lambda}_\ell^{1/2}) \ \text{ s.t. } \ \mathsf{tr}\{\tilde{\mathbf{X}}\} = P_t, \tag{47}$$

because $\tilde{\mathbf{X}}_{\text{opt}} = \mathbf{U}_0^H \mathbf{X}_{\text{opt}} \mathbf{U}_0$ for the optimal solutions $\tilde{\mathbf{X}}_{\text{opt}}$ and $\mathbf{X}_{\text{opt}}$ of (47) and (46).

With the optimal solution $\tilde{\mathbf{X}}_{\text{opt}}$ of (47), it is true that

$$\mathbf{X}^{(0)} = \mathbf{U}_0 \tilde{\mathbf{X}}_{\text{opt}} \mathbf{U}_0^H \tag{48}$$

is feasible to (38) and can be used as an initial point for the implementation of Algorithm 1.

It follows from Lemma 1 that for any positive definite matrix $\tilde{\mathbf{X}}$, one has

$$\begin{aligned}
&\mathsf{tr}\left[\left(\mathbf{I}_{M_t} \otimes \mathbf{\Lambda}_\ell^{1/2}\right)\left(\mathbf{I}_{M_t M_r} + \frac{\rho}{M_t}\left(\mathbf{\Lambda}_\ell^{1/2}\tilde{\mathbf{X}}\mathbf{\Lambda}_\ell^{1/2}\right) \otimes \mathbf{\Upsilon}_\ell\right)^{-1} \times \left(\mathbf{I}_{M_t} \otimes \mathbf{\Lambda}_\ell^{1/2}\right)\right] \\
&\geq \mathsf{tr}\left[\left(\mathbf{I}_{M_t} \otimes \mathbf{\Lambda}_\ell^{1/2}\right)(\mathbf{I}_{M_t M_r} + \frac{\rho}{M_t}\mathsf{diag}[\tilde{\mathbf{X}}(i,i)\mathbf{\Lambda}_\ell(i,i)]_{i=1,2,\ldots,M_t} \otimes \mathbf{\Upsilon}_\ell)^{-1} \times \left(\mathbf{I}_{M_t} \otimes \mathbf{\Lambda}_\ell^{1/2}\right)\right].
\end{aligned}$$

Thus the optimal solution of (47) with only one constraint $\mathsf{tr}\{\tilde{\mathbf{X}}\} = P_t$ must be in the diagonal form:

$$\tilde{\mathbf{X}} = \mathsf{diag}(y_1, y_2, \ldots, y_{M_t}).$$

This means that the matrix optimization problem (47) in the variable $\tilde{\mathbf{X}}$ is equivalent to the following vector optimization problem in $\mathbf{y} = (y_1, y_2, \ldots, y_{M_t})^T \in R^{M_t}$:

$$\min_{y_i \geq 0, i=1,2,\ldots,M_t} \sum_{i=1}^{M_t} f_i(y_i) \quad \text{s.t.} \quad \sum_{i=1}^{M_t} y_i = P_t, \tag{49}$$

where

$$f_i(y_i) = \sum_{\ell=0}^{L-1}\sum_{j=1}^{M_r} \lambda_{\ell i}\left(1 + \frac{\rho}{M_t}\lambda_{\ell i}\gamma_{\ell j} y_i\right)^{-1}. \tag{50}$$

Although being convex, (49) does not admit a closed-form optimal solution. Applying existing convex solvers for (49) is still time-consuming for large M_t. In the remainder of this section we will develop a path-following procedure of extremely low computational complexity for the optimal solution of (49).

Using the inequality

$$(1+x)^{-1} \leq (1+\bar{x})^{-1} + (1+\bar{x})^{-2}\bar{x}^2(1/x - 1/\bar{x}) \quad \forall\, x > 0, \bar{x} > 0$$

one has

$$\begin{aligned}
f_i(y_i) &\leq f_i^{(\kappa)}(y_i) \\
&:= f_i\left(y_i^{(\kappa)}\right) + \alpha_i^{(\kappa)}\left(1/y_i - 1/y_i^{(\kappa)}\right) \quad \forall\, y_i > 0, y_i^{(\kappa)} > 0
\end{aligned}$$

and $f_i\left(y_i^{(\kappa)}\right) = f_i^{(\kappa)}\left(y_i^{(\kappa)}\right)$ for

$$\alpha_i^{(\kappa)} := \sum_{\ell=0}^{L}\sum_{j=1}^{M_r} \lambda_{\ell i}\left(\frac{(\rho/M_t)\lambda_{\ell i}\gamma_{\ell j} y_i^{(\kappa)}}{(1 + (\rho/M_t)\lambda_{\ell i}\gamma_{\ell j} y_i^{(\kappa)})}\right)^2 \times \frac{1}{(\rho/M_t)\lambda_{\ell i}\gamma_{\ell j}} \tag{51}$$

[1]It is pointed out that such a special case has been addressed in [17] for extreme SNR regimes of small-scale MIMO systems only.

Algorithm 2: Faster Convex Programming (Faster CP) for Suboptimal Full-Length Training Sequences.

1: Initialize $\kappa := 0$ and $y_i^{(\kappa)} = P_t/M_t, i = 1, \ldots, M_t$.
2: **repeat**
3: Compute $\alpha_i^{(\kappa)}, i = 1, \ldots, M_t$ by (51) and then $y_i^{(\kappa+1)}, i = 1, \ldots, M_t$ by (53). Set $y^{(\kappa+1)} = (y_1^{(\kappa+1)}, \ldots, y_{M_t}^{(\kappa+1)})^T$.
4: Set $\kappa := \kappa + 1$.
5: **until** $(\sum_{i=1}^{M_t} f_i(y_i^{(\kappa+1)}) - \sum_{i=1}^{M_t} f_i(y_i^{(\kappa)}))/\sum_{i=1}^{M_t} f_i(y_i^{(\kappa)}) \leq \epsilon_{\rm tol}$.
6: Accept $\tilde{\mathbf{X}}_{\rm opt} = {\rm diag}\{y_1^{(\kappa)}, \ldots, y_{M_t}^{(\kappa)}\}$ and then $\mathbf{X}^{(0)}$ by (48).

Therefore, at the κ-th iteration we solve the following majorant minimization of (49)

$$\min_{y_i \geq 0, i=1,2,\ldots,M_t} \sum_{i=1}^{M_t} f_i^{(\kappa)}(y_i) \quad \text{s.t.} \sum_{i=1}^{M_t} y_i = P_t, \tag{52}$$

which admits a closed-form solution:

$$y_i^{(\kappa+1)} = \frac{P_t\sqrt{\alpha_i^{(\kappa)}}}{\sum_{j=1}^{M_t}\sqrt{\alpha_j^{(\kappa)}}}. \tag{53}$$

The following proposition summarizes the above result.

