EURASIP Journal on
Advances in Signal Processing
a SpringerOpen Journal

**RESEARCH**                                                          **Open Access**

# Low-mobility channel tracking for MIMO–OFDM communication systems

Srikanth Pagadarai[1*], Alexander M Wyglinski[1] and Christopher R Anderson[2]

**Abstract**

It is now well understood that by exploiting the available additional spatial dimensions, multiple-input multiple-output (MIMO) communication systems provide capacity gains, compared to a single-input single-output systems without increasing the overall transmit power or requiring additional bandwidth. However, these large capacity gains are feasible only when the perfect knowledge of the channel is available to the receiver. Consequently, when the channel knowledge is imperfect, as is common in practical settings, the impact of the achievable capacity needs to be evaluated. In this study, we begin with a general MIMO framework at the outset and specialize it to the case of orthogonal frequency division multiplexing (OFDM) systems by decoupling channel estimation from data detection. Cyclic-prefixed OFDM systems have attracted widespread interest due to several appealing characteristics not least of which is the fact that a single-tap frequency-domain equalizer per subcarrier is sufficient due to the circulant structure of the resulting channel matrix. We consider a low-mobility wireless channel which exhibits inter-block channel variations and apply Kalman tracking when MIMO–OFDM communication is performed. Furthermore, we consider the signal transmission to contain a stream of training and information symbols followed by information symbols alone. By relying on predicted channel states when training symbols are absent, we aim to understand how the improvements in channel capacity are affected by imperfect channel knowledge. We show that the Kalman recursion procedure can be simplified by the optimal minimum mean square error training design. Using the simplified recursion, we derive capacity upper and lower bounds to evaluate the performance of the system.

**Keywords:** MIMO-OFDM, Cramér-Rao lower bound, Optimal training design, Channel estimation, Kalman filtering, Capacity bounds

## 1 Introduction

In the presence of a rich scattering environment, multiple-input multiple-output (MIMO) systems enable a linear increase in capacity with no increase in bandwidth or transmit power compared to single-input single-output (SISO) systems. However, the seminal work of [1] is based on the assumption that the channel is perfectly known to the receiver. In practical systems, the estimated channel using training sequences can be imperfect. As a result, there is potentially a mutual information loss between the input and the output of the channel. Given a power budget and a desired data rate, the time and power spent on training versus information symbols have to be judiciously

selected since there is an interesting interplay involving information throughput and the quality of the channel estimates. If a large fraction of the time and/or power is spent on training, excellent channel estimates can be obtained at the expense of poor information throughput. Conversely, expending too little time and/or power on training results in poor channel estimates that lead to error-prone information symbol transmission. Receivers that rely on channel estimates to perform information symbol decoding are termed as "mismatched" receivers [2-5]. In this article, we study this scenario involving a transmitter with no channel state information (CSI) communicating with a receiver that relies on imperfect channel estimates. A different problem which deals with transmit and receive precoder design under the assumption that CSI is available at the transmitter has been studied extensively in the published literature e.g., see [6-8].

*Correspondence: srikanthp@alum.wpi.edu
[1] Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, Worcester, MA 01609, USA
Full list of author information is available at the end of the article

The problem of channel estimation has been studied in numerous contexts. Here, we list a few relevant studies. For an exhaustive survey of the area of channel estimation using known pilot sequences, see [9]. One of the earliest works in formulating training designs to obtain channel estimates for orthogonal frequency division multiplexing (OFDM) systems was [10]. In [11], optimal training designs have been designed for single-carrier and OFDM systems by maximizing a tight lower bound on the ergodic training-based independent and identically distributed (i.i.d.) capacity. Optimal pilot symbol design and their placement in a packet were addressed for both SISO and MIMO systems in [12] by minimizing the Bayesian Cramer-Rao bound (CRB) of a semi-blind channel estimator. In [13], a general affine-precoding framework [14] was considered and it was shown that decoupling channel estimation from symbol detection and optimizing a least-squares channel estimator naturally leads to an OFDM system with information and training symbols on disjoint subcarriers. Considering the same framework, Ohno and Giannakis [15] provide a link between optimal training designs and maximizing the channel capacity lower bound similar to [11]. This work was extended in [16] to include a MIMO communications setup. Furthermore, by considering block-processing of transmitted symbols with a cyclic-prefix or zero-padding, optimal training designs are provided that maximize the channel capacity lower bound when a linear minimum mean square error (LMMSE) estimator is employed.

The impact of receiver estimation error from an information theoretic viewpoint has also extensively been studied. One of the earliest studies was conducted in [4] where the relationship between lower and upper bounds on the mutual information between transmitted and estimated Gaussian symbols is derived by modeling a time-varying frequency-selective channel as a random component with a known mean and a covariance that accounts for the channel estimation error. Specifically, it was shown that signal-to-noise ratio in the mutual information lower bound is lowered as a result of imperfect channel knowledge. In [17], the achievable data rate of a flat-fading interleaved MIMO channel is related to the LMMSE covariance matrix. In [5], the transmission of Gaussian symbols through a flat-fading channel was considered and it was demonstrated that when the Gaussianity assumption on the additive noise is rendered invalid due to channel estimation errors, scaled nearest neighbor detection is suboptimal. In [18], a lower bound on the capacity of a time-multiplexed training scheme in the presence of a flat-fading channel was studied and related to the variance of an LMMSE channel estimator. In [19], two pilot arrangement schemes were considered and the impact of the receiver estimation error was analyzed when CSI is available only at the receiver and when it is also fed back to

the transmitter. In both cases, maximum likelihood channel estimation was considered. The relationship between the symbol Bayesian CRB and the mutual information between estimated and transmitted symbols was shown in [20]. In this study, two strategies were considered. One, when the receiver obtains joint Bayesian channel and symbol estimates and two, when the receiver computes channel estimates followed by their utilization in obtaining symbol estimates. The model presented in [18] was generalized in [21] by considering a superimposed training scheme of which time multiplexed training can be termed as a special case. Based on the mutual information bounds derived, a comparison between the superimposed training and the conventional time multiplexed training is performed by optimizing over training design, number of transmit antennas, and a training/information symbol power budget. While Hassibi and Hochwald [18] provide the optimal noise covariance matrix that maximizes a tight lower bound on the mutual information between the input and the output when both the transmitter and the receiver have imperfect CSI, Ding and Blostein [22] provide the optimal signal covariance matrix and show that the uniform power allocation scheme is suboptimal.

Our most important contribution in this article is the result provided in Theorem 1. Although this result is similar to that provided in ([23], Lemma 1), the approach that we have taken, i.e., the application of the theory of complex-valued differentials to compute the Bayesian Fisher Information Matrix (FIM), is novel. Second, although the decoupling of channel and symbol estimation has been noted based on the structure of the least-squares channel estimator for SISO systems [13] as well as MIMO–OFDM systems [24], we arrive at this conclusion by maximizing the Bayesian FIM of a general affine-precoded MIMO system. Moreover, we extend the analysis conducted regarding Kalman channel tracking of SISO–OFDM systems in [25] to MIMO–OFDM systems. In the process, we extend the discussion in [18,21] by considering a "slowly" time-varying frequency-selective channel. In other words, while both [18,21] consider a block-invariant frequency-flat channel, we consider a frequency-selective channel that is correlated over successive symbol blocks.

The system model considered is described in Section 2. Based on this system model, we derive the Bayesian FIM of a general MIMO communications system that employs affine precoding at the transmitter in Section 3. We then show that in order to decouple channel estimation from data detection, an orthogonality constraint has to be met between the training and linear precoder matrices. A solution to the orthogonality constraint is the MIMO–OFDM system with frequency domain multiplexing (FDM) training symbols. We consider a MIMO channel that undergoes block-wise variations according

to a first-order autoregressive (AR) model. Moreover, in order to improve the information throughput while understanding the impact of imperfect channel estimates, we formulate a scheme where during the training phase, the OFDM symbol contains training and information symbols, whereas in the data transmission phase, only information symbols are transmitted. Consequently, we consider a scheme wherein during the training phase, channel tracking is performed by a Kalman filter followed by the estimation of information symbols during the data phase based on channel state prediction in Section 4. Using this setup, we derive the capacity upper and lower bounds in Section 5 based on a training scheme that has been derived in an MMSE minimizing sense. We then provide simulation examples in Section 6 to support the theoretical results.

