Training Signal Design for Estimation of Correlated MIMO Channels with Colored Interference

Yong Liu [†], Tan F. Wong [*†], and William. W. Hager [‡]

[†] Wireless Information Networking Group

Department of Electrical & Computer Engineering

University of Florida, Gainesville, Florida 32611-6130

Phone: 352-392-2665, Fax: 352-392-0044

yl295@ecel.ufl.edu, twong@ece.ufl.edu

[‡] Department of Mathematics

University of Florida, Gainesville, Florida 32611-8105

Phone: 352-392-0281, Fax: 352-392-8357

hager@math.ufl.edu

EDICS: MSP-CEST, SPC-APPL

Abstract

In this paper, we study the problem of estimating correlated multiple-input multiple-output (MIMO) channels in the presence of colored interference. The linear minimum mean square error (MMSE) channel estimator is derived and the optimal training sequences are designed based on the MSE of channel estimation. We propose an algorithm to estimate the long-term channel statistics in the construction of the optimal training sequences. We also design an efficient scheme to feed back the required information to the transmitter where we can approximately construct the optimal sequences. Numerical results show that the optimal training sequences provide substantial performance gain for channel estimation when compared with other training sequences.

Index Terms

MIMO system, channel estimation, optimal training sequence, MSE

This work was supported by the National Science Foundation under Grants 0203270 and ANI-0020287.

I. Introduction

Many multiple antenna communication systems are designed to perform coherent detection that requires channel state information (CSI) in the demodulation process. For practical wireless communication systems, it is common that the channel parameters are estimated by sending known training symbols to the receiver. The performance of this training-based channel estimation scheme depends on the design of training signals which has been extensively investigated in the literature [1]-[9].

It is well known that imperfect knowledge of the channel has a detrimental effect on the achievable rate it can sustain [10]. Training sequences can be designed based on information theoretic metrics such as the ergodic capacity and outage capacity of a MIMO channel [1] [2] [3]. The mean square error (MSE) is another commonly used performance measure for channel estimation. Many works [4]-[9] have been carried out to investigate the training sequence design problem based on MSE for MIMO fading channels. In [5], the authors study the problem of training sequence design for multiple-antenna systems over flat fading MIMO channels in the presence of colored interference. The MIMO channels are assumed to be spatially white, i.e., there is no correlation among the transmit and receive antennas. The optimal training sequences are designed to minimize the channel estimation MSE under a total transmit power constraint. The optimal training sequence design result implied that we should intentionally assign transmit power to the subspace with less interference. In [6], the problem of transmit signal design is investigated for the estimation of spatial correlated MIMO Rayleigh flat fading channels. The optimal training signal is designed to optimize two criteria: the minimization of the channel estimation MSE and the maximization of the conditional mutual information (CMI) between the channel and the received signal. The authors adopted the virtual channel representation model [11] for MIMO correlated channels. It is shown that the optimal training signal should be transmitted along the strong directions in which more scatters are present. The power transmitted along these directions is determined by the water-filling solutions based on the minimum MSE and maximum CMI criteria.

In the present work, we investigate the problem of estimating correlated MIMO channels with colored interference. We adopt the correlated MIMO channel model from [12] [13] which expresses the channel matrix as a product of the receive correlation matrix, a white “channel” matrix with identically and independent distributed (i.i.d.) entries, and the transmit correlation matrix. This model implies that transmit and receive correlation can be separated. This fact has been verified by field measurements. The colored interference model used here is more suitable than the white noise model when jamming signals and/or co-channel interference are present in the wireless communication system. We consider an interference

limited wireless communication system, i.e., we ignore the thermal noise which is insignificant compared to the interference. Then we show that the covariance matrix of the interference has a Kronecker product form which implies that the temporal and spatial correlations of the interference are separable. The channel estimation MSE is used as a performance metric for the design of training sequences. The optimization problem formulated here minimizes the channel estimation MSE under a power constraint. This is a generalization of two previous optimization problems which are encountered widely in the signal processing area [5], [8], [9], [14].

In [7], the authors encounter essentially the same optimization problem in a different form. According to previous optimization results for the special case in [8], the authors choose to optimize the training sequence matrix in a particular set of matrices which have the same solution structure and eigenvector ordering as our solution. Here we rigorously prove that this particular solution structure and eigenvector ordering result are optimal for arbitrary matrices under the power constraint. The optimal training sequence design assigns more power to the transmission direction constructed by the eigen-direction with larger channel gains and the interference subspace with less interference. In order to implement the channel estimator and construct the optimal training sequences, we propose an algorithm to estimate long-term channel statistics and design an efficient feedback scheme so that we can approximately construct the optimal sequences at the transmitter. Numerical results show that with the optimal training sequences, the channel estimation MSE can be reduced substantially when compared with the use of other training sequences.

II. System Model

We consider a single user link with multiple interferers. The desired user has n_t transmit antennas and n_r receive antennas. We assume that there are M_I interfering signals and the ith interferer has n_i transmit antennas. The MIMO channel is assumed to be quasi-static (block fading) in that it varies slowly enough to be considered invariant over a block. However, the channel changes to independent values from block to block. We assume that the users employ a frame-based transmission protocol which comprises training and payload data. The received baseband signals at the receive antennas during the training period are given in matrix form by

$$\mathbf{Y} = \mathbf{H}\mathbf{S}^T + \underbrace{\sum_{i=1}^{M_I} \mathbf{H}_i \mathbf{S}_i^T}_{\mathbf{E}}, \tag{1}$$

where T denotes the transpose of a matrix. The $n_r \times n_t$ matrix $\mathbf{H}$ and the $n_r \times n_i$ matrix $\mathbf{H}_i$ are the channel gain matrices from the transmitter and the ith interferer to the receiver, respectively. $\mathbf{S}$ is

the $N \times n_t$ training symbol matrix known to the receiver for estimating the channel gain matrix $\mathbf{H}$ of the desired user during the training period. N is the number of training symbols from each transmit antenna and N is usually much larger than n_t. $\mathbf{S}_i$ is the $N \times n_i$ interference symbol matrix from the ith interferer. We assume that the elements in $\mathbf{S}_i$ are identically distributed zero-mean complex random variables, correlated across both space and time. The interference processes are assumed to be wide-sense stationary in time. We consider an interference limited wireless communication system. Hence we ignore the effect of the thermal noise [15].

We adopt the correlated MIMO channel model [12], [13] which models the channel gain matrix $\mathbf{H}$ as $\mathbf{H} = \mathbf{R}_r^{1/2}\mathbf{H}_w\mathbf{R}_t^{1/2}$, where $\mathbf{R}_t$ models the correlation between the transmit antennas and $\mathbf{R}_r$ models the correlation between the receive antennas, respectively. We assume that both $\mathbf{R}_r$ and $\mathbf{R}_t$ are of full rank. The notation $(\cdot)^{1/2}$ stands for the Hermitian square root of a matrix. $\mathbf{H}_w$ is a matrix whose elements are independent and identical distributed zero-mean circular-symmetric complex Gaussian random variables with unit variance. Let $\mathbf{h}_w = vec(\mathbf{H}_w)$, where $vec(\mathbf{X})$ is the vector obtained by stacking the columns of $\mathbf{X}$ on top of each other, then we have $\mathbf{h} = vec(\mathbf{H}) = (\mathbf{R}_t^{1/2} \otimes \mathbf{R}_r^{1/2})\mathbf{h}_w$, with $\mathbf{h} \sim \mathcal{CN}(0, \mathbf{R}_t \otimes \mathbf{R}_r)$ where $\mathcal{CN}$ denotes complex Gaussian distribution, $\otimes$ denotes the Kronecker product, and $\sim$ means "distributed as". Similarly, the channel gain matrix from the ith interferer to the receiver is $\mathbf{H}_i = \mathbf{R}_r^{1/2}\mathbf{H}_{wi}\mathbf{R}_{ti}^{1/2}$ and $\mathbf{h}_i = vec(\mathbf{H}_i) = (\mathbf{R}_{ti}^{1/2} \otimes \mathbf{R}_r^{1/2})\mathbf{h}_{wi}$. Using the vec operator, we can write the received signal in (1) in vector form as

$$\mathbf{y} = vec(\mathbf{Y}) = (\mathbf{S} \otimes \mathbf{I}_{n_r})\mathbf{h} + \mathbf{e}, \tag{2}$$

where $\mathbf{I}_{n_r}$ denotes the $n_r \times n_r$ identity matrix and $\mathbf{e} = vec(\mathbf{E})$.

To derive the linear MMSE channel estimator, we need the following lemma.

Lemma 2.1: $\mathrm{E}(\mathbf{e}) = 0$ and the covariance matrix of $\mathbf{e}$ is

$$\mathrm{E}(\mathbf{e}\mathbf{e}^H) = \sum_{i=1}^{M_I} \mathbf{Q}_{Ni} \otimes \mathbf{R}_r = \mathbf{Q}_N \otimes \mathbf{R}_r$$

where

$$\mathbf{Q}_{Ni} = \begin{bmatrix} \sum_{k=1}^{n_i} R_{k,k}^i(0) & \cdots & \sum_{k=1}^{n_i} R_{k,k}^i(N-1) \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{n_i} R_{k,k}^i(N-1) & \cdots & \sum_{k=1}^{n_i} R_{k,k}^i(0) \end{bmatrix},$$

$\mathbf{Q}_N = \sum_{i=1}^{M_I} \mathbf{Q}_{Ni}$, $\hat{\mathbf{S}}_i = \mathbf{S}_i\mathbf{R}_{ti}^{1/2}$ which is called the transformed interference symbol matrix, $R_{k,k}^i(\tau) = E\left[[\hat{\mathbf{S}}_i]_{m,k}[\hat{\mathbf{S}}_i]_{m+\tau,k}\right]$ represents the correlation between $[\hat{\mathbf{S}}_i]_{m,k}$ and $[\hat{\mathbf{S}}_i]_{m+\tau,k}$, and H denotes the conjugate transpose of a matrix.

Proof: See Appendix A. ■

We note that $\mathbf{Q}_N$ captures the temporal correlation of the interference and $\mathbf{R}_r$ represents the spatial correlation. The covariance matrix of the interference has the Kronecker product form which implies that the temporal and spatial correlations of the interference are separable.

Since (2) is a linear model, based on the Bayesian Gauss-Markov Theorem [16], the linear minimum mean square error estimator (LMMSE) for $\mathbf{h}$ is given as:

$$\begin{aligned}\hat{\mathbf{h}} &= [(\mathbf{S}^H \otimes \mathbf{I}_{n_r})(\mathbf{Q}_N \otimes \mathbf{R}_r)^{-1}(\mathbf{S} \otimes \mathbf{I}_{n_r}) + (\mathbf{R}_t \otimes \mathbf{R}_r)^{-1}]^{-1}(\mathbf{S}^H \otimes \mathbf{I}_{n_r})(\mathbf{Q}_N \otimes \mathbf{R}_r)^{-1}\mathbf{y} \\ &= [(\mathbf{S}^H\mathbf{Q}_N^{-1}\mathbf{S} + \mathbf{R}_t^{-1})^{-1}\mathbf{S}^H\mathbf{Q}_N^{-1} \otimes \mathbf{I}_{n_r}]\mathbf{y}.\end{aligned}$$

Using the equality $vec(\mathbf{AYB}) = (\mathbf{B}^T \otimes \mathbf{A})vec(\mathbf{Y})$, we can rewrite the channel estimator in the more compact matrix form as

$$\hat{\mathbf{H}} = \mathbf{Y}\left[(\mathbf{S}^H\mathbf{Q}_N^{-1}\mathbf{S} + \mathbf{R}_t^{-1})^{-1}\mathbf{S}^H\mathbf{Q}_N^{-1}\right]^T.$$

Hence the channel estimator does not depend on the receive channel correlation matrix $\mathbf{R}_r$.