Proposition 2: Under any computational tolerance $\epsilon_{\rm tol} > 0$, Algorithm 2 will terminate after finitely many iterations, yielding the global optimal solution of the optimization problem (47), which results in a feasible solution (48) of the large-scale optimization problem (24).

Remark: It can be seen from (50) that the rank of the optimal solution $\tilde{\mathbf{X}}_{\rm opt}$ of (47) is no more than $\tau_{\max}$, which is the maximum rank of $\mathbf{R}_{{\rm T},\ell}$, $\ell = 0, \ldots, L-1$. Thus under the special structure (45), not more than $\tau_{\max}$ out of M_t antennas are needed to send the optimal training sequences.

V. Successive Quadratic Programming for Nonconvex Optimization of Reduced-Length Training Sequences

For the case that the length of the training sequence is short such that $N < LM_t$, the convex optimization problem (24) is only a lower bound minimization of the nonconvex optimization problem (18). Consequently, FCP developed in the previous section does not provide an optimal solution. In this section we develop a successive quadratic programming (SQP) for solving (18) for this case, which, by (13), is equivalent to the following optimization problem:

$$\max_{\mathbf{S}\in\mathbf{C}^{N\times M_t}} f(\mathbf{S}) \quad \text{s.t.} \quad (14), \tag{54}$$

where

$$f(\mathbf{S}) = \mathsf{tr}\left[\mathbf{R}_{\mathbf{h}}\mathbf{M}^H(\mathbf{S})\left(\frac{\rho}{M_t}\mathbf{M}(\mathbf{S})\mathbf{R}_{\mathbf{h}}\mathbf{M}^H(\mathbf{S}) + \mathbf{I}_{NM_r}\right)^{-1} \times \mathbf{M}(\mathbf{S})\mathbf{R}_{\mathbf{h}}\right].$$

At any $\mathbf{S}^{(\kappa)}$ satisfying the power constraint (14), by using Lemma 4, one has

$$\begin{aligned} f(\mathbf{S}) \geq\ & f^{(\kappa)}(\mathbf{S}) \\ :=\ & 2\Re\{\mathsf{tr}(\mathcal{L}^H(\mathbf{S}^{(\kappa)})\mathbf{M}(\mathbf{S}))\} - \frac{\rho}{M_t}\mathsf{tr}(\Phi(\mathbf{S}^{(\kappa)}) \\ & \times \mathbf{M}(\mathbf{S})\mathbf{R}_{\mathbf{h}}\mathbf{M}^H(\mathbf{S})) - \mathsf{tr}(\Phi(\mathbf{S}^{(\kappa)})), \end{aligned}$$

where

$$\begin{aligned} \mathcal{L}(\mathbf{S}^{(\kappa)}) =\ & \left(\frac{\rho}{M_t}\mathbf{M}(\mathbf{S}^{(\kappa)})\mathbf{R}_{\mathbf{h}}\mathbf{M}^H(\mathbf{S}^{(\kappa)}) + \mathbf{I}_{NM_r}\right)^{-1} \\ & \times \mathbf{M}(\mathbf{S}^{(\kappa)})\mathbf{R}_{\mathbf{h}}^2, \\ \Phi(\mathbf{S}^{(\kappa)}) =\ & \left(\frac{\rho}{M_t}\mathbf{M}(\mathbf{S}^{(\kappa)})\mathbf{R}_{\mathbf{h}}\mathbf{M}^H(\mathbf{S}^{(\kappa)}) + \mathbf{I}_{NM_r}\right)^{-1} \\ & \times \mathbf{M}(\mathbf{S}^{(\kappa)})\mathbf{R}_{\mathbf{h}}^2\mathbf{M}^H(\mathbf{S}^{(\kappa)}) \\ & \times \left(\frac{\rho}{M_t}\mathbf{M}(\mathbf{S}^{(\kappa)})\mathbf{R}_{\mathbf{h}}\mathbf{M}^H(\mathbf{S}^{(\kappa)}) + \mathbf{I}_{NM_r}\right)^{-1} \end{aligned}$$

At the κ-th iteration, the following convex quadratic optimization is solved to generate $\mathbf{S}^{(\kappa+1)}$

$$\begin{aligned} \max_{\mathbf{S}\in\mathbf{C}^{N\times M_t}} \quad & \Big[2\Re\{\mathsf{tr}(\mathcal{L}^H(\mathbf{S}^{(\kappa)})\mathbf{M}(\mathbf{S}))\} - \frac{\rho}{M_t}\mathsf{tr}(\Phi(\mathbf{S}^{(\kappa)}) \\ & \times\ \mathbf{M}(\mathbf{S})\mathbf{R}_{\mathbf{h}}\mathbf{M}^H(\mathbf{S}))\Big] \\ \text{s.t.} \quad & (14). \end{aligned} \tag{55}$$

where, by (11),

$$\mathbf{M}(\mathbf{S})\mathbf{R}_{\mathbf{h}}\mathbf{M}^H(\mathbf{S})) = \sum_{\ell=0}^{L-1}(\mathbf{F}_\ell\mathbf{S}\mathbf{R}_{{\rm T},\ell}\mathbf{S}^H\mathbf{F}_\ell^H) \otimes \mathbf{R}_{{\rm R},\ell}.$$

Now, partition $\Phi(\mathbf{S}^{(\kappa)})$ into N^2 blocks of size $M_r \times M_r$,

$$\begin{aligned} \Phi(\mathbf{S}^{(\kappa)}) &= [\Phi_{i,j}(\mathbf{S}^{(\kappa)})]_{i=1,\ldots,N;j=1,\ldots,N}, \\ \Phi_{i,j}(\mathbf{S}^{(\kappa)}) &\in \mathbb{C}^{M_r\times M_r}, \end{aligned}$$

and then represent

$$\begin{aligned} &\mathbf{x} = {\rm vec}(\mathbf{S}), \quad x_{n+(t-1)N} = \mathbf{S}(n,t), \\ &1 \leq n \leq N, 1 \leq t \leq M_t. \end{aligned}$$

Furthermore, let $\mathbf{I}_{n,t} \in R^{N\times M_t}$ be a matrix with all zero entries except $\mathbf{I}_{n,t}(n,t) = 1$ and define $\mathbf{I}_{n+(t-1)N} := \mathbf{I}_{n,t}$. Then define $\mathbf{b} \in \mathbf{C}^{NM_t}$ such that

$$\mathbf{b}^*_{n+(t-1)N} = \mathsf{tr}(\mathcal{L}^H(\mathbf{S}^{(\kappa)})\mathbf{M}(\mathbf{I}_{t,n})), 1 \leq n \leq N, 1 \leq t \leq M_t,$$

Algorithm 3: Successive Quadratic Programming (SQP) for Nonconvex Optimization of Reduced-Length Training sequences.