## 2 System model

In our analysis, we consider a MIMO communications system consisting of $K$ transmit antennas that transmit $N$ training and information symbols over a time-varying frequency-selective block fading channel. We design super-imposed training symbols optimally such that the channel estimates from $N_t$ consecutive blocks of training symbols are utilized in the data detection of the following $N_d$ information symbol blocks. We assume that the receiver also has $K$ receive antennas without loss of generality. The maximum order of the discrete-time complex baseband wireless channels, $L$, is assumed to be known.

### 2.1 Training phase

In the training phase, training symbols and information symbols are *affinely*-precoded [14] and transmitted over $K$ antennas. A matrix formulation of this system for an arbitrary time index, $n$, is as follows. Assuming that the information symbol vector at each antenna is of size $M$, we stack the symbols transmitted across $K$ transmit antennas as shown below

$$\tilde{\mathbf{x}}_n \triangleq \text{vec}\left(\left[\tilde{\mathbf{x}}_{n,1}\ \tilde{\mathbf{x}}_{n,2}\ \dots\ \tilde{\mathbf{x}}_{n,K}\right]\right), \tag{1}$$

where the $n$th block of $M$ symbols from $k$th transmit antenna is represented as

$$\tilde{\mathbf{x}}_{n,k} \triangleq \left[\tilde{x}_{n,k}(0)\ \tilde{x}_{n,k}(1)\ \dots\ \tilde{x}_{n,k}(M-1)\right]^{\mathsf{T}}. \tag{2}$$

The affine-precoder output vector is similarly arranged as

$$\mathbf{x}_n \triangleq \text{vec}\left(\left[\mathbf{x}_{n,1}\ \mathbf{x}_{n,2}\ \dots\ \mathbf{x}_{n,K}\right]\right), \tag{3a}$$

$$\mathbf{x}_{n,k} \triangleq \left[x_{n,k}(0)\ x_{n,k}(1)\ \dots\ x_{n,k}(P-1)\right]^{\mathsf{T}}. \tag{3b}$$

Denoting the precoder matrix of size $KP \times KM$ as $\mathbf{Q}$ and the additive pilot-symbol vector of size $KP \times 1$ as $\mathbf{t}$, we can

now write the equation for the transmitted symbol vector during the training mode as follows:

$$\mathbf{x}_n = \mathbf{t} + \mathbf{Q}\tilde{\mathbf{x}}_n, \tag{4}$$

where $\mathbf{t} \triangleq \text{vec}\left(\left[\mathbf{t}_1\ \ \mathbf{t}_2\ \ \dots\ \ \mathbf{t}_K\right]\right)$. Furthermore, the matrix $\mathbf{Q}$ is such that the data stream transmitted from an antenna is precoded independently of the data-streams from the other antennas. In other words, $\mathbf{Q}$ has a block diagonal structure and hence $\mathbf{Q} \triangleq \text{diag}\left(\left[\mathbf{Q}_1\ \mathbf{Q}_2\ \dots\ \mathbf{Q}_K\right]\right)$. Despite this restriction on the structure of $\mathbf{Q}$, it is still general enough to encapsulate not only a MIMO system employing $K$ antennas but also a multi-user system, e.g., $K_U$ users utilizing $K$ antennas in total and communicating with a base-station equipped with $K$ antennas. Also, restricting the structure of $\mathbf{Q}$ to be block diagonal simplifies an orthogonality condition (cf., Theorem 2) that helps in the design of the linear precoder and the training vector.

After pre-multiplying the above vector by $\mathbf{I}_K \otimes \mathbf{C}_T$ where $\mathbf{C}_T \triangleq \left[\left[\mathbf{0}_{L \times (P-L)}\ \mathbf{I}_L\right]^{\mathsf{T}}\ \mathbf{I}_P\right]^{\mathsf{T}}$ with $\bar{P} = P + L$, the $K\bar{P} \times 1$ vector ($\bar{P} = P + L$) undergoes a digital-to-analog conversion followed by pulse-shaping to yield a continuous-time signal. Assuming perfect timing and carrier synchronization at the receiver, the signal is sampled to obtain the received symbol vector. Subsequently, the cyclic-prefix is removed by a pre-multiplication operation with $\mathbf{I}_K \otimes \mathbf{C}_R$ ($\mathbf{C}_R \triangleq \left[\mathbf{0}_{P \times L}\ \mathbf{I}_P\right]^{\mathsf{T}}$) and an ISI-free received vector of size $KP \times 1$ is available for processing:

$$\mathbf{y}_n = \mathcal{H}_n \mathbf{t} + \mathcal{H}_n \mathbf{Q}\tilde{\mathbf{x}}_n + \mathbf{z}_n, \tag{5}$$

where the channel matrix $\mathcal{H}_n$ is

$$\mathcal{H}_n \triangleq \begin{bmatrix} \mathbf{H}_{1,1}^{(n)} & \mathbf{H}_{1,2}^{(n)} & \dots & \mathbf{H}_{1,K}^{(n)} \\ \mathbf{H}_{2,1}^{(n)} & \mathbf{H}_{2,2}^{(n)} & \dots & \mathbf{H}_{2,K}^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{K,1}^{(n)} & \mathbf{H}_{K,2}^{(n)} & \dots & \mathbf{H}_{K,K}^{(n)} \end{bmatrix}. \tag{6}$$

Each matrix in the set, $\{\mathbf{H}_{i,j}^{(n)}\}$ for $1 \le i \le K,\ 1 \le j \le K$ is circulant with the first column, $\left[h_{i,j}^{(n)}[0]\ \dots\ h_{i,j}^{(n)}[L]\ \mathbf{0}\right]^{\mathsf{T}}$ and first row, $\left[h_{i,j}^{(n)}[0]\ \mathbf{0}\ h_{i,j}^{(n)}[L]\ \dots\ h_{i,j}^{(n)}[1]\right]$. We now define a channel vector $\mathbf{h}_n$ such that

$$\mathbf{h}_n \triangleq \text{vec}\left(\left[\mathbf{h}_{n,1}\ \mathbf{h}_{n,2}\ \dots\ \mathbf{h}_{n,K}\right]\right), \tag{7a}$$

$$\mathbf{h}_{n,i} \triangleq \text{vec}\left(\left[\mathbf{h}_{i,1}^{(n)}\ \mathbf{h}_{i,2}^{(n)}\ \dots\ \mathbf{h}_{i,K}^{(n)}\right]\right), \tag{7b}$$

$$\mathbf{h}_{i,k}^{(n)} \triangleq \left[h_{i,k}^{(n)}[0]\ \dots\ h_{i,k}^{(n)}[L]\right]^{\mathsf{T}}. \tag{7c}$$

By exploiting the commutativity property of discrete convolution, (5) can now be written in a different form in terms of the MIMO channel vector, $\mathbf{h}_n$ and the pilot symbol matrix, $\mathbf{T} \triangleq \left[\mathbf{T}_1\ \mathbf{T}_2\ \dots\ \mathbf{T}_K\right]$ as,

$$\mathbf{y}_n = \left(\mathbf{I}_K \otimes \mathbf{T}\right)\mathbf{h}_n + \mathcal{H}_n \mathbf{Q}\tilde{\mathbf{x}}_n + \mathbf{z}_n, \tag{8}$$

where the circulant matrices, $\{\mathbf{T}_k\}$ are constructed such that $[\,t_k[0] \ \ldots \ t_k[P-1]\,]^\mathsf{T}$ is the first column and $[\,t_k[0]\ t_k[P-L]\ldots t_k[P-1]\,]$ is the first row. In (5) and (8), we use the subscript $n$ in $\mathcal{H}_n$ and $\mathbf{h}_n$ to indicate the time-dependence of the random channel. The system model described above needs to satisfy the following conditions.

(C1) The $KP \times KM$ dimensional linear-precoder matrix, $\mathbf{Q}$ is of full column-rank and strictly-tall, i.e., $P > M$. Also, $V \triangleq P - M$.

(C2) The $P \times K(L+1)$ dimensional training matrix, $\mathbf{T}$ is a tall matrix, i.e., $P \geq K(L+1)$.

(C3) The matrix, $\mathbf{T}$ is of full column-rank, i.e., $\mathrm{rank}(\mathbf{T}) = K(L+1)$.

*Remark*: Condition (C1) is enforced as a simple way of introducing redundancy in the precoding process [7,26]. Condition (C2) ensures that enough dimensions are available for the identification of the unknown channel coefficients in a linear least-squares sense. As we shall show in Theorem 2, the extra dimensions that are available as a result of employing a full-column rank, strictly-tall precoding matrix are useful in designing the training vector. Condition (C2) also suggests that given the knowledge of the channel order and for a fixed number of transmit antennas, the data-block size has to be at least equal to the product of the number of channel taps and the number of transmit antennas. Condition (C3) which complements (C2) implies that each element of the set, $\{\mathbf{T}_k\}$ is also of full column-rank.