The performance of the channel estimator is measured by the estimation error $\epsilon = \mathbf{h} - \hat{\mathbf{h}}$ whose mean is zero and whose covariance matrix is

$$\begin{aligned}\mathbf{C}_\epsilon &= E[(\mathbf{h} - \hat{\mathbf{h}})(\mathbf{h} - \hat{\mathbf{h}})^H] \\ &= [(\mathbf{S}^H \otimes \mathbf{I}_{n_r})(\mathbf{Q}_N \otimes \mathbf{R}_r)^{-1}(\mathbf{S} \otimes \mathbf{I}_{n_r}) + (\mathbf{R}_t \otimes \mathbf{R}_r)^{-1}]^{-1} \\ &= (\mathbf{S}^H\mathbf{Q}_N^{-1}\mathbf{S} + \mathbf{R}_t^{-1})^{-1} \otimes \mathbf{R}_r.\end{aligned}$$

The diagonal elements of the error covariance matrix $\mathbf{C}_\epsilon$ yields the minimum Bayesian MSE and their sum is usually referred to as the total MSE. The total MSE is a commonly used performance measure for MIMO channel estimation. By using the fact that $\text{tr}(\mathbf{A} \otimes \mathbf{B}) = \text{tr}\mathbf{A}\text{tr}\mathbf{B}$ where tr denotes the trace of a matrix, we have

$$\text{tr}(\mathbf{C}_\epsilon) = \text{tr}((\mathbf{S}^H\mathbf{Q}_N^{-1}\mathbf{S} + \mathbf{R}_t^{-1})^{-1} \otimes \mathbf{R}_r) = \text{tr}((\mathbf{S}^H\mathbf{Q}_N^{-1}\mathbf{S} + \mathbf{R}_t^{-1})^{-1})\text{tr}(\mathbf{R}_r).$$

Thus the minimization of the total MSE over training sequences does not depend on the receive channel correlation matrix. Only the temporal interference correlation matrix $\mathbf{Q}_N$ and the transmit correlation matrix $\mathbf{R}_t$ need to be considered in obtaining the optimal training sequences.

III. Optimal Training Sequence Design

In the section, we investigate the problem of designing optimal training sequence for the channel estimation approach considered in the previous section. With the total MSE as the performance measure,

the optimization of training sequences can be formulated as follows

$$\begin{aligned} \min_{\mathbf{S}} \quad & \text{tr}(\mathbf{S}^H \mathbf{Q}_N^{-1} \mathbf{S} + \mathbf{R}_t^{-1})^{-1} \\ \text{subject to} \quad & \text{tr}\{\mathbf{S}^H \mathbf{S}\} \leq P \end{aligned} \tag{3}$$

where $\text{tr}\{\mathbf{S}^H \mathbf{S}\} \leq P$ specifies the power constraint.

Some special cases of this optimization problem (with either $\mathbf{Q}_N$ or $\mathbf{R}_t$ equal to the identity matrix) have been encountered in joint linear transmitter-receiver design [8], [14], [17] and training sequence design for channel estimation in MIMO systems [5], [9]. The solution in the special case $\mathbf{R}_t = \mathbf{I}$, found for example in [5] and [14], can be expressed in terms of the eigenvalues and eigenvectors of $\mathbf{Q}_N$ and a Lagrange multiplier associated with the power constraint. Similarly, the solution in the special case $\mathbf{Q}_N = \mathbf{I}$, found for example in [6], [8] and [9], can be expressed in terms of the eigenvalues and eigenvectors of $\mathbf{R}_t$ and a Lagrange multiplier associated with the power constraint. The optimization of the MSE problem introduced here is more difficult. We will show that (3) has a solution that can be expressed as $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$ where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices of eigenvectors for $\mathbf{Q}_N$ and $\mathbf{R}_t$ respectively, and $\mathbf{\Sigma}$ is diagonal. Solving (3) involves computing diagonalizations of $\mathbf{Q}_N$ and $\mathbf{R}_t$, and finding an ordering for the columns of $\mathbf{U}$ and $\mathbf{V}$.

The optimal training sequences should be designed according to the following theorem which summarizes the solution to the optimization problem (3).

Theorem 3.1: Suppose that $\mathbf{Q}_N$ and $\mathbf{R}_t$ are Hermitian positive definite matrices, and let $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$ and $\mathbf{V}\mathbf{\Delta}\mathbf{V}^H$ be the associated diagonalizations where the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal eigenvectors, the corresponding eigenvalues $\{\lambda_i\}$ of $\mathbf{Q}_N$ are arranged in an increasing order, and the corresponding eigenvalues $\{\delta_i\}$ of $\mathbf{R}_t$ are arranged in a decreasing order. Then the optimal solution of (3) is given by

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{4}$$

where $\mathbf{\Sigma}$ specifies the power allocation, which is a diagonal matrix with diagonal elements given by

$$\sigma_i = \max\left\{\sqrt{\frac{\lambda_i}{\mu}} - \frac{\lambda_i}{\delta_i},\ \ 0\right\}^{1/2} \qquad \text{for } 1 \leq i \leq M \triangleq \min\{n_t, N\}, \tag{5}$$

with the parameter μ chosen so that $\sum_{i=1}^{M} \sigma_i^2 = P$.

Proof: See Appendix B. ■

With the optimal training sequences, the channel estimator simplifies to $\hat{\mathbf{H}} = \mathbf{Y}\mathbf{U}_M^* \mathbf{\Gamma} \mathbf{V}_M^T$, where $\mathbf{\Gamma} = \text{diag}\{\gamma_1, \ldots, \gamma_M\}$ with $\gamma_i = \frac{\sigma_i \delta_i}{\sigma_i^2 \delta_i + \lambda_i}$, the columns of $\mathbf{U}_M$ are the eigenvectors of $\mathbf{Q}_N$ corresponding to

its M smallest eigenvalues, and the columns of $\mathbf{V}_M$ are the eigenvectors of $\mathbf{R}_t$ corresponding to its M largest eigenvalues.

The design of the optimal training sequences summarized in the above theorem has a clear physical interpretation. Each eigenvector of the transmit correlation matrix $\mathbf{R}_t$ represents a transmit direction and the associated eigenvalue indicates the channel gain in that direction. More power should be assigned to the signals transmitted along the directions with larger channel gains. On the other hand, each eigenvector of the interference temporal correlation matrix $\mathbf{Q}_N$ represents an interference subspace and the corresponding eigenvalue indicates the amount of interference in that subspace. Hence, we should choose the subspaces with the least amount of interference for transmission. The power assignment is determined by the water-filling argument under a finite power constraint.

To facilitate the understanding of the water-filling interpretation for the power assignment, we can rewrite the optimal power assignment solution in an alternative way as:

$$\tilde{\sigma}_i = \max\left\{\tilde{\mu} - \frac{\sqrt{\lambda_i}}{\delta_i},\ \ 0\right\}^{1/2} \qquad \text{for } 1 \le i \le M,$$

with $\tilde{\mu} = 1/\sqrt{\mu}$, $\tilde{\sigma}_i = \sigma_i/\lambda_i^{\frac{1}{4}}$, and $\tilde{\mu}$ chosen so that $\sum_{i=1}^{M} \sqrt{\lambda_i}\tilde{\sigma}_i^2 = P$, where $\tilde{\mu}$ represents the water level, $\{\sqrt{\lambda_i}/\delta_i\}$ specifies the depth profile which can be visualized as the surface over which the water is poured, and the volume of each subchannel is weighted by $\sqrt{\lambda_i}$ for the calculation of the total water volume.

A simple algorithm [18], which terminates in at most N steps, can be used to calculate the optimal power assignment solution:

Input: set of pairs $\{(\lambda_i, \delta_i)\}$ and the power constraint P.

Output: the water level $\tilde{\mu}$ and $\{\tilde{\sigma}_i\}$.

1. Choose $\tilde{\mu}$ as the maximum of $\{\frac{\sqrt{\lambda_i}}{\delta_i}\}$ and set $L_{new} = M + 1$.

2. Set $L_{old} = L_{new}$. Let I be the set of indices with $\tilde{\mu} - \frac{\sqrt{\lambda_i}}{\delta_i} \ge 0$ and L_{new} be the cardinal number of the set I. Compute $\tilde{\mu} = (P + \sum_{i\in I} \frac{\lambda_i}{\delta_i})/\sum_{i\in I}\sqrt{\lambda_i}$.

3. If $L_{new} < L_{old}$, go to step 2.

4. Compute $\tilde{\sigma}_i = \max\left\{\tilde{\mu} - \frac{\sqrt{\lambda_i}}{\delta_i},\ \ 0\right\}^{1/2}$, and output $\tilde{\mu}$ and $\{\tilde{\sigma}_i\}$.

IV. Estimation of Channel Statistics and Feedback Design

To implement the channel estimator and construct the optimal training sequences, we need knowledge of the transmit antenna correlation matrix $\mathbf{R}_t$ and the interference covariance matrix $\mathbf{Q}_N$ at both the receiver and transmitter. Since these two matrices are long-term channel characteristics, they can be estimated by

using the observed training signals at the receiver and then fed back to the transmitter for the construction of the optimal training sequences. In this section, we propose an algorithm to estimate these long-term channel characteristics and design an efficient feedback scheme so that we can approximately construct the optimal training sequences at the transmitter.

Let us assume that the training signal matrix $\mathbf{S}$ is sent over a sequence of K packets. During the transmission of each packet, the channel is assumed to be invariant. Then the received training signals for the nth packet are given as

$$\begin{aligned}\mathbf{y}(n) &= (\mathbf{S}\otimes\mathbf{I}_{n_r})\mathbf{h}(n)+\mathbf{e}(n)\\ &= (\mathbf{S}\mathbf{R}_t^{1/2}\otimes\mathbf{R}_r^{1/2})\mathbf{h}_w(n)+\mathbf{e}(n).\end{aligned}$$

We note that the correlation matrix of the received signal also has the Kronecker product form:

$$\begin{aligned}\mathbf{R} &= E[\mathbf{y}(n)\mathbf{y}(n)^H]\\ &= (\mathbf{S}\mathbf{R}_t^{1/2}\otimes\mathbf{R}_r^{1/2})E(\mathbf{h}_w(n)\mathbf{h}_w(n)^H)(\mathbf{R}_t^{1/2}\mathbf{S}^H\otimes\mathbf{R}_r^{1/2})+E(\mathbf{e}(n)\mathbf{e}(n)^H)\\ &= \mathbf{R}_q\otimes\mathbf{R}_r,\end{aligned}$$

where $\mathbf{R}_q=\mathbf{S}\mathbf{R}_t\mathbf{S}^H+\mathbf{Q}_N$. We calculate the sample average correlation matrix of the received signal from the previous K packets:

$$\hat{\mathbf{R}}=\frac{1}{K}\sum_{n=1}^{K}\mathbf{y}(n)\mathbf{y}(n)^H.$$

If $\mathbf{e}(n)$ is Gaussian, $\hat{\mathbf{R}}$ is a sufficient statistic for the estimation of the correlation matrix $\mathbf{R}$.