1: Initialize $\kappa := 0$ and $X^{(0)}$ as a feasible point for (18)
2: **repeat**
3: Compute $\mathbf{x}_{\rm opt}$ by (57) and $\mathbf{X}^{(\kappa+1)}$ by matricization of $\mathbf{x}_{\rm opt}$.
4: Set $\kappa := \kappa + 1$.
5: **until** $|f(\mathbf{S}^{(\kappa+1)}) - f(\mathbf{S}^{(\kappa)}))/f(\mathbf{S}^{(\kappa)})| \leq \epsilon_{\rm tol}$ for some tolerance $\epsilon_{\rm tol}$.
6: Accept $\mathbf{S} = \mathbf{S}^{(\kappa)}$ and $f(\mathbf{S}^{(\kappa)})$ as the performance.

and $0 \preceq \mathbf{A} \in \mathbb{C}^{NM_t \times NM_t}$ such that, for $1 \leq n' \leq N, 1 \leq n \leq N, 1 \leq t \leq M_t, 1 \leq t' \leq M_t$,

$$\begin{aligned} &\mathbf{A}(n' + (t'-1)N, n + (t-1)N) \\ = &\ \mathsf{tr}\left(\Phi(\mathbf{S}^{(\kappa)}) \left(\sum_{\ell=0}^{L-1} (\mathbf{F}_\ell \mathbf{I}_{n+(t-1)N} \mathbf{R}_{{\rm T},\ell} \mathbf{I}^T_{n'+(t'-1)N} \mathbf{F}^H_\ell) \right.\right. \\ &\left.\left. \otimes\ \mathbf{R}_{{\rm R},\ell}\right)\right) \\ = &\ \sum_{\ell=0}^{L-1} W_N^{(n-n')\ell} \mathbf{R}_{{\rm T},\ell}(t,t') \mathsf{tr}(\Phi_{n',n}(\mathbf{S}^{(\kappa)}) \mathbf{R}_{{\rm R},\ell}). \end{aligned}$$

It follows that

$$\begin{aligned} &\mathsf{tr}(\mathcal{L}^H(\mathbf{S}^{(\kappa)}) \mathbf{M}(\mathbf{S})) = \mathbf{b}^H \mathbf{x}, \\ &\mathsf{tr}(\Phi(\mathbf{S}^{(\kappa)}) \mathbf{M}(\mathbf{S}) \mathbf{R_h} \mathbf{M}^H(\mathbf{S})) = \mathbf{x}^H \mathbf{A} \mathbf{x}. \end{aligned}$$

Thus, problem (55) is in the form

$$\max_{\mathbf{x} \in \mathbf{C}^{NM_t}} \left[2\Re\{\mathbf{b}^H \mathbf{x}\} - \frac{\rho}{M_t} \mathbf{x}^H \mathbf{A} \mathbf{x}\right] \quad \text{s.t.} \quad ||\mathbf{x}||^2 = P_t, \tag{56}$$

which admits the optimal solution in a closed form:

$$\mathbf{x}_{\rm opt} = (\mathbf{A} + \mu \mathbf{I}_{NM_t})^{-1} \mathbf{b}, \tag{57}$$

where $\mu = 0$ if $||\mathbf{A}^{-1}\mathbf{b}||^2 \leq P_t$, otherwise $\mu > 0$ such that

$$||(\mathbf{A} + \mu \mathbf{I}_{NM_t})^{-1} \mathbf{b}||^2 = P_t. \tag{58}$$

The value of μ can be easily found by the following golden search.

Golden search: Start from $\mu_{\min} = 0$ and $\mu_{\max} = ||\mathbf{b}||/\sqrt{P_t}$ and $\mu = (\mu_{\min} + \mu_{\max})/2$. If $||(\mathbf{A} + \mu \mathbf{I}_{NM_t})^{-1}\mathbf{b}||^2 < P_t$ reset $\mu_{\max} = \mu$, otherwise reset $\mu_{\min} = \mu$. Continue search till $||(\mathbf{A} + \mu \mathbf{I}_{NM_t})^{-1}\mathbf{b}||^2 = P_t$.

Finally, $\mathbf{S}^{(\kappa+1)}$ can be recovered from $\mathbf{x}_{\rm opt}$, for which

$$f(\mathbf{S}^{(\kappa+1)}) \leq f(\mathbf{S}^{(\kappa)}), \tag{59}$$

i.e., the objective in (18) is decreasing.

Proposition 3: Algorithm 3 converges to a local solution of the nonconvex optimization problem (18), which satisfies the KKT condition.

Proof: It follows from (59) that the sequence $\{\mathbf{S}^{(\kappa)}\}$ is a sequence of improved solutions of (18), which either terminates or converges to point $\bar{\mathbf{S}}$ to satisfy the KKT condition for the convex optimization problem (55) (with setting $\mathbf{S}^{(\kappa)} = \bar{\mathbf{S}}$), which is also the KKT condition for the nonconvex optimization problem (18) [33]. ■

It is useful to initialize Algorithm 3 from a good feasible point $\mathbf{X}^{(0)}$, which can be found through the following convex relaxation for (18). For $N = \lceil L/2 \rceil M_t < LM_t$, the convex optimization problem (24) is only an upper bound of the nonconvex optimization problem (18). The value of K defined by (30) is $K = \lceil L/2 \rceil$, which means $L - 2K = 0$ if L is even and $L - 2K = 1$ if L is odd. Therefore, $K \leq L - 1$ and $2K > L - 1$ for L even but but $K \leq L - 1$ and $2K = L - 1$ for L odd. Then

$$\begin{aligned} &\mathbf{q}_i^H \mathbf{F}_\alpha \mathbf{q}_j = 0, \quad \text{for } i, j = 1, 2, \ldots, M_t, \\ &\alpha \in \{1, 2, \ldots, L-1\} \backslash \{K, (L-2K)2K\} \end{aligned} \tag{60}$$

and

$$\mathbf{q}_i^H \mathbf{F}_K \mathbf{q}_j = \begin{cases} 1, & \text{for} \quad |i-j| = 1 \quad \text{or} \\ & |i-j| = 2(L-2K) \\ 0, & \text{otherwise} \end{cases} \tag{61}$$

$$\mathbf{q}_i^H \mathbf{F}_{(L-2K)2K} \mathbf{q}_j = \begin{cases} 1, & \text{for} \quad |i-j| = L-2K \\ 0, & \text{otherwise.} \end{cases} \tag{62}$$

Moreover, $|K(i-j)| < N$ particularly implies $K(i-j) \neq 0 \pmod N$. Thus (33) is verified by Lemma 2, which means $\{\mathbf{q}_i\}_{i=1}^{M_t}$ are orthogonal and $\mathbf{Q}$ is full-rank. This means that among $L-1$ off-diagonal blocks $(\mathbf{S}^H \mathbf{F}_\alpha \mathbf{S}) \otimes \mathbf{I}_{M_r}$, $0 \leq \alpha \leq L-1$ of $\mathbf{M}^H(\mathbf{S})\mathbf{M}(\mathbf{S})$ in (11) there is only one nonzero block, namely $(\mathbf{S}^H \mathbf{F}_K \mathbf{S}) \otimes \mathbf{I}_{M_r}$ (when L is even), or two nonzero blocks, namely $(\mathbf{S}^H \mathbf{F}_K \mathbf{S}) \otimes \mathbf{I}_{M_r}$ and $(\mathbf{S}^H \mathbf{F}_{2K} \mathbf{S}) \otimes \mathbf{I}_{M_r}$ (when L is odd). These nonzero blocks are off-diagonal sparse according to (61) and (62) since M_t is large. Therefore, the convex optimization problem (24) is a tight approximation of the nonconvex optimization problem (18).