## 2.2 Data transmission phase

Due to the fact that no training symbols are available in the data transmission phase, we can write the system model as follows:

$$\mathbf{r}_n = \mathcal{H}_n \mathbf{s}_n + \mathbf{z}_n, \tag{9}$$

where $\mathbf{s}_n = \bar{\mathbf{Q}}\,\tilde{\mathbf{s}}_n$ and $\tilde{\mathbf{s}}_n$ are obtained in a manner similar to (1).

A few assumptions on the system model shown in (8) and (9) are now in order.

(A1) The channel vector, $\mathbf{h}_n$ is zero-mean, i.i.d complex Gaussian with variance $\sigma_h^2$, i.e., $\mathbf{h}_n \sim \mathcal{CN}(\mathbf{0}, \sigma_h^2\, \mathbf{I}_{K^2(L+1)})$. Moreover, each channel tap gain is assumed to be an independent AR process. We only consider a first-order AR model (cf. Appendix 1.1 for a brief discussion of the general AR model) for each tap gain so that

$$\mathbf{h}_n = a\mathbf{h}_{n-1} + \mathbf{u}_n, \tag{10}$$

where $a \in [\,0,\,1\,]$ is the AR coefficient for the $l$th channel tap gain and the excitation noise, $\mathbf{u}_n \sim \mathcal{CN}(\mathbf{0}, \sigma_u^2\, \mathbf{I}_{K^2(L+1)})$. In order to match the

correlation functions at lags 0 and 1 and thus make the random process WSS for $n \geq 0$, we select $\sigma_u^2 = (1-a^2)\,\sigma_h^2$.

(A2) The transmitted symbol vectors, $\tilde{\mathbf{x}}_n$ and $\tilde{\mathbf{s}}_n$ are i.i.d complex Gaussian with variance $\sigma_x^2$ and $\sigma_s^2$, i.e., $\tilde{\mathbf{x}}_n \sim \mathcal{CN}(\mathbf{0}, \sigma_x^2\, \mathbf{I}_{KP})$ and $\tilde{\mathbf{s}}_n \sim \mathcal{CN}(\mathbf{0}, \sigma_s^2\, \mathbf{I}_{KP})$ respectively.

(A3) The additive noise vector, $\mathbf{z}_n$ is zero-mean, circularly-symmetric i.i.d complex Gaussian noise with variance $\sigma_z^2$, i.e., $\mathbf{z}_n \sim \mathcal{CN}(\mathbf{0}, \sigma_z^2\, \mathbf{I}_{KP})$.

*Remark*: Assumption (A1) indicates that the channel is modeled as Rayleigh-fading random vector. This assumption represents a standard model for a rich scattering environment in the absence of line-of-sight. An expression for $a$ in terms of the channel Doppler spread and the transmission bandwidth was shown in [4]. However, the first-order AR model possibly incurs considerable estimation error and results in numerous erroneous symbol decisions [27,28]. One reason for making assumption (A2) is to satisfy the regularity conditions related to the evaluation of the Bayesian FIM described below. They require that the joint distribution of $p(\mathbf{y}_n, \tilde{\mathbf{x}}_n, \mathbf{h}_n)$ be absolutely continuous with respect to $x_{n,k}(p)$. A data vector modeled as Gaussian meets this criterion. For those transmit symbol vectors modeled on other distributions, the Theorem 1 gives an approximation. Another reason for making this assumption lies in the fact that a signal that is a zero-mean uncorrelated complex Gaussian distributed maximizes the lower bound (which is given with respect to a zero-mean uncorrelated complex Gaussian noise vector) on the mutual information between the input and the output for of MIMO channels [18,29]. Moreover, for a block transmission scheme such as an OFDM system with large number of subcarriers, the transmit symbol vector obtained by linear-precoding the information-symbol vector with an IDFT matrix can be claimed to be Gaussian by an appeal to the central limit theorem ([30], Figure 4.21). Hence, (A2) is not a particularly restrictive assumption.

## 3 Decoupled channel and symbol estimation

An observation of (8) reveals that the knowledge of the MIMO channel vector is contained not only in the known training symbols, but also in the unknown information symbols. However, the joint estimation of the channel vector and the detection of the information symbol vector is a non-linear problem, and its solution may not exist in certain cases [13]. On the other hand, a sub-optimal approach is to decouple the channel estimation problem from the data detection process. In order to do so, we may consider the channel vector to be a deterministic unknown within the classical approach to statistical estimation or as a random vector by adopting the Bayesian viewpoint. In this study, we consider the latter approach and derive the FIM

of the channel vector based on the (8). That is, we derive the Bayesian FIM concerning the estimation of the channel vector using $KP \times 1$ observations gathered from all the receive antennas at an arbitrary time instant, $n$. We then maximize the Bayesian FIM, which is equivalent to minimizing the Bayesian Cramer-Rao lower bound and obtain an orthogonality criterion. Finally, we formulate an affine precoder scheme that meets this condition.

### 3.1 Strategy: Bayesian FIM maximization

**Theorem 1.** Assuming that the likelihood function of $p(\mathbf{y}_n; \mathbf{h}_n)$ for the system model given in (8) satisfies the regularity condition, the complex FIM for estimating the MIMO channel is

$$
\begin{aligned}
\mathcal{I}(\mathbf{h}_n) = {} & \sigma_z^{-2}(\mathbf{I}_K \otimes \mathbf{T}^\mathsf{H}\mathbf{T}) - \sigma_z^{-4}\sigma_x^2 \, \mathbf{\Xi}(\mathbf{t}, \mathbf{Q}) \\
& + \sigma_x^4 \, \mathbf{\Xi}(\mathbf{Q}) + \sigma_h^{-2}\mathbf{I}_{K^2(L+1)},
\end{aligned}
\tag{11}
$$

where,

$$
\mathbf{\Xi}(\mathbf{t}, \mathbf{Q}) \triangleq \mathbb{E}_\mathbf{h}\left\{ (\mathbf{I}_K \otimes \mathbf{T})^\mathsf{H} \, \mathcal{H}_n \, \mathbf{Q} \, \mathbf{G} \, \mathbf{Q}^\mathsf{H}\mathcal{H}_n^\mathsf{H}(\mathbf{I}_K \otimes \mathbf{T}) \right\},
\tag{12a}
$$

$$
\begin{aligned}
\mathbf{\Xi}(\mathbf{Q}) \triangleq \mathbb{E}_\mathbf{h}\bigg\{ & \sum_{j=0}^{KM-1} (\mathcal{Q}_j^\mathsf{H} \otimes \mathbf{h}_n^\mathsf{T}\mathcal{Q}_j^\mathsf{T})(\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1} \otimes \mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-\mathsf{T}}) \\
& \times \sum_{j=0}^{KM-1} (\mathcal{Q}_j^* \otimes \mathcal{Q}_j\mathbf{h}_n) \bigg\},
\end{aligned}
\tag{12b}
$$

$$
\mathbf{G} \triangleq \left(\mathbf{I}_{KM} + \sigma_z^{-2}\sigma_x^2\mathbf{Q}^\mathsf{H}\mathcal{H}_n^\mathsf{H}\,\mathcal{H}_n\,\mathbf{Q}\right)^{-1},
\tag{12c}
$$

$$
\mathcal{Q}_j \triangleq \mathbf{I}_K \otimes [\, \mathcal{Q}_{j,1} \;\; \mathcal{Q}_{j,2} \; \dots \; \mathcal{Q}_{j,K} \,].
\tag{12d}
$$

*Proof.* See Appendix 1.2. □

*Remark*: Since we will be considering a non-decision-aided setup where any information about the channel coefficients that is contained in data symbols is discarded, the term $\mathbf{\Xi}(\mathbf{Q})$ represents potentially useful information that is not utilized. Consequently, as it is independent of $\mathbf{t}$, the maximization of $\mathcal{I}(\mathbf{h}_n)$ in such a scenario is possible by working with $\mathbf{\Xi}(\mathbf{t}, \mathbf{Q})$ alone. The maximization of $\mathcal{I}(\mathbf{h}_n)$ leads us to the orthogonality condition shown in Theorem 2.