If $\mathbf{R}=\mathbf{R}_q\otimes\mathbf{R}_r$, then $\mathbf{R}=\alpha\mathbf{R}_q\otimes\frac{1}{\alpha}\mathbf{R}_r$ for any $\alpha\neq 0$. Hence, $\mathbf{R}_q$ and $\mathbf{R}_r$ can not be uniquely identified from observing $\mathbf{y}(n)$. Fortunately, the channel estimator and the design of optimal sequences are invariant to scaling of the estimates of $\mathbf{R}_t$ and $\mathbf{Q}_N$ because

$$\hat{\mathbf{H}}^{'}(n)=\mathbf{Y}(n)\left[(\mathbf{S}^H(\alpha\mathbf{Q}_N)^{-1}\mathbf{S}+(\alpha\mathbf{R}_t)^{-1})^{-1}\mathbf{S}^H(\alpha\mathbf{Q}_N)^{-1}\right]^T=\hat{\mathbf{H}}(n)$$

and

$$\text{tr}((\mathbf{S}^H(\alpha\mathbf{Q}_N)^{-1}\mathbf{S}+(\alpha\mathbf{R}_t)^{-1})^{-1}=\alpha\text{tr}((\mathbf{S}^H\mathbf{Q}_N^{-1}\mathbf{S}+\mathbf{R}_t^{-1})^{-1}).$$

For the estimation of $\mathbf{R}_q$ and $\mathbf{R}_r$, we need to impose an additional constraint on $\mathbf{R}_r$. Here we force $\text{tr}(\mathbf{R}_r)=n_r$. Then an iterative flip-flop algorithm [19] [20] [21] can be used to estimate $\mathbf{R}_q$ and $\mathbf{R}_r$. If the received interference signal $\mathbf{e}(n)$ is Gaussian distributed, the flip-flop algorithm, when converges, provides the maximum likelihood estimates (MLEs) of $\mathbf{R}_q$ and $\mathbf{R}_r$ [19]. When $\mathbf{e}(n)$ is not Gaussian, the

algorithm gives the estimates of $\mathbf{R}_q$ and $\mathbf{R}_r$ in the least square sense. For fixed $\widehat{\mathbf{R}}_r(j-1)$, the MLE of $\mathbf{R}_q$ is obtained as

$$\widehat{\mathbf{R}}_q(j) = \frac{1}{n_r}\sum_{u=1}^{n_r}\sum_{v=1}^{n_r}\sigma_{uv}^r\Big\{\frac{1}{K}\sum_{n=1}^{K}\mathbf{Y}_u^T(n)\mathbf{Y}_v^*(n)\Big\} \tag{6}$$

where σ_{uv}^r is the (u,v)th element of $\widehat{\mathbf{R}}_r^{-1}(j-1)$ and $\mathbf{Y}_u(n)$ is the uth row vector of the received signal matrix $\mathbf{Y}(n)$. Similarly, for fixed $\widehat{\mathbf{R}}_q(j)$, the MLE of $\mathbf{R}_r$ is obtained as

$$\widehat{\mathbf{R}}_r(j) = \frac{1}{N}\sum_{u=1}^{N}\sum_{v=1}^{N}\sigma_{uv}^q\Big\{\frac{1}{K}\sum_{n=1}^{K}\mathbf{W}_u(n)\mathbf{W}_v^H(n)\Big\} \tag{7}$$

where σ_{uv}^q is the (u,v)th element of $\widehat{\mathbf{R}}_q^{-1}(j)$ and $\mathbf{W}_u(n)$ is the uth column vector of the received signal $\mathbf{Y}(n)$. Then to get uniquely identifiable $\mathbf{R}_q$ and $\mathbf{R}_r$, we need to scale $\widehat{\mathbf{R}}_r(j)$ to make $\text{tr}(\widehat{\mathbf{R}}_r(j)) = n_r$. We note that the terms inside the braces in (6) and (7) can be computed before the running of the iterative estimation algorithm to reduce computational complexity. To start the iterative algorithm, an initial value of either $\widehat{\mathbf{R}}_q$ or $\widehat{\mathbf{R}}_r$ should be assigned. A natural choice is to initially make $\widehat{\mathbf{R}}_r(0) = \mathbf{I}_{n_r}$. Then the iterative algorithm alternates between the calculations of $\widehat{\mathbf{R}}_q$ and $\widehat{\mathbf{R}}_r$ until convergence. While it is difficult to analytically prove that the algorithm converges to the MLE, extensive data experiments in statistics [19] show that it always converges to the MLE for situations of practical sample sizes. The convergence in our case is also verified by the numerical results in Section V.

Then we need to estimate $\mathbf{R}_t$ and $\mathbf{Q}_N$ based on $\widehat{\mathbf{R}}_q$. Before doing so, let $\mathcal{R}$ denote the range space of a matrix, $\mathcal{R}^\perp$ denote the orthogonal complementary subspace of the range of a matrix, and consider the following lemma:

Lemma 4.1: Let $\mathcal{L}$ be the linear map defined by $\mathcal{L}(\mathbf{R}_t, \mathbf{Q}_N) = \mathbf{S}\mathbf{R}_t\mathbf{S}^H + \mathbf{Q}_N$ where $\mathbf{R}_t$ and $\mathbf{Q}_N$ are Hermitian positive semi-definite matrices and $\mathbf{S}$ is of full rank. Let $\mathbb{D}$ be defined by $\mathbb{D} = \{(\mathbf{R}_t, \mathbf{Q}_N) : \mathcal{R}(\mathbf{Q}_N) \subset \mathcal{R}^\perp(\mathbf{S})\}$. Then $\mathcal{L} : \mathbb{D} \to \mathbb{C}^{N\times N}$ is one-to-one. Moreover, given any $(\mathbf{R}_t, \mathbf{Q}_N)$ and $\mathbf{R}_q = \mathcal{L}(\mathbf{R}_t, \mathbf{Q}_N)$, there exists $(\mathbf{R}_t^{'}, \mathbf{Q}_N^{'}) \neq (\mathbf{R}_t, \mathbf{Q}_N)$ such that $\mathcal{L}(\mathbf{R}_t^{'}, \mathbf{Q}_N^{'}) = \mathbf{R}_q$.

Proof: See Appendix C. ■

Based on the above lemma, we see that estimating $\mathbf{Q}_N$ and $\mathbf{R}_t$ simultaneously from $\widehat{\mathbf{R}}_q$ is not possible. We can only uniquely determine $\mathbf{Q}_N$ up to $\mathcal{R}^\perp(\mathbf{S})$ from $\mathbf{R}_q$. Fortunately, this is not much a limitation when N is large as shown in Lemma 4.2 below. Let $|\mathbf{Q}_N|_w$ be the weak norm of $\mathbf{Q}_N$ which is defined by $|\mathbf{Q}_N|_w = \sqrt{\text{tr}(\mathbf{Q}_N^H\mathbf{Q}_N)/N}$.

Lemma 4.2: With the assumption that $\mathbf{Q}_N$ is an absolutely summable Hermitian Toeplitz matrix, the difference between the two sequences of matrices $\mathbf{Q}_N$ and $\mathbf{P}_{\mathbf{S}}^\perp\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}^\perp$ approaches zero in weak norm as N increases, i.e., $\lim_{N\to\infty}|\mathbf{Q}_N - \mathbf{P}_{\mathbf{S}}^\perp\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}^\perp|_w = 0$.

Proof: See Appendix D. ■

Since $\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{R}_q\mathbf{P}_{\mathbf{S}}^{\perp} = \mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}^{\perp}$, we can estimate $\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}^{\perp}$ from $\mathbf{P}_{\mathbf{S}}^{\perp}\widehat{\mathbf{R}}_q\mathbf{P}_{\mathbf{S}}^{\perp}$. For notational simplicity, let $\mathbf{A}$ denote $\mathbf{P}_{\mathbf{S}}^{\perp}\widehat{\mathbf{R}}_q\mathbf{P}_{\mathbf{S}}^{\perp}$. Since the interference signals are wide-sense stationary in time, $\mathbf{Q}_N$ has the form of a Topelitz matrix which can be represented by a sequence $\{q_k; k = 0, \pm 1, \ldots \pm (N-1)\}$ with $[\mathbf{Q}_N]_{k,j} = q_{k-j}$. Then the (i,j)th element of $\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}^{\perp}$ is given by $\sum_l \sum_k p_{il} q_{l-k} p_{kj}$ with p_{ij} denoting the (i,j)th element of $\mathbf{P}_{\mathbf{S}}^{\perp}$. Equating the (i,j)th element of $\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}^{\perp}$ with a_{ij}, we have a set of linear equations in $\{q_k\}$. Noticing the Hermitian nature of $\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}^{\perp}$ and $\mathbf{A}$ and separating the real and imaginary parts of q_k and a_{ij}, we have N^2 linear equations with $2N-1$ unknowns in $\mathbf{q}_r = [q_0, \mathrm{Re}(q_1), \mathrm{Im}(q_1), \ldots, \mathrm{Re}(q_{N-1}), \mathrm{Im}(q_{N-1})]^T$. This set of linear equations can be solved by employing the least square approach. Then an estimate of $\mathbf{Q}_N$ which is denoted as $\widehat{\mathbf{Q}}_N$ can be constructed based on $\mathbf{q}_r$. Although this $\widehat{\mathbf{Q}}_N$ is only unique up to $\mathcal{R}^{\perp}(\mathbf{S})$, Lemma 4.2 tells us that this is not too severe a deficiency when N is large. In addition, when N is large, $\mathbf{Q}_N$ can be approximated by the circulant matrix [22] with fixed eigenvectors as:

$$\tilde{\mathbf{Q}}_N = \mathbf{F}_N \mathbf{\Psi}_N \mathbf{F}_N^H, \tag{8}$$

where $\mathbf{F}_N$ is the $N \times N$ FFT matrix and $\mathbf{\Psi}_N$ is a diagonal matrix containing eigenvalues $\{\psi_i\}$ of $\tilde{\mathbf{Q}}_N$. Note that we only require the n_t smallest eigenvalues of $\mathbf{Q}_N$ and their corresponding eigenvectors in constructing the optimal training sequences. With the circulant matrix approximation (8), it is equivalent to estimating the n_t smallest eigenvalues ψ_i and identifying the corresponding columns of $\mathbf{F}_N$. If we arrange the eigenvalues $\{\psi_i\}$ of $\tilde{\mathbf{Q}}_N$ and the eigenvalues $\{\lambda_i\}$ of $\mathbf{Q}_N$ in increasing orders, we have [23] $\lim_{N\to\infty} |\psi_i - \lambda_i| = 0$. Thus the n_t smallest eigenvalues of $\widehat{\mathbf{Q}}_N$ can be used as the estimates of the n_t smallest ψ_i's, and the corresponding columns of $\mathbf{F}_N$ are chosen as those closest (in terms of the Euclidean norm) to the eigenvectors associated with the n_t smallest eigenvalues of $\widehat{\mathbf{Q}}_N$.

The estimates of the n_t smallest ψ_i's and the n_t indices of the chosen columns of $\mathbf{F}_N$ are then fed back to the transmitter for the optimal training sequence construction. We notice that it is bandwidth efficient to just feed back these indices of $\mathbf{F}_N$ instead of the whole eigenvectors of $\widehat{\mathbf{Q}}_N$ because the number of training symbols N during the training period is usually large.

To derive an estimator of $\mathbf{R}_t$, we need to use Lemma 4.2 again. When N is large, $\mathbf{R}_q \approx \mathbf{S}\mathbf{R}_t\mathbf{S}^H + \mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}^{\perp}$, and hence $\mathbf{P}_{\mathbf{S}}\mathbf{R}_q\mathbf{P}_{\mathbf{S}} \approx \mathbf{P}_{\mathbf{S}}\mathbf{S}\mathbf{R}_t\mathbf{S}^H\mathbf{P}_{\mathbf{S}} + \mathbf{P}_{\mathbf{S}}\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{P}_{\mathbf{S}} = \mathbf{S}\mathbf{R}_t\mathbf{S}^H$. Then with a full rank $\mathbf{S}$, we can estimate the transmit channel correlation matrix $\mathbf{R}_t$ using

$$\widehat{\mathbf{R}}_t = (\mathbf{S}^H\mathbf{S})^{-1}\mathbf{S}^H\widehat{\mathbf{R}}_q\mathbf{S}(\mathbf{S}^H\mathbf{S})^{-1}.$$

V. Numerical Results

In this section, we present some numerical results to show the performance gain achieved by the optimal training sequences. We consider a MIMO system with 3 transmit antennas and 3 receive antennas. The antennas form uniform linear arrays at both the transmitter and receiver. For a small angle spread, the correlation coefficient between the ith and the jth transmit antenna [12] can be approximated as:

$$[\mathbf{R}_t]_{i,j} \approx \frac{1}{2\pi}\int_0^{2\pi} \exp\{-j2\pi|i-j|\sin\Delta\frac{d_t}{\lambda}\sin\theta\}d\theta = J_0(2\pi|i-j|\sin\Delta\frac{d_t}{\lambda}),$$

where $J_0(x)$ is the zeroth-order Bessel function of the first kind, Δ is the angle spread, d_t is the antenna spacing and λ is the wavelength of the narrow-band signal. We set $d_t = 0.5\lambda$. In the simulations, we consider two channels with different transmit channel correlations: a high spatial correlation channel with $\Delta = 5^\circ$ and a low spatial correlation channel with $\Delta = 25^\circ$. The receive correlation matrix $\mathbf{R}_r$ is calculated similarly as the transmit correlation matrix with $\Delta = 25^\circ$. We have assumed that the channel characteristics are estimated based on the observed training signals from 50 previous packets, i.e., $K = 50$.