VI. Simulation Results

This section presents simulation results to illustrate the performance of our proposed methods. Considered in the simulation are large-scale MIMO-OFDM systems employing $M = 1024$ sub-carriers. Due to the BS form factor limitation in practice[1], [2], the BS is equipped with $M_h \times M_v$ uniform planar arrays (UPAs) on a half-wavelength lattice, where there are M_h rows in the horizontal dimension and M_v columns in the vertical dimension. For example, 4×4, 8×4 and 8×8 corresponding to $M_t = 16$, $M_t = 32$ and $M_t = 64$, respectively. There are $M_r = 2$ antennas at each user terminal that are separated with a half wavelength. The wideband frequency-selective channel is simulated with $L = 4$ taps, where the power gains are set as $(\sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2) = (0.3, 0.3, 0.2, 0.2)$. Here, the spatial correlation matrix at the BS is written as $\mathbf{R}_{{\rm T},\ell} = \mathbf{R}_{h,\ell} \otimes \mathbf{R}_{v,\ell}$, where $\mathbf{R}_{h,\ell}$ and $\mathbf{R}_{v,\ell}$ are the horizontal and vertical channel

covariance matrices, respectively. $\mathbf{R}_{h,\ell}$ and $\mathbf{R}_{v,\ell}$ can be represented by one-ring model [34] that is extended to two-dimensional or planar array [25], and is given as

$$[\mathbf{R}_{s,\ell}]_{p,q} = \frac{1}{2\triangle_{s,\ell}} \int_{\bar{\theta}_{s,\ell}-\triangle_{s,\ell}}^{\bar{\theta}_{s,\ell}+\triangle_{s,\ell}} \mathrm{e}^{-j\pi(p-q)\sin(\alpha)} d\alpha,$$

where $s \in \{h, v\}$, $\triangle_{h,\ell}$ and $\triangle_{v,\ell}$ are the angle spreads in the horizontal and vertical dimensions, respectively, $\bar{\theta}_{h,\ell}$ and $\bar{\theta}_{v,\ell}$ are the angles of departure from the transmit array. Likewise, the spatial correlation matrix at the receiver is expressed as

$$[\mathbf{R}_{\mathrm{R},\ell}]_{p,q} = \frac{1}{2\triangle_{r,\ell}} \int_{\bar{\theta}_{r,\ell}-\triangle_{r,\ell}}^{\bar{\theta}_{r,\ell}+\triangle_{r,\ell}} \mathrm{e}^{-j\pi(p-q)\sin(\alpha)} d\alpha,$$

where $\triangle_{r,\ell}$ and $\bar{\theta}_{r,\ell}$ are the angle spread and angle of arrival at the receive array. In the simulations, the angle spreads are the same and equal to $\sigma_\theta = 12°$. The cyclic prefix is chosen to be $L-1$ which is long enough to cancel inter-block interference. The power budget is[2]

$$P_{\mathrm{tot}} = MM_t. \tag{63}$$

Thus, the total power for data symbols is

$$P_{\mathrm{sym}} = MM_t - P_t, \tag{64}$$

which is equally located for all data symbols. Specifically, the received signal over the kth data sub-carrier is given as

$$\mathbf{r}(k) = \sqrt{\frac{\rho_s}{M_t}}\mathbf{H}_f(k)\mathbf{s}(k) + \mathbf{n}(k),\ k \notin \{t_0, \ldots, t_{N-1}\}, \tag{65}$$

where $\rho_s = P_{\mathrm{sym}}/(M-N)M_t$ and

$$\mathbf{H}_f(k) := \mathbf{H}(W_M^k) = \sum_{\ell=0}^{L-1} \mathbf{H}_\ell W_M^{\ell k}.$$

Example 1: This simulation example numerically evaluates the performance of optimal training sequences designed in Sections III for downlink channel state estimation of large-scale MIMO-OFDM systems. The angles of departure are set to be

$$\begin{aligned}(\bar{\theta}_{h,0}, \bar{\theta}_{h,1}, \bar{\theta}_{h,2}, \bar{\theta}_{h,3}) &= (13°, 16°, 20°, 24°) \quad \text{and} \\ (\bar{\theta}_{v,0}, \bar{\theta}_{v,1}, \bar{\theta}_{v,2}, \bar{\theta}_{v,3}) &= (20°, 25°, 28°, 33°),\end{aligned} \tag{66}$$

and the angles of arrival are

$$(\bar{\theta}_{r,0}, \bar{\theta}_{r,1}, \bar{\theta}_{r,2}, \bar{\theta}_{r,3}) = (290°, 300°, 315°, 320°). \tag{67}$$

We first focus on the case of using the optimum training length, namely $N = LM_t$. In our discussion, the term "effective rank" is defined as the number of eigenvalues that are larger than the threshold 10^{-6}. The maximum effective ranks $\tau_{\max}$ corresponding to $M_t = 16$, $M_t = 32$ and $M_t = 64$ are found to be 13, 20 and 30, respectively. It is the small angle spread ($\sigma_\theta = 12°$) that results into the singularity of $\mathbf{R}_{\mathrm{T},\ell}$. The MSE with optimal training sequences $\mathbf{X}_{\mathrm{opt}}$ found by the proposed FCP in (41) with $\epsilon_{\mathrm{tol}} = 10^{-5}$ and $\epsilon = 10^{-6}$ is plotted in Fig. 2, which is also compared with the equi-power solution ($\mathbf{S} = \sqrt{N}\mathbf{Q}$). The optimal training sequences $\mathbf{X}_{\mathrm{opt}}$ clearly and significantly outperform the equi-power solution. Furthermore, the effective rank of $\mathbf{X}_{\mathrm{opt}}$ is provided in Table I, which can be seen to be smaller than the effective rank $\tau_{\max}$. According to (25), no more than $\tau_{\mathrm{opt}} \leq \tau_{\max}$ out of M_t antennas need to send the training sequence.

Next, we compare the performance of training sequences obtained with FCP and Faster CP. According to the definition of $\mathbf{R}_{\mathrm{T},\ell}$, the matrix $\mathbf{U}_\ell$ depends on the angles of departure and the angle spreads. Here, we consider two cases of the angles of departure. When the angles of departure are set as (66), Fig. 3 plots the MSE for training sequences obtained with FCP and Faster

[2]Recall that the noise component in (5) is modeled to have unit variance.