**Theorem 2.** If the affine precoder scheme, $(\mathbf{t}, \mathbf{Q})$ satisfies conditions, (C1) and (C2), then the following orthogonality condition is necessary and sufficient for a non decision-aided training-only estimator to maximize the Bayesian FIM, $\mathcal{I}(\mathbf{h}_n)$ obtained in (11):

$$
\mathbf{T}_i^\mathsf{H}\mathbf{Q}_{j,m} = \mathbf{0} \quad 1 \le i, j \le K \quad 0 \le m \le M-1. \tag{13}
$$

*Proof.* See Appendix 1.3. □

The expression for the Bayesian FIM that we have obtained in Theorem 1 is analogous to the result provided in ([23], Lemma 1). Moreover, as shown in Appendix 1.2, we have not based this result on the block-diagonal structure of $\mathbf{Q}$. Hence, the result in Theorem 1 is a general one. Also, the result that we derived in Theorem 2 was showed previously in [13] for SISO systems and in [24] for MIMO OFDM systems using minimum least-squares estimation error variance arguments. In [16,23], the orthogonality condition was derived within a Bayesian framework with the former relying on an LMMSE channel estimator while the latter uses a Bayesian FIM expression similar to this study. However, while we focus on a block diagonal structure of the linear precoder, Vosoughi and Scaglione [23] focus on a general linear precoder matrix.

### 3.2 OFDM with FDM training: an orthogonal affine precoder scheme

We see from ([13], Theorem 1) for the case of a SISO system that the affine precoder scheme which uses linearly precoded OFDM along with an FDM training sequence that modulates a disjoint set of tones not used for data transmission meets the orthogonality criterion. Similarly, for MIMO systems, Theorem 3 establishes that although the training symbols and information-bearing symbols overlap in time domain, orthogonality between the subcarriers in frequency domain satisfies (13).

**Theorem 3.** The affine precoder scheme, $(\mathbf{t}, \mathbf{Q})$ that satisfies the orthogonality condition given in (13) irrespective of the FIR channel provides a non-data-aided channel estimator if it is selected from the class

$$
\mathbf{Q}_k = \mathcal{W}^\mathsf{H}\mathbf{P}^{(\mathbf{Q})}\begin{bmatrix} \Theta_{M \times M} \\ \mathbf{0}_{V \times M} \end{bmatrix},
\tag{14a}
$$

$$
\mathbf{t}_k = \mathcal{W}^\mathsf{H}\mathbf{P}^{(\mathbf{t})}\begin{bmatrix} \tilde{\mathbf{t}}_k \\ \mathbf{0} \end{bmatrix}.
\tag{14b}
$$

In the above equations, $\Theta_{M \times M}$ is any full-rank matrix and $\mathbf{P}^{(\mathbf{t})}$ is a permutation matrix that places the $L + 1$ possible non-zero entries of $\tilde{\mathbf{t}}_k$ on non data-bearing subcarriers, whereas $\mathcal{W}_{m,n} = \frac{1}{P}exp(j2\pi mn/P)$.

*Proof.* The proof is a straight-forward generalization of ([13], Appendix I). □

In the subsequent sections, we focus our analysis on a MIMO-OFDM communication system. That is, we assume $\tilde{\mathbf{x}}_{n,\bar{k}}$ to be the result of a linear-precoding operation involving a general full-rank matrix, $\Theta_{M \times M}$ before it is IDFT-modulated. Moreover, the same set of subcarriers are used for transmitting training symbols across all the antennas.

### 3.2.1 Training phase

By substituting the result of Theorem 3 in (5), and considering the signal at an arbitrary receive antenna, $k$, we write the following equation:

$$
\begin{aligned}
\mathbf{y}_{n,k} &= \sum_{\bar{k}=1}^{K} \mathbf{H}_{k,\bar{k}}^{(n)} \mathbf{t}_{\bar{k}} + \sum_{\bar{k}=1}^{K} \mathbf{H}_{k,\bar{k}}^{(n)} \mathbf{Q}_{\bar{k}} \tilde{\mathbf{x}}_{n,\bar{k}} + \mathbf{z}_{n,k} \\
&= \sum_{\bar{k}=1}^{K} \mathbf{H}_{k,\bar{k}}^{(n)} \mathcal{W}^{\mathsf{H}} \mathbf{P}^{(\mathbf{t})} \begin{bmatrix} \tilde{\mathbf{t}}_{\bar{k}} \\ \mathbf{0} \end{bmatrix} \\
&\quad + \sum_{\bar{k}=1}^{K} \mathbf{H}_{k,\bar{k}}^{(n)} \mathcal{W}^{\mathsf{H}} \mathbf{P}^{(\mathbf{Q})} \begin{bmatrix} \mathbf{I}_{M \times M} \\ \mathbf{0}_{V \times M} \end{bmatrix} \tilde{\mathbf{x}}_{n,\bar{k}} + \mathbf{z}_{n,k}. \quad (15)
\end{aligned}
$$

By multiplying the above equation with $\mathbf{P}_{0:V-1}^{(\mathbf{t})\mathsf{T}} \mathcal{W}$, we notice that the channel estimation is decoupled from data detection so that the following expression is obtained

$$
\begin{aligned}
\mathbf{P}_{0:V-1}^{(\mathbf{t})\mathsf{T}} \mathcal{W} \mathbf{y}_{n,k} &= \sum_{\bar{k}=1}^{K} \mathbf{P}_{0:V-1}^{(\mathbf{t})\mathsf{T}} \bar{\mathbf{H}}_{k,\bar{k}}^{(n)} \mathbf{P}_{0:V-1}^{(\mathbf{t})} \tilde{\mathbf{t}}_{\bar{k}} \\
&\quad + \mathbf{P}_{0:V-1}^{(\mathbf{t})\mathsf{T}} \mathcal{W} \mathbf{z}_{n,k}, \quad (16)
\end{aligned}
$$

where $\bar{\mathbf{H}}_{k,\bar{k}}^{(n)} \triangleq \mathcal{W} \mathbf{H}_{k,\bar{k}}^{(n)} \mathcal{W}^{\mathsf{H}}$ is a diagonal matrix and $\mathbf{P}_{0:V-1}^{(\mathbf{t})}$ is the result of disregarding the zero entries in $[\tilde{\mathbf{t}}_{\bar{k}}^{\mathsf{H}} \quad \mathbf{0}^{\mathsf{H}}]^{\mathsf{H}}$. Moreover, we have utilized the fact that $\mathbf{P}_{0:V-1}^{(\mathbf{t})\mathsf{T}} \mathbf{P}_{0:M}^{(\mathbf{Q})} = \mathbf{0}$ i.e., the submatrices, $\mathbf{P}_{0:V-1}^{(\mathbf{t})\mathsf{T}}$ and $\mathbf{P}_{0:M}^{(\mathbf{Q})}$ are orthogonal to each other. We also recognize that the following relationship holds due to the diagonal nature of $\bar{\mathbf{H}}_{k,\bar{k}}^{(n)}$:

$$
\begin{aligned}
\mathbf{P}_{0:V-1}^{(\mathbf{t})\mathsf{T}} \bar{\mathbf{H}}_{k,\bar{k}}^{(n)} \mathbf{P}_{0:V-1}^{(\mathbf{t})} \tilde{\mathbf{t}}_{\bar{k}} &= \sqrt{P} \tilde{\mathbf{T}}_{\bar{k}} \mathbf{P}_{0:V-1}^{(\mathbf{t})\mathsf{T}} \mathcal{W}_{0:L} \mathbf{h}_{k,\bar{k}}^{(n)} \\
&= \tilde{\mathbf{T}}_{\bar{k}} \hat{\mathcal{W}}_{0:L} \mathbf{h}_{k,\bar{k}}^{(n)}, \quad (17)
\end{aligned}
$$

where $\tilde{\mathbf{T}}_{\bar{k}} \triangleq \operatorname{diag}(\tilde{\mathbf{t}}_{\bar{k}})$ and $\hat{\mathcal{W}}_{0:L} \triangleq \sqrt{P} \mathbf{P}_{0:V-1}^{(\mathbf{t})\mathsf{T}} \mathcal{W}_{0:L}$. As a result, (16) can be written as

$$
\tilde{\mathbf{y}}_{n,k}^{(t)} = \tilde{\mathbf{T}} \mathbf{h}_{n,k} + \tilde{\mathbf{z}}_{n,k}, \quad (18)
$$

where $\tilde{\mathbf{y}}_{n,k}^{(t)} \triangleq \hat{\mathcal{W}}_{0:L} \mathbf{y}_{n,k}$ and

$$
\tilde{\mathbf{T}} \triangleq [\tilde{\mathbf{T}}_1 \hat{\mathcal{W}}_{0:L} \quad \tilde{\mathbf{T}}_2 \hat{\mathcal{W}}_{0:L} \quad \dots \quad \tilde{\mathbf{T}}_K \hat{\mathcal{W}}_{0:L}]. \quad (19)
$$