We consider two kinds of interference: co-channel interference from other users in the same wireless system and jamming signals which are usually modeled by autoregressive (AR) random processes. We compare the channel estimation performance in terms of the total MSE for systems using different sets of training sequences. The following training sequence sets are considered for comparison: 1) the optimal training sequences described in Section III; 2) the approximate optimal training sequence constructed based on the channel and interference statistics obtained by using the proposed estimation algorithm in Section IV; 3) the temporally optimal training sequences for which the transmit channel correlation matrix is assumed to be an identity matrix and only temporal interference correlation is considered in designing the optimal training sequences (we also consider the approximate temporally optimal sequences which are constructed based on the channel statistics obtained by using the proposed algorithm); 4) the spatially optimal training sequences for which the interference is assumed to be temporally white and only transmit correlation is considered in designing the optimal training sequences (we also consider the approximate spatially optimal sequences which are constructed based on the channel statistics obtained by using the proposed algorithm); 5) binary orthogonal sequences which are generated by using the first n_t columns of the Hadamard matrix; and 6) random sequences where the training symbols are i.i.d. binary random variables with zero mean and unit variance.

A. Co-channel Interference

In a cellular wireless communication system, co-channel interference (CCI) from other cells exists due to frequency reuse. Hence, the interfering signals have the same signal format as that of the desired user. We can express the interfering signal transmitted from the ith transmit antenna of the mth interferer as

$$s_i^{(m)}(t) = \sqrt{\frac{P_m}{n_i N}} \sum_{l=-\infty}^{\infty} b_{i,l}^{(m)} \psi(t - lT - \tau_m),$$

where P_m is the transmit power of the mth interferer, and $\{b_{i,l}^{(m)}\}$ are data symbols transmitted from the ith transmit antenna of the mth interferer. The data symbols are assumed to be i.i.d. binary random variables with zero mean and unit variance. In addition, $\psi(t)$ is the symbol waveform and T is the symbol duration. It is assumed that the receiver is synchronized to the desired user but not necessarily to the interfering signals and τ_m is the symbol timing difference between the mth interferer and the desired user signal. Without loss of generality, we assume $0 \leq \tau_m < T$. The elements of the interference symbol matrix $\mathbf{S}_i$ are samples at the matched filter output at the receiver at time index jT. The (j,i)th element of $\mathbf{S}_i$ is

$$s_{j,i}^{(m)} = \sqrt{\frac{P_m}{n_i N}} \sum_{l=-\infty}^{\infty} b_{i,l}^{m} \hat{\psi}((j-l)T - \tau_m),$$

where $\hat{\psi}(t) = \int_{-\infty}^{\infty} \psi(t-s)\psi^*(s)ds$ is the autocorrelation of the symbol waveform. For the co-channel interference, the temporal interference correlation is due to the intersymbol interference in the sampled interfering signals.

In the simulations, it is assumed that there are two interfering signals with two transmit antennas in the system and the signal-to-interference ratio ($P/\sum P_m$) is set to be 0dB. The ISI-free symbol waveform with raised cosine spectrum [24] is chosen as the symbol waveform. For this case, we have

$$\psi(t) = \text{sinc}(\pi t/T) \frac{\cos(\pi\beta t/T)}{1 - 4\beta^2 t^2/T^2}.$$

We set the roll-off factor $\beta = 0.5$, $\tau_1 = 0.2T$ and $\tau_2 = 0.5T$.

In Figs. 1 and 2, we show the total channel estimation MSEs for the high spatial correlation channel and low spatial correlation channel, respectively. For both cases, the optimal sequences outperform the orthogonal sequences and random sequences significantly. For the high spatial correlation channel, the optimal sequences provide a substantial performance gain over both the spatially optimal sequences and the temporally optimal sequences. The approximate optimal sequences achieve most of the performance gain obtained by the optimal sequences. For the low spatial correlation channel, the temporally optimal sequences achieve an estimation performance similar to that achieved by the optimal sequences. These

two optimal sequences provide significant performance gains over the spatially optimal sequences. In this case, the temporal correlation has a stronger impact on channel estimation than the spatial channel correlation due to the fact that the length of training sequences N is much larger than the number of transmit antennas n_t. The MSE performance of the approximate optimal sequences is a little worse than that of the temporally optimal seqeunces because of the errors in estimating $\mathbf{Q}_N$ and $\mathbf{R}_t$. Note that the approximate temporally optimal sequences give performance that is in turn slightly worse than that given by the approximate optimal sequences.

B. Jamming Signals

We assume that there are two jammers, each with one transmit antenna, in the system. The jamming signals are modeled as two first-order AR processes driven by temporally white Gaussian processes $\{u_{i,t}\}$, i.e.,

$$s_{i,t} = \alpha_i s_{i,t-1} + u_{i,t}$$

where $s_{i,t}$ represents the jamming signal transmitted by the ith jammer at the tth time index, α_i is the temporal correlation coefficient, and $u_{i,t}$ has zero mean with variance $\sigma_{u,i}^2$ which decides the transmit power of the ith jammer. The signal-to-interference ratio is set to be 0 dB. We choose $\alpha_1 = 0.4$ and $\alpha_2 = 0.5$. In Figs. 3 and 4, we show the total channel estimation MSEs for the high spatial correlation channel and low spatial correlation channel, respectively. For the AR jammers, similar conclusions on the estimation performance achieved by different training sequences can be made as in the case of co-channel interference.

APPENDICES

A. Proof of Lemma 2.1

Let $\mathbf{E} = \sum_{i=1}^{M_I} \mathbf{E}_i = \sum_{i=1}^{M_I} \mathbf{H}_i \mathbf{S}_i^T$ and $\mathbf{e}_i = vec(\mathbf{E}_i)$. Since $\mathbf{h}_{wi} \sim \mathcal{CN}(0, \mathbf{I}_{n_r n_t})$, $\mathrm{E}(\mathbf{e}_i) = 0$. Then we have $\mathrm{E}(\mathbf{e}) = 0$. The received signal from the ith interferer can be written as

$$\mathbf{E}_i = \mathbf{H}_i \mathbf{S}_i^T = \mathbf{R}_r^{1/2} \mathbf{H}_{wi} \underbrace{\mathbf{R}_{ti}^{1/2} \mathbf{S}_i^T}_{\hat{\mathbf{S}}_i^T} = \mathbf{R}_r^{1/2} \mathbf{H}_{wi} \hat{\mathbf{S}}_i^T.$$

Since $\mathbf{S}_i$ is wide-sense stationary in time, $\hat{\mathbf{S}}_i$ is also wide-sense stationary in time. Using the vec operator, we can rewrite the interfering signal from the ith interferer as

$$\mathbf{e}_i = vec(\mathbf{E}_i) = (\mathbf{I}_N \otimes \mathbf{R}_r^{1/2}) vec(\mathbf{H}_{wi} \hat{\mathbf{S}}_i^T).$$

The covariance matrix of $\mathbf{e}_i$ is given as

$$\begin{aligned}\mathrm{E}(\mathbf{e}_i\mathbf{e}_i^H) &= \mathrm{E}[(\mathbf{I}_N \otimes \mathbf{R}_r^{1/2})vec(\mathbf{H}_{wi}\hat{\mathbf{S}}_i^T)vec(\mathbf{H}_{wi}\hat{\mathbf{S}}_i^T)^H(\mathbf{I}_N \otimes \mathbf{R}_r^{1/2})^H] \\ &= (\mathbf{I}_N \otimes \mathbf{R}_r^{1/2})\mathrm{E}[vec(\mathbf{H}_{wi}\hat{\mathbf{S}}_i^T)vec(\mathbf{H}_{wi}\hat{\mathbf{S}}_i^T)^H](\mathbf{I}_N \otimes \mathbf{R}_r^{1/2}).\end{aligned}$$

Let $\mathbf{e}_i^{'} = vec(\mathbf{H}_{wi}\hat{\mathbf{S}}_i^T)$, it is easy to see that the covariance matrix of $\mathbf{e}_i^{'}$ is

$$\begin{aligned}\mathrm{E}[\mathbf{e}_i^{'}\mathbf{e}_i^{'H}] &= \begin{bmatrix} \sum_{k=1}^{n_i} R_{k,k}^i(0)\mathbf{I}_r & \dots & \sum_{k=1}^{n_i} R_{k,k}^i(N-1)\mathbf{I}_r \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{n_i} R_{k,k}^i(N-1)\mathbf{I}_r & \dots & \sum_{k=1}^{n_i} R_{k,k}^i(0)\mathbf{I}_r \end{bmatrix} \\ &= \mathbf{Q}_{Ni} \otimes \mathbf{I}_{n_r}.\end{aligned}$$

Then we have

$$\begin{aligned}\mathrm{E}[\mathbf{e}_i\mathbf{e}_i^H] &= (\mathbf{I}_N \otimes \mathbf{R}_r^{1/2})(\mathbf{Q}_{Ni} \otimes \mathbf{I}_{n_r})(\mathbf{I}_N \otimes \mathbf{R}_r^{1/2}) \\ &= \mathbf{Q}_{Ni} \otimes \mathbf{R}_r.\end{aligned}$$

The covariance matrix of $\mathbf{e}$ is then given as

$$\mathrm{E}[\mathbf{e}\mathbf{e}^H] = \sum_{i=1}^{M_I} \mathbf{Q}_{Ni} \otimes \mathbf{R}_r = \mathbf{Q}_N \otimes \mathbf{R}_r.$$

B. Solution of the optimization problem (3)

We solve the optimization problem by using the method introduced in [25]. First, we analyze the optimal structure of the solution by using the Lagrangian method, then find the optimal power allocation scheme, and finally determine the optimal ordering for the related eigenvector matrices.

1) Solution Structure: We begin by analyzing the structure of an optimal solution to (3). Let us define

$$\mathbf{T} = \mathbf{U}^H\mathbf{S}\mathbf{V}. \tag{9}$$

Substituting $\mathbf{S} = \mathbf{U}\mathbf{T}\mathbf{V}^H$ in (3) gives the following equivalent optimization problem:

$$\min\ \mathrm{tr}(\mathbf{T}^H\mathbf{\Lambda}^{-1}\mathbf{T} + \mathbf{\Delta}^{-1})^{-1} \quad \text{subject to } \mathrm{tr}(\mathbf{T}^H\mathbf{T}) \le P, \quad \mathbf{T} \in \mathbb{C}^{N \times n_t}. \tag{10}$$

We now show that the solution to (10) has at most one nonzero in each row and column.

Theorem 5.1: There exists a solution of (10) of the form $\mathbf{T} = \mathbf{\Pi}_1\mathbf{\Sigma}\mathbf{\Pi}_2$ where $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ are permutation matrices and $\sigma_{ij} = 0$ for all $i \neq j$.

Proof: We argue that it suffices to prove the theorem under the following nondegeneracy assumption:

$$\delta_i \neq \delta_j > 0 \text{ and } \lambda_i \neq \lambda_j > 0 \text{ for all } i \neq j. \tag{11}$$

Indeed, since the cost function of (10) is a continuous function of $\mathbf{\Delta}$ and $\mathbf{\Lambda}$, and since any $\lambda > \mathbf{0}$ and $\delta > \mathbf{0}$ can be approximated arbitrarily closely by vectors δ and λ satisfying the nondegeneracy conditions (11), we conclude that the theorem holds for arbitrary $\lambda > \mathbf{0}$ and $\delta > \mathbf{0}$.