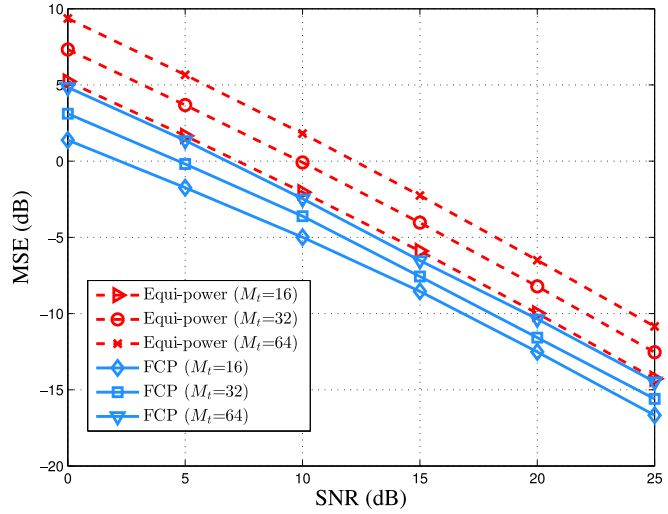

Fig. 2. MSE of FCP and equi-power solutions with $N = LM_t$ in the downlink transmission: (66) and (67).

TABLE I
THE EFFECTIVE RANK OF $\mathbf{X}_{\mathrm{opt}}/\tilde{\mathbf{X}}_{\mathrm{opt}}$ IN (41)/(47)

SNR (dB)	0	5	10	15	20	25
$M_t = 16$	4/5	6/6	6/6	8/8	8/8	8/9
$M_t = 32$	7/8	8/9	9/10	11/11	11/12	12/15
$M_t = 64$	11/13	13/13	14/15	15/19	16/19	18/20

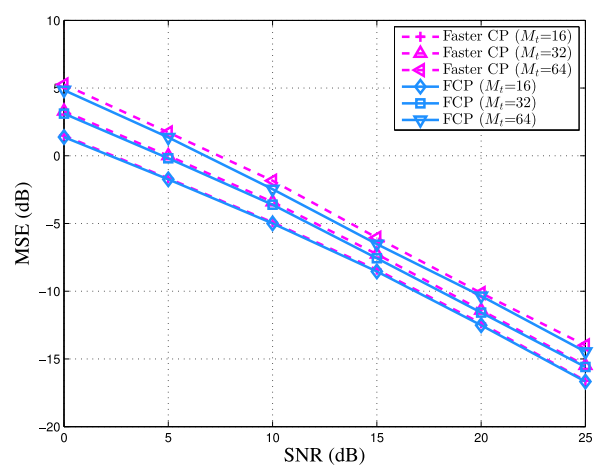

Fig. 3. MSE for training sequences obtained with FCP and Faster CP in the downlink transmission: $N = LM_t$, (66) and (67).

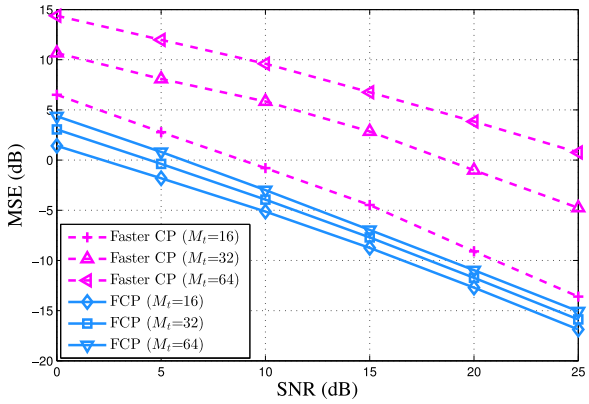

Fig. 4. MSE for training sequences obtained with FCP and Faster CP in the downlink transmission: $N = LM_t$, (67) and (68).

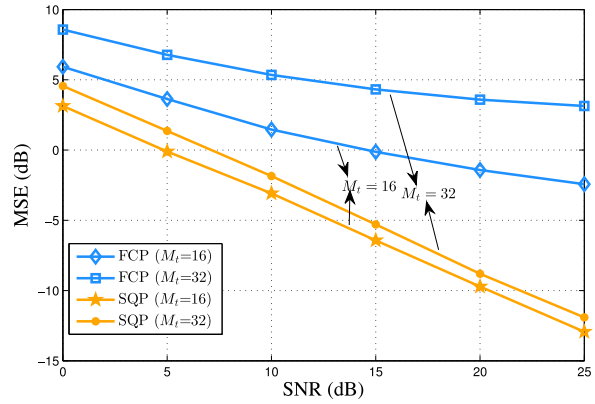

Fig. 5. MSE for training sequences obtained with FCP and SQP in the downlink transmission: $N = LM_t/2$, (66) and (67).

CP. There are only small improvements in using FCP instead of Faster CP. This is due to the fact that such small angles of departure $(\bar{\theta}_{h,0}, \bar{\theta}_{h,1}, \bar{\theta}_{h,2}, \bar{\theta}_{h,3})$ and $(\bar{\theta}_{v,0}, \bar{\theta}_{v,1}, \bar{\theta}_{v,2}, \bar{\theta}_{v,3})$ result in similar $\mathbf{U}_\ell$ $(\ell = 0, 1, \ldots, L-1)$. We also provide the effective rank of $\tilde{\mathbf{X}}_{\text{opt}}$ in Table I.

For larger angles of departure, such as

$$\begin{aligned}(\bar{\theta}_{h,0}, \bar{\theta}_{h,1}, \bar{\theta}_{h,2}, \bar{\theta}_{h,3}) &= (13^\circ, 20^\circ, 36^\circ, 50^\circ) \quad \text{and} \\ (\bar{\theta}_{v,0}, \bar{\theta}_{v,1}, \bar{\theta}_{v,2}, \bar{\theta}_{v,3}) &= (20^\circ, 35^\circ, 49^\circ, 65^\circ), \end{aligned} \tag{68}$$

the unitary matrices $\mathbf{U}_\ell$ $(\ell = 0, 1, \ldots, L-1)$ become different. With these particular settings of angles, the maximum effective ranks $\tau_{\max}$ corresponding to $M_t = 16$, $M_t = 32$ and $M_t = 64$ are still 13, 20 and 30, respectively. Fig. 4 shows the MSE for training sequences obtained with FCP and Faster CP. In this case, FCP clearly outperforms Faster CP.