Also, it can be showed that $\mathbf{T}_k^{\mathsf{H}} \mathbf{T}_k = \hat{\mathcal{W}}_{0:L}^{\mathsf{H}} \tilde{\mathbf{T}}_k^* \tilde{\mathbf{T}}_k \hat{\mathcal{W}}_{0:L}$. With, $\tilde{\mathbf{y}}_n^{(t)} \triangleq \operatorname{vec}([\tilde{\mathbf{y}}_{n,1}^{(t)} \quad \tilde{\mathbf{y}}_{n,2}^{(t)} \quad \dots \quad \tilde{\mathbf{y}}_{n,K}^{(t)}])$, we can write the MIMO system model for the measured signal across all receive antennas due to the pilot tones as follows:

$$
\tilde{\mathbf{y}}_n^{(t)} = (\mathbf{I}_K \otimes \tilde{\mathbf{T}}) \mathbf{h}_n + \tilde{\mathbf{z}}_n. \quad (20)
$$

It can be showed that enforcing conditions (C1), (C2), and (C3) naturally result in two more standard conditions

regarding the structure of $\tilde{\mathbf{T}}$ and thus satisfy the dimensionality of (20).

(C4) The $V \times K(L+1)$ dimensional training matrix, $\tilde{\mathbf{T}}$ is a tall matrix, i.e., $V \geq K(L+1)$.

(C5) The matrix, $\tilde{\mathbf{T}}$ is of full column-rank, i.e., $\operatorname{rank}(\tilde{\mathbf{T}}) = K(L+1)$.

By employing operations similar to those that helped in obtaining (20), the equation for the observation vector affected by the information symbols alone is as follows:

$$
\tilde{\mathbf{y}}_n^{(dt)} = \widetilde{\mathcal{H}}_n \tilde{\mathbf{x}}_n + \tilde{\mathbf{z}}_n, \quad (21)
$$

where the $KM \times KM$ channel matrix, $\widetilde{\mathcal{H}}_n$ is as follows:

$$
\widetilde{\mathcal{H}}_n \triangleq \begin{bmatrix} \tilde{\mathbf{H}}_{1,1}^{(n)} & \tilde{\mathbf{H}}_{1,2}^{(n)} & \dots & \tilde{\mathbf{H}}_{1,K}^{(n)} \\ \tilde{\mathbf{H}}_{2,1}^{(n)} & \tilde{\mathbf{H}}_{2,2}^{(n)} & \dots & \tilde{\mathbf{H}}_{2,K}^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{H}}_{K,1}^{(n)} & \tilde{\mathbf{H}}_{K,2}^{(n)} & \dots & \tilde{\mathbf{H}}_{K,K}^{(n)} \end{bmatrix}. \quad (22)
$$

Each matrix in the set, $\{\tilde{\mathbf{H}}_{i,j}^{(n)}\}$ for $1 \leq i \leq K$, $1 \leq j \leq K$ is a diagonal matrix and is obtained by performing the operation,

$$
\tilde{\mathbf{H}}_{i,k}^{(n)} = \mathbf{P}_{0:M-1}^{(\mathbf{Q})\mathsf{T}} \bar{\mathbf{H}}_{i,k}^{(n)} \mathbf{P}_{0:M-1}^{(\mathbf{Q})}. \quad (23)
$$

### 3.2.2 Data transmission phase

Although the linear precoder matrix, $\bar{\mathbf{Q}}$ can be any full-column rank matrix in general, we focus on a block diagonal structure. We consider each element in the set, $\{\bar{\mathbf{Q}}_k\}$ to be a $P \times P$ IDFT matrix that modulates an information symbol vector which has been linearly precoded by a general full-rank matrix, $\bar{\Theta}_{P \times P}$.

$$
\tilde{\mathbf{r}}_n = \overline{\mathcal{H}}_n \tilde{\mathbf{s}}_n + \tilde{\mathbf{z}}_n, \quad (24)
$$

where $\tilde{\mathbf{r}}_n \triangleq \mathcal{W} \mathbf{r}_n$ and the $KP \times KP$ channel matrix, $\overline{\mathcal{H}}_n$ is defined as follows:

$$
\overline{\mathcal{H}}_n \triangleq \begin{bmatrix} \bar{\mathbf{H}}_{1,1}^{(n)} & \bar{\mathbf{H}}_{1,2}^{(n)} & \dots & \bar{\mathbf{H}}_{1,K}^{(n)} \\ \bar{\mathbf{H}}_{2,1}^{(n)} & \bar{\mathbf{H}}_{2,2}^{(n)} & \dots & \bar{\mathbf{H}}_{2,K}^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\mathbf{H}}_{K,1}^{(n)} & \bar{\mathbf{H}}_{K,2}^{(n)} & \dots & \bar{\mathbf{H}}_{K,K}^{(n)} \end{bmatrix}. \quad (25)
$$

*Remark*: By enforcing the orthogonality condition and by choosing MIMO–OFDM with FDM training symbols as the affine precoder scheme, we have broken down (8) into (20) and (21). As a result, the impact of overlapping data-bearing symbols on the channel estimator has been

circumvented. Moreover, we carryover the linear precoder from the training phase to the data transmission phase by introducing a simple modification on the dimensions of the IDFT matrix.

Before we study the MMSE characteristics during training and data transmission phases when a Kalman filter is employed to track the time-varying channel vector, $\mathbf{h}_n$ we note that the following time and power budget constraints are enforced over (20), (21), and (24),

$$N = N_t + N_d, \tag{26a}$$

$$\mathcal{P} N = (\mathcal{P}_t + \mathcal{P}_{dt}) N_t + \mathcal{P}_d N_d, \tag{26b}$$

where $\mathcal{P}$ is the total average transmit power that is split into $\mathcal{P}_t$, the average power allocated for training, $\mathcal{P}_{dt}$, the average power allocated for information symbols in the training phase, and $\mathcal{P}_d$, the average power allocated for information symbols in the data transmission phase. In addition, $\mathcal{P}_t$ is distributed equally among the transmit antennas, i.e.,

$$\mathcal{P}_t = \|\mathbf{t}\|^2 = \sum_{k=1}^{K} \|\mathbf{t}_k\|^2, \tag{27}$$

where $\|\mathbf{t}_k\|^2 = \sum_{v=0}^{V-1} |\tilde{\mathbf{T}}_k[v]|^2 = \mathcal{P}_t/K, \quad \forall 1 \le k \le K$.

## 4 Blockwise Kalman tracking

Due to the AR(1) random process model for time-variations on the channel vector, in order to compute the channel estimator in the MMSE sense, we have to utilize the past and the current observations, $\{\tilde{\mathbf{y}}_{nN+k}^{(t)} : k \in [0, N_t - 1], n = 1, 2, \ldots\}$. An MMSE channel estimator can then be given as,

$$\hat{\mathbf{h}}_{nN+k} = \mathbb{E}\{\mathbf{h}_{nN+k} \,|\, \{\tilde{\mathbf{y}}_{nN+j}^{(t)} : j \le k, j \in [0, N_t - 1]\},$$
$$\{\tilde{\mathbf{y}}_{(n-m)N+j}^{(t)} : j \in [0, N_t - 1], m = 1, 2, \ldots\}\}. \tag{28}$$

However, a batch processing approach would necessitate the use of large datasets. A natural choice is the sequential MMSE approach and is implemented by a Kalman filter. A Kalman filter is well known for its computational efficiency which results from the fact that only the most recent estimate need to be stored in order to refine the MMSE estimate of the unknown parameter of interest based on the new observations. For the current problem at hand, we compute the channel estimate during the training phase based on (20) and utilize the predicted channel vector in the data transmission phase. The Kalman filter recursion algorithm for estimating the MIMO channel vector in the setup considered is summarized in (29a)–(29e) [31].