There exists an optimal solution of (10) since the feasible set is compact and the cost function is a continuous function of $\mathbf{T}$. Since the eigenvalues of $\mathbf{\Delta}^{\frac{1}{2}}\mathbf{T}^H\mathbf{\Lambda}^{-1}\mathbf{T}\mathbf{\Delta}^{\frac{1}{2}}$ are nonnegative, the eigenvalues of $(\mathbf{\Delta}^{\frac{1}{2}}\mathbf{T}^H\mathbf{\Lambda}^{-1}\mathbf{T}\mathbf{\Delta}^{\frac{1}{2}} + \mathbf{I})^{-1}$ are less than or equal to 1. Also, by [28, Chap. 9, H.1.g], the trace of the product of two positive semi-definite Hermitian matrices is bounded by the dot product of the eigenvalues of the two matrices which are arranged in decreasing order. It follows that for any choice of $\mathbf{T}$,

$$\text{tr}(\mathbf{T}^H\mathbf{\Lambda}^{-1}\mathbf{T} + \mathbf{\Delta}^{-1})^{-1} = \text{tr}\mathbf{\Delta}(\mathbf{\Delta}^{\frac{1}{2}}\mathbf{T}^H\mathbf{\Lambda}^{-1}\mathbf{T}\mathbf{\Delta}^{\frac{1}{2}} + \mathbf{I})^{-1} \leq \text{tr}(\mathbf{\Delta}),$$

with equality when $\mathbf{T} = \mathbf{0}$. Hence, there exists a nonzero optimal solution of (10), which is denoted $\bar{\mathbf{T}}$. The first-order necessary condition for an optimal solution is the following: There exists a scalar $\gamma \geq 0$ such that

$$\frac{d}{d\mathbf{T}}\text{tr}\left((\mathbf{T}^H\mathbf{\Lambda}^{-1}\mathbf{T} + \mathbf{\Delta}^{-1})^{-1} + \gamma\mathbf{T}^H\mathbf{T}\right)_{\mathbf{T}=\bar{\mathbf{T}}} = \mathbf{0}. \tag{12}$$

Let $\mathbf{M} = \bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}} + \mathbf{\Delta}^{-1}$. Since the derivative [26] of the invertible matrix $\mathbf{M}$ is given by $\frac{d\mathbf{M}^{-1}}{dt} = -\mathbf{M}^{-1}\left(\frac{d\mathbf{M}}{dt}\right)\mathbf{M}^{-1}$ for every element t of the matrix $\mathbf{T}$, (12) is equivalent to:

$$\text{tr}\left(\gamma[\bar{\mathbf{T}}^H\delta\mathbf{T} + \delta\mathbf{T}^H\bar{\mathbf{T}}] - \mathbf{M}^{-1}[\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\delta\mathbf{T} + \delta\mathbf{T}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}}]\mathbf{M}^{-1}\right) = 0$$

for all matrices $\delta\mathbf{T} \in \mathbb{C}^{N\times n_t}$.

Let Real(z) denote the real part of $z \in \mathbb{C}$. Based on the fact that tr $(\mathbf{A} + \mathbf{A}^H) = 2(\text{Real}\ [\text{tr}\ (\mathbf{A})])$ and tr $(\mathbf{AB}) =$ tr $(\mathbf{BA})$, we have Real $\left[\text{tr}\left(\gamma\bar{\mathbf{T}}^H\delta\mathbf{T} - \mathbf{M}^{-2}\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\delta\mathbf{T}\right)\right] = 0$. By taking $\delta\mathbf{T}$ either pure real or pure imaginary, we deduce that tr $\left([\gamma\bar{\mathbf{T}}^H - \mathbf{M}^{-2}\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}]\delta\mathbf{T}\right) = 0$ for all $\delta\mathbf{T}$. By choosing $\delta\mathbf{T}$ to be completely zero except for a single nonzero entry, we conclude that

$$\gamma\bar{\mathbf{T}}^H - \mathbf{M}^{-2}\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1} = \mathbf{0}. \tag{13}$$

If $\gamma = 0$, then $\bar{\mathbf{T}} = \mathbf{0}$ since both $\mathbf{\Delta}$ and $\mathbf{\Lambda}$ are invertible. Hence, $\gamma > 0$.

We multiply (13) on the right by $\bar{\mathbf{T}}$ to obtain

$$\gamma\bar{\mathbf{T}}^H\bar{\mathbf{T}} = \mathbf{M}^{-2}\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}} = (\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}} + \mathbf{\Delta}^{-1})^{-2}\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}} \tag{14}$$

Since $\bar{\mathbf{T}}^H\bar{\mathbf{T}}$ is Hermitian, we have

$$(\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}} + \mathbf{\Delta}^{-1})^{-2}\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}} = \bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}}(\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}} + \mathbf{\Delta}^{-1})^{-2}.$$

Then we will show that $\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}}$ and $\mathbf{\Delta}^{-1}$ commute with each other. We need the following lemma [27, P. 249]:

Lemma 5.1: If $\mathbf{A}$ and $\mathbf{B}$ are diagonalizable, they share the same eigenvector matrix if and only if $\mathbf{AB} = \mathbf{BA}$.

Let $\mathbf{A} = \bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}}$ and $\mathbf{B} = \mathbf{\Delta}^{-1}$. Then we have $(\mathbf{A}+\mathbf{B})^{-2}\mathbf{A} = \mathbf{A}(\mathbf{A}+\mathbf{B})^{-2}$. According to Lemma 5.1, $\mathbf{A}$ and $(\mathbf{A}+\mathbf{B})^{-2}$ share the same eigenvector matrix. Since $\mathbf{A}+\mathbf{B}$ and $(\mathbf{A}+\mathbf{B})^{-2}$ have the same eigenvector matrix, $\mathbf{A}$ and $\mathbf{A}+\mathbf{B}$ share the same eigenvector matrix. Then we have $\mathbf{A}(\mathbf{A}+\mathbf{B}) = (\mathbf{A}+\mathbf{B})\mathbf{A}$. Hence, $\mathbf{AB} = \mathbf{BA}$, which implies that $\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}}$ and $\mathbf{\Delta}^{-1}$ commute with each other. Since $\mathbf{\Delta}^{-1}$ is diagonal, it follows from the nondegeneracy assumption that $\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}}$ is diagonal. Since $\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}}$ is diagonal, $\bar{\mathbf{T}}^H\bar{\mathbf{T}}$ is diagonal by (14).

Since $\bar{\mathbf{T}}^H\mathbf{\Lambda}^{-1}\bar{\mathbf{T}}$ and $\mathbf{\Delta}^{-1}$ are diagonal, both $\mathbf{M}$ and $\mathbf{M}^{-1}$ are diagonal. Hence, the factor $\mathbf{M}^{-2}$ in (13) is diagonal with real diagonal elements denoted e_j, $1 \le j \le n_t$. By (13), we have $\gamma\bar{t}_{ij} = \frac{e_j\bar{t}_{ij}}{\lambda_i}$. If $\bar{t}_{ij} \neq 0$, then this further implies that $\frac{e_j}{\lambda_i} = \gamma \neq 0$. By the nondegeneracy condition (11), no two diagonal elements of $\mathbf{\Lambda}$ are equal. If for any fixed j, $\bar{t}_{ij} \neq 0$ for $i = i_1$ and i_2, then the identity $\frac{e_j}{\lambda_i} = \gamma$ yields a contradiction since $\gamma \neq 0$ and $\lambda_{i_1} \neq \lambda_{i_2}$. Hence, each column of $\bar{\mathbf{T}}$ has at most one nonzero. Since $\bar{\mathbf{T}}^H\bar{\mathbf{T}}$ is diagonal, two different columns cannot have their single nonzero in the same row. This implies that each column and each row of $\mathbf{T}$ have at most one nonzero. A suitable permutation of the rows and columns of $\bar{\mathbf{T}}$ gives a diagonal matrix $\mathbf{\Sigma}$, which completes the proof. ■

Combining the relationship (9) between $\mathbf{T}$ and $\mathbf{S}$ and Theorem 5.1, we conclude that problem (3) has a solution of the form $\mathbf{S} = \mathbf{U\Pi}_1\mathbf{\Sigma\Pi}_2\mathbf{V}^H$, where $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ are permutation matrices. We will show that we can eliminate one of these two permutation matrices. Substituting $\mathbf{S} = \mathbf{U\Pi}_1\mathbf{\Sigma\Pi}_2\mathbf{V}^H$ in (3), the equivalent optimization problem is obtained as:

$$\begin{aligned} \min_{\mathbf{\Sigma},\mathbf{\Pi}_1,\mathbf{\Pi}_2} \quad & \mathrm{tr}\left(\mathbf{\Sigma}^H(\mathbf{\Pi}_1^H\mathbf{\Lambda}^{-1}\mathbf{\Pi}_1)\mathbf{\Sigma} + \mathbf{\Pi}_2\mathbf{\Delta}^{-1}\mathbf{\Pi}_2^H\right)^{-1} \\ \text{subject to} \quad & \sum_{i=1}^{M}\sigma_i^2 \le P \end{aligned} \tag{15}$$

where M represents the minimum of N and n_t. In the above optimization problem, the minimization is over diagonal matrices $\mathbf{\Sigma}$ with $\sigma_1, \ldots, \sigma_M$ as the diagonal elements, and two permutation matrices $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$. Since the symmetric permutations $\mathbf{\Pi}_1^H\mathbf{\Lambda}^{-1}\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2\mathbf{\Delta}^{-1}\mathbf{\Pi}_2^H$ essentially interchange diagonal elements of $\mathbf{\Lambda}$ and $\mathbf{\Delta}$, (15) is equivalent to

$$\min_{\sigma,\pi_1,\pi_2} \quad \sum_{i=1}^{M}\frac{1}{{\sigma_i}^2/\lambda_{\pi_1(i)} + 1/\delta_{\pi_2(i)}}$$

$$\text{subject to} \quad \sum_{i=1}^{M} \sigma_i^2 \leq P, \quad \pi_1 \in \mathcal{P}_N, \quad \pi_2 \in \mathcal{P}_{n_t} \tag{16}$$

where $\mathcal{P}_N$ is the set of bijections of $\{1, 2, \ldots, N\}$ onto itself.

We will now show that the optimal solution only depends on the smallest eigenvalues of $\mathbf{Q}_N$ and the largest eigenvalues of $\mathbf{R}_t$.

Lemma 5.2: Let $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$ and $\mathbf{V}\mathbf{\Delta}\mathbf{V}^H$ be diagonalizations of $\mathbf{Q}_N$ and $\mathbf{R}_t$ respectively where the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal eigenvectors. Let σ, π_1, and π_2 denote an optimal solution of (16) and define the sets $\mathcal{M} = \{i : \sigma_i > 0\}$, $\mathcal{Q} = \{\lambda_{\pi_1(i)} : i \in \mathcal{M}\}$, and $\mathcal{R} = \{\delta_{\pi_2(i)} : i \in \mathcal{M}\}$. If $\mathcal{M}$ has l elements, then the elements of the set $\mathcal{Q}$ constitute the l smallest eigenvalues of $\mathbf{Q}_N$, and the elements of $\mathcal{R}$ constitute the l largest eigenvalues $\mathbf{R}_t$, respectively.

Proof: Assume $k \notin \mathcal{M}$ and $\lambda_{\pi_1(k)} < \lambda_{\pi_1(i)}$ for some $i \in \mathcal{M}$. It is easy to see that by interchanging the values of $\pi_1(i)$ and $\pi_1(k)$, the new ith term in the cost function is smaller than the previous ith term. This contradicts the assumption that σ and π are optimal. Hence, $\lambda_{\pi_1(k)} \geq \lambda_{\pi_1(i)}$.

Suppose that $k \notin \mathcal{M}$ and $\delta_{\pi_2(k)} > \delta_{\pi_2(i)}$ for some $i \in \mathcal{M}$. Let C denote the cost value due to the sum of the ith term and the kth term before the interchange. Similarly, let C^+ denote the cost value due to the sum of the ith term and the kth term after the interchange of the values of $\pi_2(i)$ and $\pi_2(k)$. We have

$$C = \frac{1}{{\sigma_i}^2/\lambda_{\pi_1(i)} + 1/\delta_{\pi_2(i)}} + \delta_{\pi_2(k)}$$

and

$$C^+ = \frac{1}{{\sigma_i}^2/\lambda_{\pi_1(i)} + 1/\delta_{\pi_2(k)}} + \delta_{\pi_2(i)}$$

Since $\delta_{\pi_2(k)} > \delta_{\pi_2(i)}$, we have

$$C^+ - C = -\frac{(\delta_{\pi_2(k)} - \delta_{\pi_2(i)})({\sigma_i}^4 \delta_{\pi_2(k)}\delta_{\pi_2(i)}/\lambda^2_{\pi_1(i)} + {\sigma_i}^2\delta_{\pi_2(k)}/\lambda_{\pi_1(i)} + {\sigma_i}^2\delta_{\pi_2(i)}/\lambda_{\pi_1(i)})}{({\sigma_i}^2\delta_{\pi_2(k)}/\lambda_{\pi_1(i)} + 1)({\sigma_i}^2\delta_{\pi_2(i)}/\lambda_{\pi_1(i)} + 1)} < 0.$$

The cost is reduced by interchanging the values of $\pi_2(i)$ and $\pi_2(k)$, which violates the optimality of σ and π. Hence, $\delta_{\pi_2(k)} \leq \delta_{\pi_2(i)}$. ■

Using Lemma 5.2, we now show that one of the permutations in (16) can be deleted if the eigenvalues of $\mathbf{Q}_N$ and $\mathbf{R}_t$ are arranged in a particular order.