Example 2: As mentioned earlier, in the case of $N = LM_t/2$, the convex optimization problem (24) becomes a lower bound minimization of the nonconvex optimization problem (18). Hence, training sequences found by FCP in Section III are not optimized. However, the SQP algorithm proposed in Section IV is able to find the optimal training sequences. First, for the angles of departure and arrival set as (66) and (67) and $N = LM_t/2$, the values of MSE achieved by optimal training sequences found by SQP and FCP are compared in Fig. 5. As expected, the optimal training sequences found by SQP significantly outperform training sequences found by FCP with $M_t \in \{16, 32\}$. Fig. 6 plots the MSEs of SQP in (54) with $M_t = 16$ and $N = LM_t/2$ versus iteration numbers for different SNR values and when $\epsilon_{\text{tol}} = 10^{-5}$ is set for the stop criterion. It is observed that the MSE is monotonically decreasing with increasing iteration number. Fig. 7 plots the MSE for optimal training sequences obtained by FCP with $N = LM_t$ and SQP with $N = LM_t/2$. It can be seen that the optimal training sequences by FCP with full length $N = LM_t$ achieves better MSE than that achieved by the training sequences with the reduced, half-length $N = LM_t/2$ found by SQP. This is expected since when longer training sequences are used for channel estimation, the channel estimation becomes more accurate.

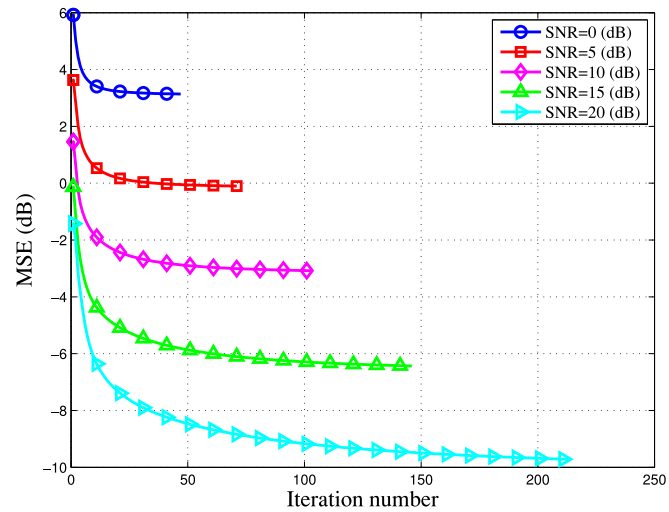

Fig. 6. The MSE of SQP in (54) with $M_t = 16$ and $N = LM_t/2$ versus the number of iterations in the downlink transmission with (66) and (67).

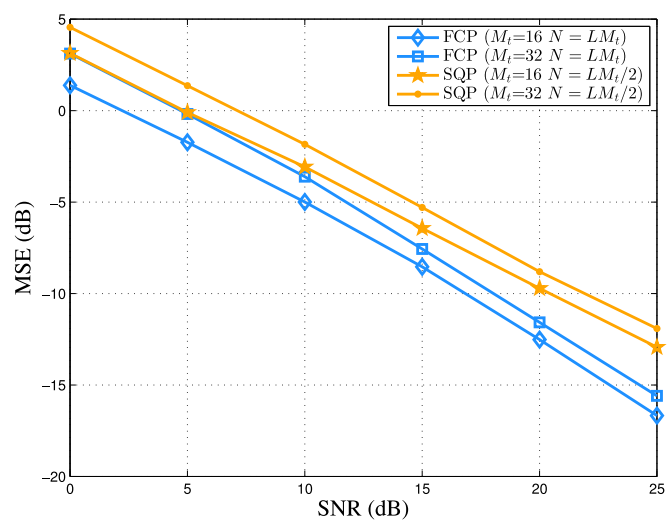

Fig. 7. MSE achieved by training sequences obtained with FCP when $N = LM_t$ and with SQP when $N = LM_t/2$: Downlink transmission with (66) and (67).

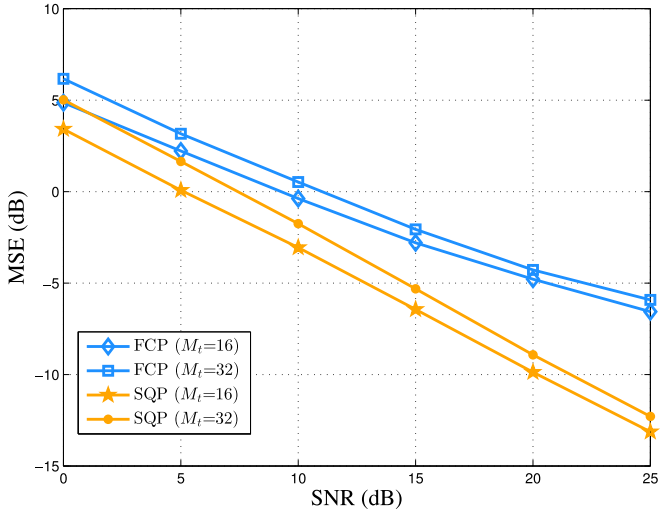

Fig. 8. MSE for training sequences obtained with FCP and SQP in the downlink transmission: $N = LM_t/2$, (67) and (68).

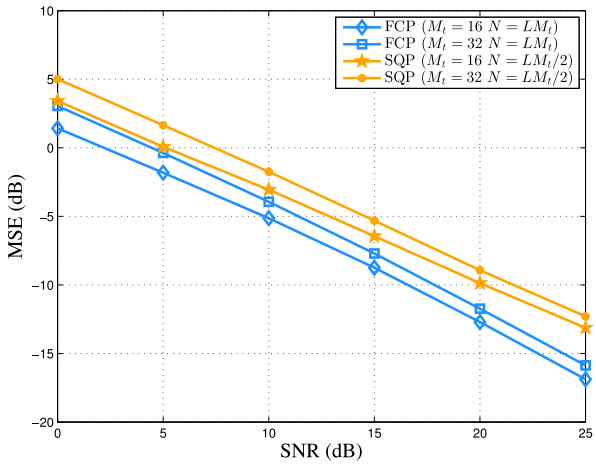

Fig. 9. MSE achieved by training sequences obtained with FCP when $N = LM_t$ and with SQP when $N = LM_t/2$: Downlink transmission with (67) and (68).

Second, we further investigate the MSE performance in the case of (67) and (68) and $N = LM_t/2$. Fig. 8 shows that better MSE performance can be obtained by the optimal training sequences of SQP than the optimal training sequences of FCP. It also can be seen that optimal training sequences obtained by FCP with $N = LM_t$ outperforms SQP with $N = LM_t/2$ in Fig 9.