**Prediction:**

$$\hat{\mathbf{h}}_{n|n-1} = a\hat{\mathbf{h}}_{n-1|n-1}, \tag{29a}$$

**Minimum Prediction MSE Matrix:**

$$\mathbf{M}_{n|n-1} = a^2 \mathbf{M}_{n-1|n-1} + (1 - a^2) \sigma_h^2 \mathbf{I}_{K^2(L+1)}, \tag{29b}$$

**Kalman Gain Matrix:**

$$\mathbf{K}_n = \mathbf{M}_{n|n-1} (\mathbf{I}_K \otimes \tilde{\mathbf{T}}^{\mathsf{H}})$$
$$\times (\sigma_z^2 \mathbf{I}_{KP} + (\mathbf{I}_K \otimes \tilde{\mathbf{T}}) \mathbf{M}_{n|n-1} (\mathbf{I}_K \otimes \tilde{\mathbf{T}}^{\mathsf{H}}))^{-1}, \tag{29c}$$

**Correction:**

$$\hat{\mathbf{h}}_{n|n} = \hat{\mathbf{h}}_{n|n-1} + \mathbf{K}_n (\tilde{\mathbf{y}}_n^{(t)} - (\mathbf{I}_K \otimes \tilde{\mathbf{T}}) \hat{\mathbf{h}}_{n|n-1}), \tag{29d}$$

**Minimum MSE Matrix:**

$$\mathbf{M}_{n|n} = (\mathbf{I} - \mathbf{K}_n (\mathbf{I}_K \otimes \tilde{\mathbf{T}})) \mathbf{M}_{n|n-1}. \tag{29e}$$

### 4.1 Kalman filter recursion

It can be noticed that when the system converges to a steady state, the MMSE of the channel estimator is not stationary during each cycle of $N$ blocks. In the data transmission phase, the MMSE associated with the channel estimator's predicted state increases monotonically from the $N_t$th block to the $(N - 1)$th block. Thus, the maximum steady-state MMSE in the data transmission phase occurs at the last information symbol block of each cycle. On the other hand, since the channel estimator computed based on the observations of the 0th block in $n$th cycle refines the predicted channel state at the end of the last information symbol block of $(n - 1)$th cycle, the steady-state MMSE decreases monotonically from the 0th training block to the $(N_t - 1)$th training block. Before we derive the steady-state MMSE expressions for the two cases described above, we derive the steady-state MMSE when all the blocks are training symbols and make an interesting observation. The steady-state MMSE when all blocks are training symbols is given by the solution to the Ricatti equation (based on (29e)),

$$\mathbf{M}^{(\infty)} = (\mathbf{I} - \mathbf{K}^{(\infty)} (\mathbf{I}_K \otimes \tilde{\mathbf{T}})) \mathbf{M}_1^{(\infty)}, \tag{30}$$

where $\mathbf{M}^{(\infty)} \triangleq \lim_{n \to \infty} \mathbf{M}_{n|n}$, $\mathbf{M}_1^{(\infty)} \triangleq \lim_{n \to \infty} \mathbf{M}_{n|n-1}$, and $\mathbf{K}^{(\infty)} \triangleq \lim_{n \to \infty} \mathbf{K}_n$. Although several techniques have been proposed in the published literature such as eigenvector solutions [32], Schur vector approaches [33], iterative solving for scalar polynomials [34], etc., to solve the system of equations obtained in (30), we will show that by utilizing the following lemma describing the optimal design of the training symbols in the MMSE sense, the above system of equations is greatly simplified for MIMO-OFDM systems.

**Lemma 1.** For the system model shown in (20), the minimum error variance of the MMSE channel estimator is,

$$\sigma^2_{\Delta \mathbf{h}_n} = \frac{K^3 (L+1) \sigma_z^2 \sigma_h^2}{K \sigma_z^2 + \sigma_h^2 \mathcal{P}_t}. \tag{31}$$

The optimal $\tilde{\mathbf{T}}$, $\tilde{\mathbf{T}}^{(opt)}$ that attains this error variance is,

$$\tilde{\mathbf{T}}^{(opt)} = [\ (\mathcal{E}_1 \boldsymbol{\phi}(V))^\mathsf{T} \ ... \ (\mathcal{E}_K \boldsymbol{\phi}(V))^\mathsf{T} ]^\mathsf{T}, \tag{32}$$

where

$$[\ \mathcal{E}_k\ ]_{v,v} = \exp\left[\frac{j2\pi v f_k}{V}\right], \tag{33a}$$

$$\forall \quad 0 \le v \le V-1, \ \ 1 \le k \le K$$

$$\boldsymbol{\phi}(V) = \sqrt{\frac{\mathcal{P}_t}{KV}}\ [\ \exp(j\phi_0) \quad \exp(j\phi_1) \quad ... \quad \exp(j\phi_{V-1})]^\mathsf{T}, \tag{33b}$$

$$f_k = k(V - L - 1), \tag{33c}$$

$$\{\phi_v\} \in [-\pi, \ \pi]. \tag{34}$$

*Proof.* See Appendix 1.4. □

*Remark*: By employing the training design described in (32), $\tilde{\mathbf{T}}^\mathsf{H} \tilde{\mathbf{T}}$ in (D.7) is diagonal and the MMSE of (31) is attained. The time-domain training sequences can be obtained from (14b) in a straight-forward manner by using the relation, $\mathbf{t} = (\mathbf{I}_K \otimes \mathcal{W}^\mathsf{H} \mathbf{P}^{(\mathbf{t})}_{0:V-1})\ \tilde{\mathbf{T}}$. It can be noticed that a simple way of making the term, $\tilde{\mathbf{T}}^*_{k_1}[\ v,\ v]\ \tilde{\mathbf{T}}_{k_2}[\ v,\ v]$ in (D.13) equal to zero is to allow only $(L+1)$ out of $V$ subcarriers dedicated for training symbols to be used at any given antenna. These equispaced and equipowered training symbols occupy disjoint sets of subcarriers at each transmit antenna. Clearly, such a scheme utilizes only $(L+1)$ out of $V$ subcarriers dedicated for training symbols at any given antenna. On the other hand, a general training scheme design described in (32) uses all non-data-bearing subcarriers, i.e., $V$ for channel estimation purposes.

*Remark*: In [16], disjoint sets of subcarriers were considered to reduce the MMSE channel estimation error. Training designs similar to ours were shown in [24] by minimizing the least-squares channel estimation error and in [20] by minimizing the MSE of the LMMSE channel estimate. In [35], several classes of training schemes are derived by minimizing the least-squares channel estimation error. In this study, the disjoint allocation of subcarriers for training symbols from different antennas is referred to as a *FDM* scheme and the phase-shift orthogonal design as a *code-division multiplexing in the frequency domain* scheme.

If we were to initialize the Kalman recursion by substituting the scaled identity covariance matrix of $\mathbf{h}_n$ for

$\mathbf{M}_{-1|-1}$, the one-step prediction error matrix, $\mathbf{M}_{n|n-1}$ is always a scaled identity matrix. Consequently, the matrix $\tilde{\mathbf{K}}_n \triangleq \mathbf{K}_n\ (\mathbf{I}_K \otimes \tilde{\mathbf{T}})$ is also a scaled identity matrix since $(\mathbf{I}_K \otimes \tilde{\mathbf{T}}^\mathsf{H} \tilde{\mathbf{T}})$ is designed to be a scaled identity matrix. This is better understood by writing the alternative version of the Kalman gain matrix using the matrix inversion lemma[a]:

$$\begin{aligned} \mathbf{K}_n = &\ \sigma_z^{-2}\ \mathbf{M}_{n|n-1}(\mathbf{I}_K \otimes \tilde{\mathbf{T}}^\mathsf{H}) \\ &- \sigma_z^{-4}\ \mathbf{M}_{n|n-1}\ (\mathbf{I}_K \otimes \tilde{\mathbf{T}}^\mathsf{H} \tilde{\mathbf{T}}) \\ &\times (\mathbf{M}^{-1}_{n|n-1} + (\mathbf{I}_K \otimes \tilde{\mathbf{T}}^\mathsf{H} \tilde{\mathbf{T}}))^{-1}(\mathbf{I}_K \otimes \tilde{\mathbf{T}}^\mathsf{H}). \end{aligned} \tag{35}$$

As an extension of the above fact, due to assumption, (A1) and the optimal training design described by Lemma 1, $\mathbf{M}^{(\infty)}$ is also a scaled identity matrix. It can be showed that an arbitrary diagonal element, $\mathbf{m}^{(\infty)} \triangleq \mathbf{M}^{(\infty)}[\ l, l], 0 \le l \le K^2(L+1) - 1$ is given as follows:

$$\begin{aligned} \mathbf{m}^{(\infty)} &= \frac{\sigma_z^2 (a^2\ \mathbf{m}^{(\infty)} + \sigma_u^2)}{\sigma_z^2 + \frac{\mathcal{P}_t}{K}\ (a^2\ \mathbf{m}^{(\infty)} + \sigma_u^2)} \\ &= \frac{\sigma_h^2}{\frac{1}{2}\left(1 + \frac{\sigma_h^2 \mathcal{P}_t}{\sigma_z^2 K}\right) + \sqrt{\frac{1}{4}\left(1 + \frac{\sigma_h^2 \mathcal{P}_t}{\sigma_z^2 K}\right)^2 + \frac{a^2}{1-a^2} \frac{\sigma_h^2 \mathcal{P}_t}{\sigma_z^2 K}}}. \end{aligned} \tag{36}$$

This steady-state Ricatti solution is the lower bound on the MMSE for estimating any of the $K^2(L+1)$ channel filter taps, irrespective of the particular phase being considered. To compute the steady-state MMSE characteristics, we let $n \to \infty$, and define,

$$\mathbf{M}^{(\infty)}_j \triangleq \lim_{n \to \infty} \mathbf{M}_{nN+j|nN+j}, \tag{37}$$

for $j \in [\ 0, N-1]$. We can now review the closed-form expressions for steady-state channel MMSEs in training and data transmission phases based on [36].