Theorem 5.2: Let $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$ and $\mathbf{V}\mathbf{\Delta}\mathbf{V}^H$ be diagonalizations of $\mathbf{Q}_N$ and $\mathbf{R}_t$ respectively where the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal eigenvectors, the eigenvalues of $\mathbf{Q}_N$ are arranged in an increasing order and the eigenvalues of $\mathbf{R}_t$ are arranged in a decreasing order. Then (16) is equivalent to

$$\min_{\sigma,\pi} \sum_{i=1}^{M} \frac{1}{\sigma_i^2/\lambda_{\pi(i)} + 1/\delta_i} \quad \text{subject to} \quad \sum_{i=1}^{M} \sigma_i^2 \leq P, \quad \pi \in \mathcal{P}_M, \tag{17}$$

where $\sigma_i = 0$ for $i > M$.

Proof: Since σ has at most M entries, and since the elements of $\mathcal{Q}$ are the smallest eigenvalues of $\mathbf{Q}$ and the elements of $\mathcal{R}$ are the largest eigenvalues of $\mathbf{R}_t$, we can assume that $\pi_1(i) \in [1, M]$ and $\pi_2(i) \in [1, M]$ for each $i \in \mathcal{M}$. Hence, we restrict the sum in (16) to those indices $i \in \mathcal{S} \stackrel{\triangle}{=} \{\pi_2^{-1}(j) : 1 \leq j \leq M\}$. Let us define $\sigma'_j = \sigma_{\pi_2^{-1}(j)}$ and $\pi(j) = \pi_1(\pi_2^{-1}(j))$. Since $\pi(j) \in [1, M]$ for $j \in [1, M]$, it follows that $\pi \in \mathcal{P}_M$. In (16) we restrict the summation to $i \in \mathcal{S}$ and we replace i by $\pi_2^{-1}(j)$ to obtain

$$\sum_{i \in \mathcal{S}} \frac{1}{\sigma_i^2/\lambda_{\pi_1(i)} + 1/\delta_{\pi_2(i)}} = \sum_{j=1}^{M} \frac{1}{\sigma_j'^2/\lambda_{\pi(j)} + 1/\delta_j}, \quad \text{where } \sum_{i=1}^{M} (\sigma'_j)^2 \leq P.$$

This completes the proof of (17). ■

Combining the relationship (9) between $\mathbf{T}$ and $\mathbf{S}$, Theorems 5.1 and 5.2 yields the following corollary:

Corollary 5.1: Problem (3) has a solution of the form $\mathbf{S} = \mathbf{U\Pi\Sigma V}^H$ where the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal eigenvectors of $\mathbf{Q}_N$ and $\mathbf{R}_t$ respectively with the eigenvalues of $\mathbf{Q}_N$ arranged in increasing order and the eigenvalues of $\mathbf{R}_t$ arranged in a decreasing order, $\mathbf{\Pi}$ is a permutation matrix, and $\mathbf{\Sigma}$ is diagonal.

Proof: Let σ and π be a solution of (17). For $i > M$, define $\pi(i) = i$ and $\sigma_i = 0$. If $\mathbf{\Pi}$ is the permutation matrix corresponding to π, then making a substitution $\mathbf{S} = \mathbf{U\Pi\Sigma V}^H$ in the cost function of (3) yields the cost function in (17). Since (16) and (17) are equivalent by Theorem 5.2, $\mathbf{S}$ is optimal in (3). ■

2) The Optimal $\mathbf{\Sigma}$: We now consider the optimization problem which minimizes the cost function over σ with the permutation π in (17) given. In the next subsection, we find the optimal permutation π based on the solution to the optimization problem considered here. For the sake of notational simplicity, let ρ_i denote $1/\lambda_{\pi(i)}$ and q_i denote $1/\delta_i$. Hence, for fixed π, (17) is equivalent to the following optimization problem:

$$\min_{\sigma} \sum_{i=1}^{M} \frac{1}{\rho_i \sigma_i^2 + q_i} \quad \text{subject to} \quad \sum_{i=1}^{M} \sigma_i^2 \leq P. \tag{18}$$

The solution of (18) can be expressed in terms of a Lagrange multiplier related to the power constraint. The structure of this solution has a water filling interpretation in the communication literature.

Theorem 5.3: The optimal solution of (18) is given by

$$\sigma_i = \max\left\{\sqrt{\frac{1}{\rho_i \mu}} - \frac{q_i}{\rho_i},\ 0\right\}^{1/2},$$

where the parameter μ is chosen so that $\sum_{i=1}^{M} \sigma_i^2 = P$.

Proof: Since the minimization of the cost function in (18) is over a closed and bounded set, there exists a solution. At an optimal solution to (18), the power constraint must be an equality. Otherwise,

we can multiply σ by a scalar larger than 1 to reduce to the value of the cost function. For the sake of notation simplicity, let $t_i = \sigma_i^2$. Then the reduced optimization problem (18) is equivalent to

$$\min_{\mathbf{t}} \sum_{i=1}^{M} \frac{1}{\rho_i t_i + q_i} \quad \text{subject to} \quad \sum_{i=1}^{M} t_i = P, \quad \mathbf{t} \geq \mathbf{0}. \tag{19}$$

Since the cost function is strictly convex and the constraint is convex, the optimal solution to (19) is unique.

The first-order necessary conditions (Karush-Kuhn-Tucker conditions) for an optimal solution of (19) are the following: There exists a scalar $\mu \geq 0$ and a vector $\nu \in \mathbb{R}^M$ such that

$$-\frac{\rho_i}{(\rho_i t_i + q_i)^2} + \mu - \nu_i = 0, \quad \nu_i \geq 0, \quad t_i \geq 0, \quad \nu_i t_i = 0, \quad 1 \leq i \leq M. \tag{20}$$

Due to the convexity of the cost and the constraint, any solution of these conditions is the unique optimal solution of (19).

A solution to (20) can be obtained as follows. We define the function

$$t_i(\mu) = \left(\sqrt{\frac{1}{\rho_i \mu}} - \frac{q_i}{\rho_i} \right)^+. \tag{21}$$

Here $x^+ = \max\{x, 0\}$. This particular value for t_i is obtained by setting $\nu_i = 0$ in (20) and solving for t_i; when the solution is < 0, we set $t_i(\mu) = 0$ (this corresponds to the + operator (21)). We note that $t_i(\mu)$ is a decreasing function of μ which approaches $+\infty$ as μ approaches 0 and which approaches 0 as μ grows to $+\infty$. Hence, the equation

$$\sum_{i=1}^{M} t_i(\mu) = P \tag{22}$$

has a unique positive solution. Since $t_i(\rho_i/q_i^2) = 0$, we have $t_i(\mu) = 0$ for $\mu \geq \rho_i/q_i^2$. Then we have

$$-\frac{\rho_i}{(\rho_i t_i(\mu) + q_i)^2} + \mu = -\frac{\rho_i}{q_i^2} + \mu > 0 \quad \text{for} \quad \mu > \rho_i/q_i^2.$$

We deduce that the Karush-Kuhn-Tucker conditions can be satisfied when μ is the positive solution of (22). ■

3) Optimal Eigenvector Ordering: Finally, we need to find an optimal permutation in (17), or equivalently, an optimal ordering for the eigenvalues of $\mathbf{Q}_N$ and $\mathbf{R}_t$.

Theorem 5.4: If the eigenvalues $\{\lambda_i\}$ of $\mathbf{Q}_N$ are arranged in an increasing order and the eigenvalues $\{\delta_i\}$ of $\mathbf{R}_t$ are arranged in a decreasing order, then an optimal permutation in (17) is

$$\pi(i) = i, \quad 1 \leq i \leq M. \tag{23}$$

Proof: Suppose that σ and π are optimal in (17). For convenience, let λ_i stand for λ_{π_i}. If there exist indices i and j such that $i < j$, $\sigma_i > 0$, $\sigma_j > 0$, $\lambda_i > \lambda_j$ and $\delta_i > \delta_j$, (equivalently, $\rho_i < \rho_j$ and

$q_i < q_j$), then we show that if λ_i and λ_j are interchanged in (17), the value of the objective function can be reduced.

Let us consider the following optimization problem:

$$\min_{t_i,t_j} \ \frac{1}{\rho_i t_i + q_i} + \frac{1}{\rho_j t_j + q_j} \quad \text{subject to } \ t_i + t_j = \bar{P}, \quad t_i \geq 0, \ \ t_j \geq 0, \tag{24}$$

where $\bar{P} = \sigma_i^2 + \sigma_j^2$. Since σ yields an optimal solution of (17), it follows that a solution of the above optimization problem is $t_i = \sigma_i^2$ and $t_j = \sigma_j^2$. By Theorem 5.3,

$$t_i(\mu) = \sqrt{\frac{1}{\rho_i \mu}} - \frac{q_i}{\rho_i}, \tag{25}$$

where μ is a Lagrange multiplier obtained from the power constraint $t_i + t_j = \bar{P}$:

$$\sqrt{\mu} = \frac{\frac{1}{\sqrt{\rho_i}} + \frac{1}{\sqrt{\rho_j}}}{\bar{P} + \frac{q_i}{\rho_i} + \frac{q_j}{\rho_j}}. \tag{26}$$

Let C denote the cost function for (24). Combining (25) and (26) gives

$$C = \frac{1}{\rho_i t_i + q_i} + \frac{1}{\rho_j t_j + q_j} = \frac{(\frac{1}{\sqrt{\rho_i}} + \frac{1}{\sqrt{\rho_j}})^2}{\bar{P} + \frac{q_i}{\rho_i} + \frac{q_j}{\rho_j}}.$$

Now, suppose that we interchange the values of ρ_i and ρ_j. Let C^+ denote the cost value associated with the interchange. That is, C^+ is given by

$$C^+ = \min_{t_i,t_j} \ \frac{1}{\rho_j t_i + q_i} + \frac{1}{\rho_i t_j + q_j} \quad \text{subject to } \ t_i + t_j = \bar{P}, \quad t_i \geq 0, \ \ t_j \geq 0, \tag{27}$$

Assuming the optimal solution of (24) is positive (after the exchange of ρ_i and ρ_j), we have

$$C^+ = \frac{(\frac{1}{\sqrt{\rho_i}} + \frac{1}{\sqrt{\rho_j}})^2}{\bar{P} + \frac{q_j}{\rho_i} + \frac{q_i}{\rho_j}}.$$

We need to use the following lemma [28]:

Lemma 5.3: If a_i, b_i, $i = 1, \ldots, n$ are two sets of numbers,

$$\sum_{i=1}^{n} a_{[i]} b_{[n-i+1]} \leq \sum_{i=1}^{n} a_i b_i \leq \sum_{i=1}^{n} a_{[i]} b_{[i]},$$

where $a_{[1]} \geq \ldots \geq a_{[n]}$ denote the components of a_i in a decreasing order.

By Lemma 5.3, we have $\frac{q_j}{\rho_i} + \frac{q_i}{\rho_j} > \frac{q_i}{\rho_i} + \frac{q_j}{\rho_j}$ since $\rho_i < \rho_j$ and $q_i < q_j$. This implies that $C^+ < C$.

The fact than $C^+ < C$ contradicts the optimality of σ. Hence for each i and j with $i < j$, $\rho_i < \rho_j$ and $q_i < q_j$, we can interchange the values of ρ_i and ρ_j to obtain a new permutation with the reduced value for the cost function. After the interchange, we have $\rho_i > \rho_j$ and $\lambda_i < \lambda_j$. In this way, the λ_i's are arranged in an increasing order. Since the δ_i's are arranged in a decreasing order, we conclude that the associated optimal permutation π is (23).