Example 3: In the last example, we consider optimal training sequences designed for uplink channel state estimation of large-scale MIMO-OFDM. In the uplink transmission, we set $M_t = 2$, $M_r \in \{16, 32, 64\}$, $(\bar{\theta}_{h,0}, \bar{\theta}_{h,1}, \bar{\theta}_{h,2}, \bar{\theta}_{h,3}) = (13^\circ, 16^\circ, 20^\circ, 24^\circ)$, $(\bar{\theta}_{v,0}, \bar{\theta}_{v,1}, \bar{\theta}_{v,2}, \bar{\theta}_{v,3}) = (20^\circ, 25^\circ, 28^\circ, 33^\circ)$, and $(\bar{\theta}_{t,0}, \bar{\theta}_{t,1}, \bar{\theta}_{t,2}, \bar{\theta}_{t,3}) = (290^\circ, 300^\circ, 315^\circ, 320^\circ)$. Since the length $N = LM_t = 8$ satisfying (8) is still a small number, we use it in this uplink example. The optimal training sequence is given by (25), where $\mathbf{X}_{\text{opt}} \in \mathbf{C}^{2\times 2}$ is the optimal solution of (24). Observe that the dimension 2×2 of the variable $\mathbf{X}$ is very small in comparison with the

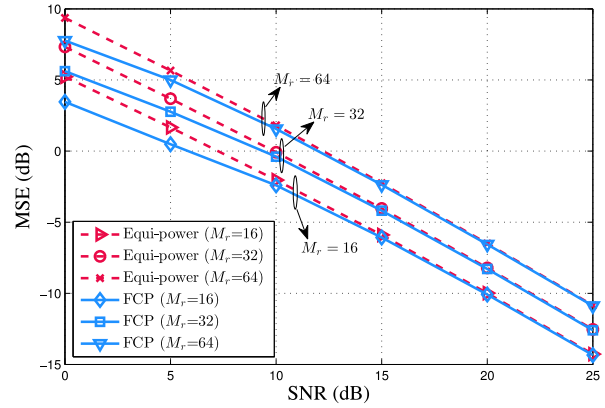

Fig. 10. MSE achieved by training sequences obtained with FCP and equi-power training sequences when $N = LM_t$: Uplink transmission.

dimension of the involved matrices in (24). Therefore it is very likely that the equi-power solution $\mathbf{X} = N\mathbf{I}_2$ provides almost optimal solution of (24). The comparison of MSE in Fig. 10 for optimal training sequences obtained with FCP and training sequences $\mathbf{S} = \sqrt{N}\mathbf{Q}$ confirms this expectation. Therefore, like massive MIMO uplink transmission [35], there is no need to solve (24) as the equi-power training sequence $\sqrt{N}\mathbf{Q}$ is practically optimal. It can be stated that the optimal training sequence design is much more challenging for the downlink than uplink of large-scale MIMO-OFDM systems.

VII. Conclusions

In this paper, we have considered the optimal design of training sequences for channel estimation in spatially-correlated large-scale MIMO-OFDM systems. The design objective is to minimize the mean square error of the channel estimate. For full-length training sequences, we showed that the design problem can be transformed into a large-scale semi-definite program. We have also developed fast computation procedures for its solutions. For reduced-length training sequences the design problem is highly nonconvex but is shown to be efficiently addressed by our proposed sequential convex programming of low computational complexity. Numerical results confirm the superior performances of the proposed training sequences.

Appendix
Useful Lemmas

Lemma 3: For $\mathbf{A}, \mathbf{B} \succeq 0, \mathbf{C}$ and $\mathbf{X} \succ 0$ and $\bar{\mathbf{X}} \succeq 0$ of appropriate size, it is true that

$$\begin{aligned}
&\text{tr}[\mathbf{C}(\mathbf{I} + (\mathbf{A}\mathbf{X}\mathbf{A}^H) \otimes \mathbf{B})^{-1}\mathbf{C}^H] \\
&\leq \text{tr}[\mathbf{C}(\mathbf{I} + (\mathbf{A}\bar{\mathbf{X}}\mathbf{A}^H) \otimes \mathbf{B})^{-1}\mathbf{C}^H] \\
&\quad + \text{tr}[(\mathbf{I} + (\mathbf{A}\bar{\mathbf{X}}\mathbf{A}^H) \otimes \mathbf{B})^{-1}\mathbf{C}^H \\
&\quad \times \mathbf{C}(\mathbf{I} + (\mathbf{A}\bar{\mathbf{X}}\mathbf{A}^H) \otimes \mathbf{B})^{-1} \\
&\quad \times ((\mathbf{A}(\bar{\mathbf{X}}\mathbf{X}^{-1}\bar{\mathbf{X}} - \bar{\mathbf{X}})\mathbf{A}^H) \otimes \mathbf{B})]. \qquad (69)
\end{aligned}$$

Proof: Function $F(\mathbf{X}) := \text{tr}[\mathbf{C}(\mathbf{I}+(\mathbf{A}\mathbf{X}^{-1}\mathbf{A}^H)\otimes\mathbf{B})^{-1}\mathbf{C}^H]$ is concave in $\mathbf{X} \succ 0$ according to [20, Lemma 1].

Therefore [36]

$$\begin{aligned} F(\mathbf{X}) &\leq F(\bar{\mathbf{X}}) + \langle \nabla F(\bar{\mathbf{X}}), \mathbf{X} - \bar{\mathbf{X}} \rangle \\ &= F(\bar{\mathbf{X}}) + \mathsf{tr}[(\mathbf{I} + (\mathbf{A}\bar{\mathbf{X}}^{-1}\mathbf{A}^H) \otimes \mathbf{B})^{-1}\mathbf{C}^H\mathbf{C}(\mathbf{I} \\ &\quad + (\mathbf{A}\bar{\mathbf{X}}^{-1}\mathbf{A}^H) \otimes \mathbf{B})^{-1}((\mathbf{A}\bar{\mathbf{X}}^{-1}(\mathbf{X} - \bar{\mathbf{X}})\bar{\mathbf{X}}^{-1} \\ &\quad \times \mathbf{A}^H) \otimes \mathbf{B})] \end{aligned}$$

for all $\mathbf{X} \succ 0$ and $\bar{\mathbf{X}} \succ 0$. Replacing $\mathbf{X} \to \mathbf{X}^{-1}$ and $\bar{\mathbf{X}} \to \bar{\mathbf{X}}^{-1}$ into the last inequality leads to

$$\begin{aligned} &\mathsf{tr}[\mathbf{C}(\mathbf{I} + (\mathbf{A}\mathbf{X}\mathbf{A}^H) \otimes \mathbf{B})^{-1}\mathbf{C}^H] \\ &\leq \mathsf{tr}[\mathbf{C}(\mathbf{I} + (\mathbf{A}\bar{\mathbf{X}}\mathbf{A}^H) \otimes \mathbf{B})^{-1}\mathbf{C}^H] \\ &\quad + \mathsf{tr}[(\mathbf{I} + (\mathbf{A}\bar{\mathbf{X}}\mathbf{A}^H) \otimes \mathbf{B})^{-1}\mathbf{C}^H \\ &\mathbf{C}(\mathbf{I} + (\mathbf{A}\bar{\mathbf{X}}\mathbf{A}^H) \otimes \mathbf{B})^{-1} \\ &((\mathbf{A}(\bar{\mathbf{X}}\mathbf{X}^{-1}\bar{\mathbf{X}} - \bar{\mathbf{X}})\mathbf{A}^H) \otimes \mathbf{B})] \\ &\forall\ \mathbf{X} \succ 0, \bar{\mathbf{X}} \succ 0. \end{aligned}$$