**Lemma 2.** When the training vectors are designed according to (32) and a Kalman filter is employed to perform channel tracking, the steady-state channel MMSEs for the system model corresponding to (20), (21), and (24) are given as follows:

$$\mathbf{m}^{(\infty)}_{N-1} = \delta^{(\infty)}_{N-1}[\ l, l] + \mathbf{m}^{(\infty)}, \tag{38a}$$

**Training phase** ($j \in [\ 0, N_t - 1]$) :

$$\mathbf{m}^{(\infty)}_j = \mathbf{m}^{(\infty)} + \frac{(1-\alpha)\ \delta^{(\infty)}_{N-1}[\ l, l]}{\beta\ (1-\alpha^j)\ \delta^{(\infty)}_{N-1}[\ l, l] + \alpha^j\ (1-\alpha)}, \tag{38b}$$

**Data transmission phase** ($j \in [\ N_t, N-1]$) :

$$\mathbf{m}^{(\infty)}_j = \frac{\mathbf{m}^{(\infty)}_{N-1} - \sigma_h^2\ (1 - a^{2(N-j)})}{a^{2(N-j)}}, \tag{38c}$$

where $\mathbf{m}_j^{(\infty)} \triangleq \mathbf{M}_j^{(\infty)}[l,l]$ and $\delta_{N-1}^{(\infty)}[l,l]$ is computed as follows:

$$\delta_{N-1}^{(\infty)}[l,l] = -b_\infty + \sqrt{b_\infty^2 + c_\infty}, \tag{39a}$$

$$b_\infty \triangleq \left( \frac{\alpha^{N_t} - a^{2(N-N_t)}}{\alpha^{N-t} - 1} \right) \left( \frac{\alpha - 1}{2\beta} \right) - \left( \frac{1 - a^{2(N-N_t)}}{2} \right) (\sigma_h^2 - \mathbf{m}^{(\infty)}), \tag{39b}$$

$$c_\infty \triangleq \alpha^{N_t} \left( \frac{1 - a^{2(N-N_t)}}{\alpha^{N_t} - 1} \right) \left( \frac{\alpha - 1}{\beta} \right) (\sigma_h^2 - \mathbf{m}^{(\infty)}), \tag{39c}$$

$$\alpha \triangleq \frac{1}{a^2} \left( 1 + (\sigma_h^2 - a^2(\sigma_h^2 - \mathbf{m}^{(\infty)})) \frac{\mathcal{P}_t}{K} \right)^2, \tag{39d}$$

$$\beta \triangleq \frac{\mathcal{P}_t}{K} \left( 1 + (\sigma_h^2 - a^2(\sigma_h^2 - \mathbf{m}^{(\infty)})) \frac{\mathcal{P}_t}{K} \right), \tag{39e}$$

*Proof.* See proof of Lemma 1 in [36]. □

# 5 Capacity bounds with sequential MMSE channel estimation

Similar to [18], we adopt the definition of capacity in bits per channel use to be the maximum over the distribution of the transmit signal of the mutual information between the known training symbols and the observations and the unknown transmitted signal. In other words, for the system model shown in (20), (21), and (24), the channel capacity averaged over the random channel is defined as follows:

$$C = \frac{1}{N} \times \frac{M}{P} \sum_{n=0}^{N_t-1} \mathbb{E} \left[ \max_{p_x(.), \mathbb{E}[||\tilde{\mathbf{x}}_n||^2] = \mathcal{P}_{dt}} I(\tilde{\mathbf{y}}_n^{(dt)}; \tilde{\mathbf{x}}_n | \widehat{\bar{\mathcal{H}}}_n) \right]$$
$$+ \frac{1}{N} \sum_{n=N_t}^{N-1} \mathbb{E} \left[ \max_{p_s(.), \mathbb{E}[||\tilde{\mathbf{s}}_n||^2] = \mathcal{P}_d} I(\tilde{\mathbf{r}}_n; \tilde{\mathbf{s}}_n | \widehat{\bar{\mathcal{H}}}_n) \right]$$

bits/channel use.

$$\tag{40}$$

## 5.1 Upper bound on the channel capacity

To benchmark the maximum achievable capacity, we consider the ideal scenario where the channel estimation is perfect. We also utilize the Gaussianity assumption on the distribution of the information symbol vectors, $\tilde{\mathbf{x}}_n$ and $\tilde{\mathbf{s}}_n$ due to (A2) in the channel capacity expression. We now have the following result:

**Theorem 4.** The upper bound on the channel capacity for the system model shown in (20), (21), and (24) is

obtained when the information symbol vectors, $\tilde{\mathbf{x}}_n$ and $\tilde{\mathbf{s}}_n$ are Gaussian distributed and is given by the expression:

$$C_u = \frac{N_t}{N} \times \frac{M}{P} \mathbb{E} \left[ \max_{p_x(.), \mathbb{E}[||\tilde{\mathbf{x}}_n||^2] = \mathcal{P}_{dt}} I(\tilde{\mathbf{y}}_0^{(dt)}; \tilde{\mathbf{x}}_0 | \widetilde{\mathcal{H}}_0) \right]$$
$$+ \frac{N_d}{N} \mathbb{E} \left[ \max_{p_s(.), \mathbb{E}[||\tilde{\mathbf{s}}_n||^2] = \mathcal{P}_d} I(\tilde{\mathbf{r}}_{N_t}; \tilde{\mathbf{s}}_{N_t} | \overline{\mathcal{H}}_{N_t}) \right]$$
$$\text{bits/channel use}$$
$$= \frac{N_t}{N} \times \frac{M}{P} \mathbb{E} \left[ \log \det \left( \mathbf{I}_{KM} + \frac{\mathcal{P}_{dt}}{\sigma_z^2} \widetilde{\mathcal{H}}_0 \widetilde{\mathcal{H}}_0^{\mathsf{H}} \right) \right]$$
$$+ \frac{N_d}{N} \mathbb{E} \left[ \log \det \left( \mathbf{I}_{KP} + \frac{\mathcal{P}_d}{\sigma_z^2} \overline{\mathcal{H}}_{N_t} \overline{\mathcal{H}}_{N_t}^{\mathsf{H}} \right) \right]$$
$$\text{bits/channel use.} \tag{41}$$

*Proof.* See Appendix 1.5. □

## 5.2 Lower bound on the channel capacity

From [4] and ([18], Theorem 1), we know that the lower bound on the mutual information between the channel input and its output is obtained when the additive noise is Gaussian distributed. In other words, when imperfect channel estimates are employed for estimating information symbols, a zero-mean uncorrelated complex Gaussian noise vector minimizes the upper bound over the distribution of the information symbol vector of the mutual information between the transmitted and observed information symbols. For the problem under consideration, the following signal model can be written by expressing the estimated channel matrix, as a sum of the conditional mean, and the random error component,

$$\tilde{\mathbf{y}}_n^{(dt)} = \widehat{\bar{\mathcal{H}}}_n \tilde{\mathbf{x}}_n + \check{\bar{\mathcal{H}}}_n \tilde{\mathbf{x}}_n + \tilde{\mathbf{z}}_n$$
$$= \widehat{\bar{\mathcal{H}}}_n \tilde{\mathbf{x}}_n + (\mathbf{I}_K \otimes \tilde{\mathbf{X}}_n) \check{\mathbf{h}}_n + \tilde{\mathbf{z}}_n. \tag{42a}$$

$$\tilde{\mathbf{r}}_n = \widehat{\overline{\mathcal{H}}}_n \tilde{\mathbf{s}}_n + \check{\overline{\mathcal{H}}}_n \tilde{\mathbf{s}}_n + \tilde{\mathbf{z}}_n$$
$$= \widehat{\overline{\mathcal{H}}}_n \tilde{\mathbf{s}}_n + (\mathbf{I}_K \otimes \tilde{\mathbf{S}}_n) \check{\mathbf{h}}_n + \tilde{\mathbf{z}}_n. \tag{42b}$$