Now, consider the case where the optimal solution of (24) is not strictly positive. Since the original solution of (24), before the exchange, is positive, it follows from (25) and (26) that

$$\bar{P} > \frac{q_i}{\sqrt{\rho_i \rho_j}} - \frac{q_j}{\rho_j} \quad \text{and} \quad \bar{P} > \frac{q_j}{\sqrt{\rho_i \rho_j}} - \frac{q_i}{\rho_i}. \tag{28}$$

After the exchange, the analogous inequalities that must be satisfied to preserve nonnegativity are

$$\bar{P} > \frac{q_j}{\sqrt{\rho_i \rho_j}} - \frac{q_i}{\rho_j}, \tag{29}$$

and

$$\bar{P} > \frac{q_i}{\sqrt{\rho_i \rho_j}} - \frac{q_j}{\rho_i}. \tag{30}$$

Note that (30) is satisfied from (28) and the fact that $\rho_i < \rho_j$ and $q_i < q_j$. If (29) is also satisfied, the proof is completed since the solution of (24) after the exchange of λ_i and λ_j is positive.

Now, suppose that (29) is violated. In this case, we have

$$\bar{P} \leq \frac{q_j}{\sqrt{\rho_i \rho_j}} - \frac{q_i}{\rho_j}. \tag{31}$$

Combining (28) and (31), it follows that

$$\max\Big\{\frac{q_i}{\sqrt{\rho_i \rho_j}} - \frac{q_j}{\rho_j}, \frac{q_j}{\sqrt{\rho_i \rho_j}} - \frac{q_i}{\rho_i}\Big\} < \bar{P} \leq \frac{q_j}{\sqrt{\rho_i \rho_j}} - \frac{q_i}{\rho_j}. \tag{32}$$

We show that for all $\bar{P}$ satisfying (32), $C^+ \leq C$. Consequently, by exchanging λ_i and λ_j, the cost cannot increase.

Let C^* be the objective function value in (27) corresponding to $t_j = 0$ and $t_i = \bar{P}$:

$$C^* = \frac{1}{\rho_j \bar{P} + q_i} + \frac{1}{q_j}.$$

We show that $C^* \leq C$. Since $C^+ \leq C^*$, we deduce that $C^+ \leq C$.

The inequality $C^* \leq C$ is equivalent to the following:

$$\frac{1}{\rho_j \bar{P} + q_i} + \frac{1}{q_j} \leq \frac{(\frac{1}{\sqrt{\rho_i}} + \frac{1}{\sqrt{\rho_j}})^2}{\bar{P} + \frac{q_i}{\rho_i} + \frac{q_j}{\rho_j}}.$$

Multiplying both sides of the above inequality by $(\rho_j \bar{P} + q_i) q_j (\bar{P} + \frac{q_i}{\rho_i} + \frac{q_j}{\rho_j})$ gives

$$q_j(\bar{P} + \frac{q_i}{\rho_i} + \frac{q_j}{\rho_j}) + (\rho_j \bar{P} + q_i)(\bar{P} + \frac{q_i}{\rho_i} + \frac{q_j}{\rho_j}) \leq (\frac{1}{\sqrt{\rho_i}} + \frac{1}{\sqrt{\rho_j}})^2 (\rho_j \bar{P} + q_i) q_j.$$

After some rearrangement, this reduces to

$$f(\bar{P}) = \rho_j \bar{P}^2 + (q_i + q_j + \frac{\rho_j q_i}{\rho_i} - \frac{\rho_j q_j}{\rho_i} - \frac{2\sqrt{\rho_j} q_j}{\sqrt{\rho_i}}) \bar{P} + (\frac{q_i}{\sqrt{\rho_i}} - \frac{q_j}{\sqrt{\rho_j}})^2 \leq 0.$$

We now evaluate f at the possible endpoints of the interval given in (32). We have

$$f(\frac{q_i}{\sqrt{\rho_i\rho_j}}-\frac{q_j}{\rho_j})=(\frac{\sqrt{\rho_j}}{\sqrt{\rho_i}}+1)(q_i-q_j)(\frac{q_i}{\sqrt{\rho_i\rho_j}}-\frac{q_j}{\rho_j}-\frac{q_j}{\sqrt{\rho_i\rho_j}}+\frac{q_i}{\rho_i})\leq 0$$

when $\frac{q_i}{\sqrt{\rho_i\rho_j}}-\frac{q_j}{\rho_j}\geq\frac{q_j}{\sqrt{\rho_i\rho_j}}-\frac{q_i}{\rho_i}$,

$$f(\frac{q_j}{\sqrt{\rho_i\rho_j}}-\frac{q_i}{\rho_i})=\frac{q_j}{q_i}(\rho_i-\rho_j)(\frac{q_j}{\sqrt{\rho_i\rho_j}}-\frac{q_i}{\rho_i}-\frac{q_i}{\sqrt{\rho_i\rho_j}}+\frac{q_j}{\rho_j})\leq 0$$

when $\frac{q_j}{\sqrt{\rho_i\rho_j}}-\frac{q_i}{\rho_i}\geq\frac{q_i}{\sqrt{\rho_i\rho_j}}-\frac{q_j}{\rho_j}$, and

$$f(\frac{q_j}{\sqrt{\rho_i\rho_j}}-\frac{q_i}{\rho_j})=q_j(q_i-q_j)\frac{1}{\rho_i\sqrt{\rho_i}\rho_j}(\sqrt{\rho_i}+\sqrt{\rho_j})(\rho_j-\rho_i)\leq 0.$$

Since f is convex and nonpositive at the ends of the interval (32), f is nonpositive on the entire interval. This implied that $C^*\leq C$ and the proof is complete. ■

C. Proof of Lemma 4.1

Let $\mathbf{P_S}=\mathbf{S}(\mathbf{S}^H\mathbf{S})^{-1}\mathbf{S}^H$ be the projection onto $\mathcal{R}(\mathbf{S})$ and $\mathbf{P_S^\perp}=\mathbf{I}-\mathbf{P_S}$ be the projection onto $\mathcal{R}^\perp(\mathbf{S})$. First, let $(\mathbf{R}_t,\mathbf{Q}_N),(\mathbf{R}_t^{'},\mathbf{Q}_N^{'})\in\mathbb{D}$. Let $\mathbf{R}_q=\mathbf{S}\mathbf{R}_t\mathbf{S}^H+\mathbf{Q}_N$ and $\mathbf{R}_q^{'}=\mathbf{S}\mathbf{R}_t^{'}\mathbf{S}^H+\mathbf{Q}_N^{'}$. Consider $\mathbf{P_S^\perp}\mathbf{R}_q=\mathbf{P_S^\perp}\mathbf{Q}_N=\mathbf{Q}_N$, $\mathbf{P_S}\mathbf{R}_q=\mathbf{S}\mathbf{R}_t\mathbf{S}^H$, and $\mathbf{P_S^\perp}\mathbf{R}_q^{'}=\mathbf{P_S^\perp}\mathbf{Q}_N^{'}=\mathbf{Q}_N^{'}$, $\mathbf{P_S}\mathbf{R}_q^{'}=\mathbf{S}\mathbf{R}_t^{'}\mathbf{S}^H$. Since $\mathbf{S}$ is of full rank, $\mathbf{P_S}\mathbf{R}_q=\mathbf{P_S}\mathbf{R}_q^{'}$ iff $\mathbf{R}_t=\mathbf{R}_t^{'}$. Also since $\mathbf{P_S}$ and $\mathbf{P_S^\perp}$ are projections onto complementary subspaces, $\mathbf{R}_q=\mathbf{R}_q^{'}$ iff $\mathbf{P_S^\perp}\mathbf{R}_q=\mathbf{P_S^\perp}\mathbf{R}_q^{'}$ and $\mathbf{P_S}\mathbf{R}_q=\mathbf{P_S}\mathbf{R}_q^{'}$, i.e. $(\mathbf{R}_t,\mathbf{Q}_N)=(\mathbf{R}_t^{'},\mathbf{Q}_N^{'})$. Moreover, given $(\mathbf{R}_t,\mathbf{Q}_N)$, choose $\mathbf{R}_t^{'}\neq\mathbf{R}_t$ and define $\mathbf{Q}_N^{'}=\mathbf{Q}_N+\mathbf{S}\mathbf{R}_t\mathbf{S}^H-\mathbf{S}\mathbf{R}_t^{'}\mathbf{S}^H$. Since $\mathbf{S}$ is of full rank, $\mathbf{Q}_N^{'}\neq\mathbf{Q}_N$. But $\mathbf{R}_q^{'}=\mathbf{S}\mathbf{R}_t^{'}\mathbf{S}^H+\mathbf{Q}_N^{'}=\mathbf{S}\mathbf{R}_t\mathbf{S}^H+\mathbf{Q}_N=\mathbf{R}_q$.

D. Proof of Lemma 4.2

In addition to the weak norm defined just right before the statement of the lemma, we are also interested in the strong norm [22], [23] of a matrix $\mathbf{A}$: $\parallel\mathbf{A}\parallel=\max_{\mathbf{x}:\mathbf{x}^H\mathbf{x}=1}[\mathbf{x}^H\mathbf{A}^H\mathbf{A}\mathbf{x}]=\sqrt{\lambda_{max}(\mathbf{A}^H\mathbf{A})}$, where λ_{max} represents the largest eigenvalues of a matrix. If $\mathbf{A}$ is Hermitian, $\parallel\mathbf{A}\parallel=|\lambda_{max}(\mathbf{A})|$.

Note that $\mathbf{Q}_N$ can be represented by a sequence $\{q_k;k=0,1,2,\ldots\}$ with $\mathbf{Q}_N=\{q_{k,j}\}=\{q_{k-j}\}$, $q_k=q_{-k}^*$ and $\sum_{k=0}^{\infty}|q_k|<\infty$. It is shown in [29] that $\parallel\mathbf{Q}_N\parallel\leq 2\,(|q_0|+2\sum_{k=1}^{\infty}|q_k|)=2M_q<\infty$. To proceed, we need the following lemma [28]:

Lemma 5.4: For two Hermitian positive semi-definite matrices $\mathbf{G}$ and $\mathbf{H}$, $\lambda_{max}(\mathbf{G}\mathbf{H})\leq\lambda_{max}(\mathbf{G})\lambda_{max}(\mathbf{H})$.

Then, we have

$$\parallel\mathbf{P_S}\mathbf{Q}_N\parallel=\sqrt{\lambda_{max}(\mathbf{Q}_N\mathbf{P_S}\mathbf{Q}_N)}\leq\sqrt{\lambda_{max}(\mathbf{Q}_N)\lambda_{max}(\mathbf{P_S})\lambda_{max}(\mathbf{Q}_N)}=\parallel\mathbf{Q}_N\parallel. \qquad (33)$$

We now show that the difference between the two matrices goes to zero asymptotically in weak norm. Using the properties of weak norm, we have

$$\begin{aligned} |\mathbf{Q}_N - \mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}^{\perp}|_w &= |\mathbf{P}_{\mathbf{S}}\mathbf{Q}_N + \mathbf{Q}_N\mathbf{P}_{\mathbf{S}} - \mathbf{P}_{\mathbf{S}}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}|_w \\ &\leq |\mathbf{P}_{\mathbf{S}}\mathbf{Q}_N|_w + |\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}|_w + |\mathbf{P}_{\mathbf{S}}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}|_w. \end{aligned} \quad (34)$$

We need the following Lemma [22], [29]:

Lemma 5.5: Given two $n \times n$ matrices $\mathbf{G}$ and $\mathbf{H}$, then $|\mathbf{GH}|_w \leq \| \mathbf{G} \| |\mathbf{H}|_w$.

First note that $|\mathbf{P}_{\mathbf{S}}|_w = \sqrt{\text{tr}[\mathbf{S}(\mathbf{S}^H\mathbf{S})^{-1}\mathbf{S}^H]/N} = \sqrt{\text{tr}[\mathbf{I}_{n_t}]/N} = \sqrt{n_t/N}$. Then using the above lemma, we have $|\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}|_w \leq \| \mathbf{Q}_N \| |\mathbf{P}_{\mathbf{S}}|_w \leq 2M_q\sqrt{n_t/N}$. Similarly, $|\mathbf{P}_{\mathbf{S}}\mathbf{Q}_N|_w = |\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}|_w \leq 2M_q\sqrt{n_t/N}$. Combining Lemma 5.5 and (33), we have $|\mathbf{P}_{\mathbf{S}}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}|_w \leq \| \mathbf{P}_{\mathbf{S}}\mathbf{Q}_N \| \sqrt{n_t/N} \leq \| \mathbf{Q}_N \| \sqrt{n_t/N} \leq 2M_q\sqrt{n_t/N}$. Thus, from (34), we have $\lim_{N\to\infty} |\mathbf{Q}_N - \mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{Q}_N\mathbf{P}_{\mathbf{S}}^{\perp}|_w = 0$.