Note that the right hand side does not involve the inverse of $\bar{\mathbf{X}}$ so by using the perturbation argument $\bar{\mathbf{X}} \to \bar{\mathbf{X}}(\epsilon) = \bar{\mathbf{X}} + \epsilon I \succ 0$ for $\epsilon \to 0+$ we can see that (69) is still true for $\bar{\mathbf{X}} \succeq 0$. ■

Lemma 4: For $\mathbf{Y} \succ 0$, $\bar{\mathbf{Y}} \succ 0$ and $\mathbf{S}$, $\bar{\mathbf{S}}$ of appropriate dimension, one has

$$\mathsf{tr}(\mathbf{S}^H\mathbf{Y}^{-1}\mathbf{S}) \geq 2\Re\{\mathsf{tr}(\bar{\mathbf{S}}^H\bar{\mathbf{Y}}^{-1}\mathbf{S})\} - \mathsf{tr}(\bar{\mathbf{S}}^H\bar{\mathbf{Y}}^{-1}\mathbf{Y}\bar{\mathbf{Y}}^{-1}\bar{\mathbf{S}}) \tag{70}$$

Proof: Using [37], function $f(\mathbf{S}, \mathbf{Y}) := \mathsf{tr}(\mathbf{S}^H\mathbf{Y}^{-1}\mathbf{S})$ is convex on the domain $\{(\mathbf{S}, \mathbf{Y}) : \mathbf{Y} \succ 0\}$. Therefore [36], $f(\mathbf{S}, \mathbf{Y}) \geq f(\bar{\mathbf{S}}, \bar{\mathbf{Y}}) + \langle \nabla f(\bar{\mathbf{S}}, \bar{\mathbf{Y}}), (\mathbf{S}, \mathbf{Y}) - (\bar{\mathbf{S}}, \bar{\mathbf{Y}}) \rangle = 2\Re\{\mathsf{tr}(\bar{\mathbf{S}}^H\bar{\mathbf{Y}}^{-1}\mathbf{S})\} - \mathsf{tr}(\bar{\mathbf{S}}^H\bar{\mathbf{Y}}^{-1}\mathbf{Y}\bar{\mathbf{Y}}^{-1}\bar{\mathbf{S}})$, which proves (70). ■

REFERENCES

[1] Y. H. Nam *et al.*, "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 162–179, Jun. 2013.

[2] Y. Kim *et al.*, "Full-dimension MIMO (FD-MIMO): The next evolution of MIMO in LTE systems," *IEEE Wireless. Commun.*, vol. 21, no. 2, pp. 26–33, Jun. 2014.

[3] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving massive MIMO spectral efficiency with a not so large number of antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3226–3239, Sep. 2012.

[4] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 264–273, Feb. 2013.

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

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

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

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

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

[10] X. Rao and V. K. N. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, Dec. 2014.

[11] S. Noh, M. D. Zoltowski, Y. Sung, and D. J. Love, "Pilot beam pattern design for channel estimation in massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 787–801, Oct. 2014.

[12] J. Choi, D. J. Love, and P. Bidigare, "Downlink training techniques for FDD massive MIMO systems: Open-loop and closed-loop training with memory," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 802–8014, Oct. 2014.

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

[14] A. Decurninge, M. Guillaud, and D. T. M. Slock, "Channel covariance estimation in massive MIMO frequency division duplex systems," in *Proc. IEEE Globecom Workshop*, 2015, pp. 1–6.

[15] B. Tomasi and M. Guillaud, "Pilot length optimization for spatially correlated multi-user MIMO channel estimation," arXiv:1602.05480, 2016.

[16] Y. G. Li, "Simplified channel estimation for OFDM systems with multiple transmit antennas," *IEEE Trans. Wireless Commun.*, vol. 1, no. 1, pp. 67–75, Aug. 2002.

[17] H. Zhang, Y. G. Li, A. Reid, and J. Terry, "Optimum training symbol design for MIMO OFDM in correlated fading channels," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2343–2347, Sep. 2006.

[18] Y. G. Li, N. Seshadi, and S. Ariyavisitakul, "Channel estimation for OFDM systems with transmitter diversity in mobile wireless channels," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 3, pp. 461–471, Mar. 1999.

[19] H. D. Tuan, V. Luong, and H. H. Nguyen, "Optimization of training sequences for spatially correlated MIMO-OFDM," in *Proc. 2009 IEEE Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 2681–2684.

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

[21] L. Dai, Z. Wang, and Z. Yang, "Spectrally efficient time-frequency training OFDM for mobile large-scale MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 251–263, Feb. 2013.

[22] Z. Gao, L. Dai, Z. Lu, C. Yuen, and Z. Wang, "Super-resolution sparse MIMO-OFDM channel estimation based on spatial and temporal correlations," *IEEE Commun. Lett.*, vol. 18, no. 7, pp. 1266–1269, Jul. 2014.

[23] H. Bölcskei, D. Gesbert, and A. J. Paulraj, "On the capacity of OFDM-based spatial multiplexing systems," *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 225–234, Feb. 2002.

[24] H. Bölcskei, "Principles of MIMO-OFDM wireless systems," in *CRC Handbook on Signal Processing for Communications*, M. Ibnkahla, Ed. Boca Raton, FL, USA: CRC Press, 2004.

[25] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[26] A. Goldsmith, *Wireless Communications*. Cambridge, U.K: Cambridge Univ. Press, 2005.

[27] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[28] I. Barhumi, G. Leus, and M. Moonen, "Optimal training design for MIMO OFDM systems in mobile wireless channels," *IEEE Trans. Signal Process.*, vol. 51, no. 6, pp. 1615–1624, Jun. 2003.

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

[30] H. D. Tuan, D. H. Pham, B. Vo, and T. Q. Nguyen, "Entropy of general Gaussian distributions and MIMO channel capacity maximizing precoder and decoder," in *Proc. IEEE Conf. Acoust., Speech, Signal Process.*, Apr. 2007, vol. 3, pp. 325–328.

[31] R. A. Horn and C. R. Jonhson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[32] G. Strang and T. Q. Nguyen, *Wavelets and Filter Banks*. Cambridge, MA, USA: Wellesley-Cambridge, 1997.

[33] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programms," *Oper. Res.*, vol. 26, pp. 681–683, Jul. 1978.

[34] D. Shiu, G. Foschini, M. Gans, and J. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 502–513, Mar. 2000.

[35] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[36] H. Tuy, *Convex Analysis and Global Optimization*. Norwell, MA, USA: Kluwer, 1997.
[37] U. Rashid, H. D. Tuan, H. H. Kha, and H. H. Nguyen, "Joint optimization of source precoding and relay beamforming in wireless MIMO relay networks," *IEEE Trans. Commun.*, vol. 62, no. 2, pp. 488–499, Feb. 2014.