REFERENCES

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

[2] F. Digham, N. Mehta, A. Molisch, and J. Zhang, "Joint pilot and data loading technique for MIMO systems operating with covariance feedback," *Intern. Conf. 3G Mobile Commun. Technol.*, Oct. 2004.

[3] X. Ma, L. Yang, and G. B. Giannakis, "Optimal training for MIMO frequency-selective fading channels," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 453-466, Mar. 2005.

[4] C. Fragouli, N. Al-Dhahir, and W. Turin, "Training based channel estimation for multiple-antenna broadband transmissions," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 384-391, Mar. 2003.

[5] T. F. Wong, and B. Park, "Training sequence optimization in MIMO systems with colored interference," *IEEE Trans. Commun.*, vol. 52, pp. 1939-1947, Nov. 2004.

[6] J. H. Kotecha, and A. M. Sayeed, "Transmit signal design for optimal estimation of correlated MIMO channels," *IEEE Trans. Signal Processing*, vol. 52, pp. 546-557, Feb. 2004.

[7] X. Cai, G. B. Giannakis and M. D. Zoltowski, "Space-time spreading and block coding for correlated fading channels in the presence of interference," *IEEE Trans. Commun.*, vol. 53, pp. 515-525, Mar. 2005.

[8] S. Zhou and G. B. Giannakis, "Optimal transmitter eigen-beamforming and space-time block coding based on channel correlations," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1673–1690, July 2003.

[9] M. Biguesh and A. B. Gershman, "MIMO channel estimation: optimal training and tradeoffs between estimation techniques," *IEEE Intern. Conf. Commun.*, 2004.

[10] M. Medard, " The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. Inform. Theory*, vol. 46, no. 3, pp. 933–946, May 2000.

[11] A. M. Sayeed, "Deconstrucing multi-antenna fading channels," *IEEE Trans. Signal Processing*, vol. 50, pp. 2563–2579, Oct. 2002.

[12] D. S. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Trans. Commun.*, vol. 48, pp. 502-513, Mar. 2000.

[13] C. Chuah, D. Tse, J. Kahn and R. Valenzuela, "Capacity Scaling in MIMO Wireless Systems under Correlated Fading," *IEEE Trans. Inform. Theory*, vol. 48, pp. 637–650, Mar. 2002.

[14] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath, "Optimal designs for space-time linear precoders and decoders," *IEEE Trans. Signal Processing*, vol. 50, pp. 1051–1064, May 2002.

[15] Y. Song and S. D. Blostein, "Channel Estimation and Data Detection for MIMO Systems under Spatially and Temporally Colored Interference," *EURASIP Journal of Applied Signal Processing*, pp. 685–695, May. 2004.

[16] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1993.

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

[18] D. Palomar and J. R. Fonollosa, "Practical algorithms for a family of waterfilling solutions," *IEEE Trans. Signal Processing*, vol. 53, no. 2, pp. 686–695, Feb. 2005.

[19] N. Lu, "Tests on multiplicative covariance structures," *Ph.D. Thesis*, University of Iowa, 2002.

[20] P. J. Brown, M. G. Kenward, and E. E. Bassett, "Bayesian discrimination with longitudinal data," *Biostatistics*, vol. 2, pp. 417–432, 2001.

[21] P. Dutilleul, "The MLE algorithm for the matrix normal distribution," *Journal of Statistical Computation and Simulation*, vol. 64, pp. 105–123, 1999.

[22] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. Inform. Theory*, vol. 18, pp. 725–730, 1972.

[23] U. Grenander and G. Szego, *Toeplitz Forms and Their Applications*, Berkeley, CA: Univ. California Press, 1958.

[24] J. G. Proakis, *Digital Communications*, New York: McGraw-Hill, 2001.

[25] W. W. Hager, Y. Liu and T. F. Wong, "Optimiztion of generalized mean square error in signal processing and communication," *Linear Algebra and Its Applications*, 2006. To appear.

[26] J. R. Magnus and H. Neudecher, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Chichester, West Sussex: Wiley, 1988.

[27] G. Strang, *Linear Algebra and Its Applications*, Thomson, 2006.

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

[29] R. M. Gray, *Toeplitz and Circulant Matrices: a Review*, Revised Aug. 2002. [Online]. Available: http://www-ee.stanford.edu/ gray/toeplitz.pdf.

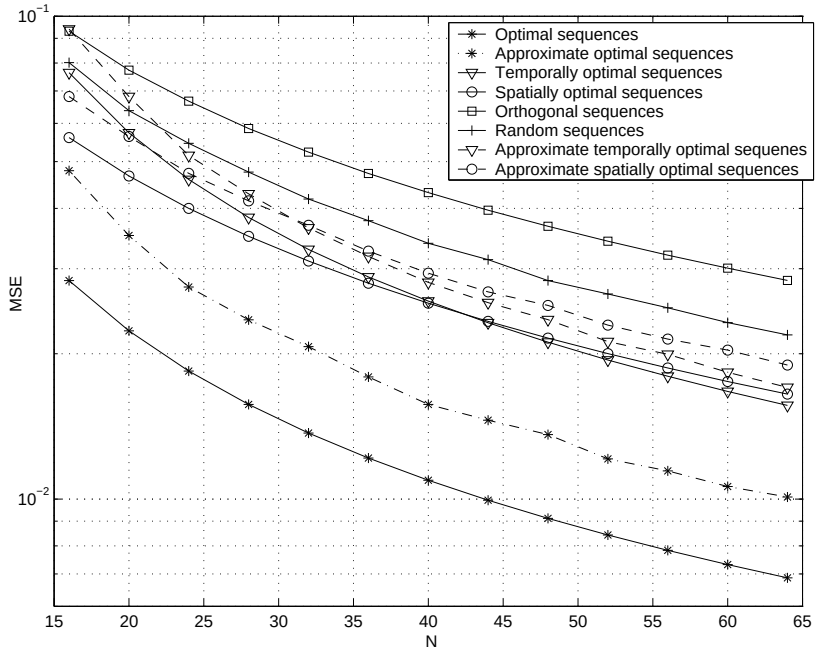

Fig. 1. Comparison of total MSEs obtained using different training sequences. ISI-free symbol waveform and high spatial correlation channel.

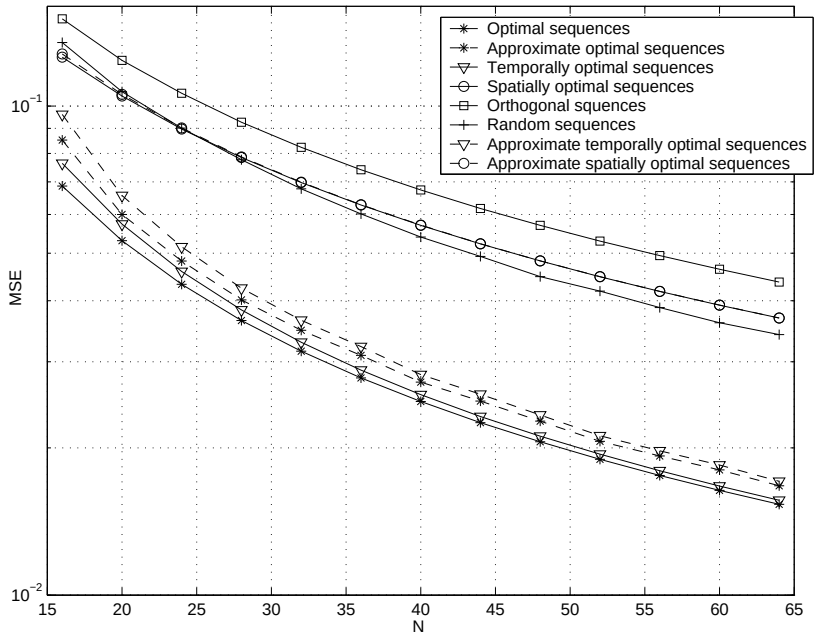

Fig. 2. Comparison of total MSEs obtained using different training sequences. ISI-free symbol waveform and low spatial correlation channel.

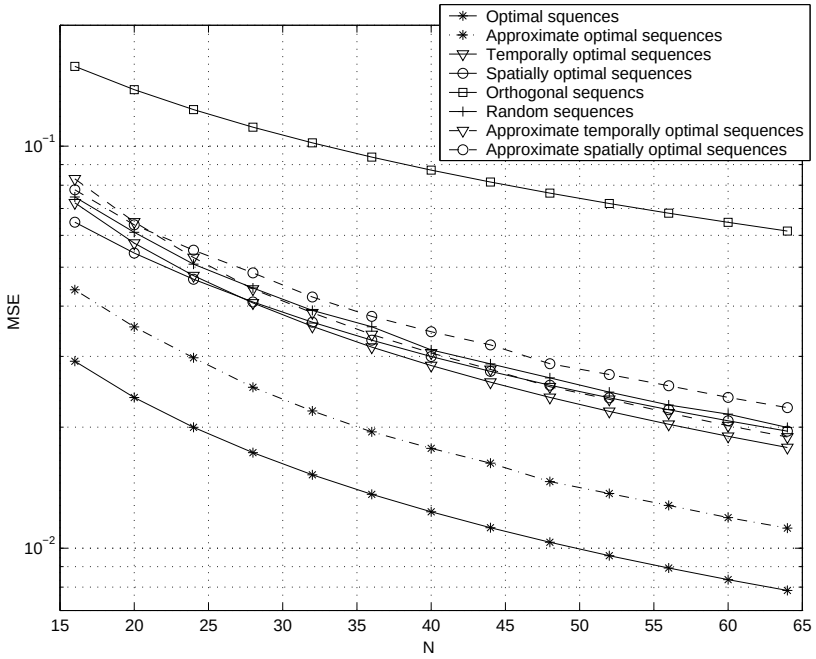

Fig. 3. Comparison of total MSEs obtained using different training sequences. AR jammers and high spatial correlation channel.

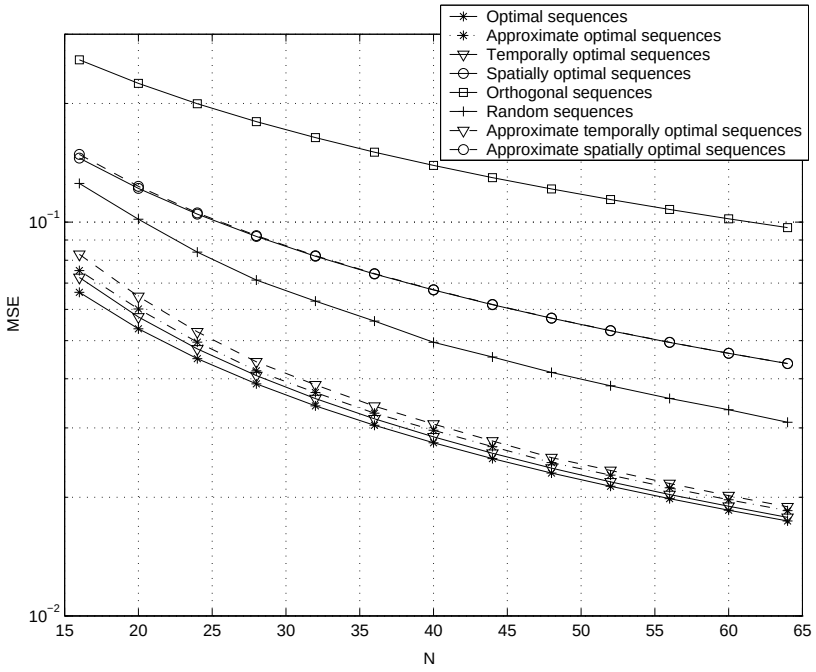

Fig. 4. Comparison of total MSEs obtained using different training sequences. AR jammers and low spatial correlation channel.