In (42a), we made use of the following relationship,

$$\mathbf{P}_{0:M-1}^{(\mathbf{Q})\mathsf{T}} \bar{\mathbf{H}}_{k,\bar{k}}^{(n)} \mathbf{P}_{0:M-1}^{(\mathbf{Q})} \tilde{\mathbf{x}}_{n,\bar{k}} = \sqrt{P} \tilde{\mathbf{X}}_{n,\bar{k}} \mathbf{P}_{0:M-1}^{(\mathbf{Q})\mathsf{T}} \mathcal{W}_{0:L} \mathbf{h}_{k,\bar{k}}^{(n)}$$
$$= \tilde{\mathbf{X}}_{n,\bar{k}} \check{\mathcal{W}}_{0:L} \mathbf{h}_{k,\bar{k}}^{(n)}, \tag{43}$$

where $\tilde{\mathbf{X}}_{n,\bar{k}} \triangleq \text{diag}(\tilde{\mathbf{x}}_{n,\bar{k}})$, $\check{\mathcal{W}}_{0:L} \triangleq \sqrt{P} \mathbf{P}_{0:M-1}^{(\mathbf{Q})\mathsf{T}} \mathcal{W}_{0:L}$ and

$$\tilde{\mathbf{X}}_n \triangleq [\tilde{\mathbf{X}}_{n,1} \check{\mathcal{W}}_{0:L} \quad \tilde{\mathbf{X}}_{n,2} \check{\mathcal{W}}_{0:L} \quad ... \quad \tilde{\mathbf{X}}_{n,K} \check{\mathcal{W}}_{0:L}]. \tag{44}$$

Similarly, in (42b), we made use of the following relationship,

$$\bar{\mathbf{H}}_{k,\bar{k}}^{(n)} \tilde{\mathbf{s}}_{n,\bar{k}} = \sqrt{P} \tilde{\mathbf{S}}_{n,\bar{k}} \mathcal{W}_{0:L} \mathbf{h}_{k,\bar{k}}^{(n)}, \tag{45}$$

where $\tilde{\mathbf{S}}_{n,\bar{k}} \triangleq \mathrm{diag}\,(\tilde{\mathbf{s}}_{n,\bar{k}})$ and

$$\tilde{\mathbf{S}}_n \triangleq [\ \sqrt{P}\,\tilde{\mathbf{S}}_{n,1}\,\boldsymbol{\mathcal{W}}_{0:L}\ \ \sqrt{P}\,\tilde{\mathbf{S}}_{n,2}\,\boldsymbol{\mathcal{W}}_{0:L}\ \ ...\ \ \sqrt{P}\,\tilde{\mathbf{S}}_{n,K}\,\boldsymbol{\mathcal{W}}_{0:L}]. \tag{46}$$

It should be observed that in (21) and (24), the channel is unknown whereas in (42a) and (42b), the channel is known. Furthermore, the additive noise in the former two equations is Gaussian and independent of the information symbols whereas in the latter two, it is possibly neither. This is due to the fact that each of the effective additive noise vectors, $\tilde{\mathbf{z}}_n^{(dt)} \triangleq (\mathbf{I}_K \otimes \tilde{\mathbf{X}}_n)\,\check{\mathbf{h}}_n + \tilde{\mathbf{z}}_n$ and $\tilde{\mathbf{z}}_n^{(d)} \triangleq (\mathbf{I}_K \otimes \tilde{\mathbf{S}}_n)\,\check{\mathbf{h}}_n + \tilde{\mathbf{z}}_n$ appear to be a sum of a Gaussian vector and a vector whose elements are obtained by summing products of Gaussian random variables. As a result, we will merely derive the lower bound by replacing the effective noise vectors, with Gaussian noise vectors that possess the same average powers. The expressions for the average noise powers in each phase are as shown below.

### 5.2.1 Training phase

$$
\begin{aligned}
\sigma_{z^{(dt)}}^2(n) &= \frac{1}{KM}\mathbb{E}\{\,\mathrm{trace}\{\,\tilde{\mathbf{z}}_n^{(dt)}\,\tilde{\mathbf{z}}_n^{(dt)\mathsf{H}}\,\}\,\} \\
&= \frac{1}{M}\,\mathbf{m}_k^{(\infty)}\mathbb{E}\{\,\mathrm{trace}\{\tilde{\mathbf{X}}_n\,\tilde{\mathbf{X}}_n^{\mathsf{H}}\}\,\} + \sigma_z^2 \\
&= \frac{1}{M}\,\mathbf{m}_n^{(\infty)}\sum_{k=1}^{K}\mathbb{E}\{\,\mathrm{trace}\{\tilde{\mathbf{X}}_{n,k}\,\check{\boldsymbol{\mathcal{W}}}_{0:L}\,\check{\boldsymbol{\mathcal{W}}}_{0:L}^{\mathsf{H}}\,\tilde{\mathbf{X}}_{n,k}^{\mathsf{H}}\}\,\} + \sigma_z^2 \\
&= \frac{L+1}{M}\,\mathbf{m}_n^{(\infty)}\sum_{k=1}^{K}\sum_{m=0}^{M-1}\mathbb{E}\{\,|\tilde{\mathbf{X}}_{n,k}[\,m,m\,]\,|^2\,\} + \sigma_z^2 \\
&= \frac{L+1}{M}\,\mathbf{m}_n^{(\infty)}M\sigma_x^2 + \sigma_z^2 \\
&= (L+1)\mathcal{P}_{dt}\,\mathbf{m}_n^{(\infty)} + \sigma_z^2 \qquad n\in[\,0,N_t-1\,],
\end{aligned}
\tag{47}
$$

where we substituted, $\sigma_x^2 = \mathcal{P}_{dt}$ to account for the power budget on the transmit symbols in the training phase.

### 5.2.2 Data transmission phase

$$
\begin{aligned}
\sigma_{z^{(d)}}^2(n) &= \frac{1}{KP}\mathbb{E}\{\,\mathrm{trace}\{\,\tilde{\mathbf{z}}_n^{(d)}\,\tilde{\mathbf{z}}_n^{(d)\mathsf{H}}\,\}\,\} \\
&= \frac{1}{P}\,\mathbf{m}_n^{(\infty)}\mathbb{E}\{\,\mathrm{trace}\{\tilde{\mathbf{S}}_n\,\tilde{\mathbf{S}}_n^{\mathsf{H}}\}\,\} + \sigma_z^2 \\
&= \frac{1}{P}\,\mathbf{m}_n^{(\infty)}\sum_{k=1}^{K}\mathbb{E}\{\,\mathrm{trace}\{P\tilde{\mathbf{S}}_{n,k}\,\check{\boldsymbol{\mathcal{W}}}_{0:L}\,\check{\boldsymbol{\mathcal{W}}}_{0:L}^{\mathsf{H}}\,\tilde{\mathbf{S}}_{n,k}^{\mathsf{H}}\}\,\} + \sigma_z^2 \\
&= \frac{L+1}{P}\,\mathbf{m}_n^{(\infty)}\sum_{k=1}^{K}\sum_{p=0}^{P-1}\mathbb{E}\{\,|\tilde{\mathbf{S}}_{n,k}[\,p,p\,]\,|^2\,\} + \sigma_z^2 \\
&= \frac{L+1}{P}\,\mathbf{m}_n^{(\infty)}K\sigma_s^2 + \sigma_z^2 \\
&= (L+1)\mathcal{P}_d\,\mathbf{m}_n^{(\infty)} + \sigma_z^2 \qquad n\in[\,N_t,N-1\,],
\end{aligned}
\tag{48}
$$

where we substituted, $\sigma_s^2 = \mathcal{P}_d$ to account for the power budget on the transmit symbols in the data transmission phase.

The lower bound on the channel capacity when the estimated MIMO channels are taken to be the true channels is now given by the following result.

**Theorem 5.** The worst-case lower bound on the channel capacity for the system model shown in (20), (21), and (24) is obtained when the additive noise is Gaussian distributed and is maximized when the information symbol vectors, $\tilde{\mathbf{x}}_n$ and $\tilde{\mathbf{s}}_n$ are Gaussian distributed. It is given by the expression:

$$
\begin{aligned}
C_l &= \frac{1}{N}\times\frac{M}{P}\sum_{n=0}^{N_t-1}\mathbb{E}\left[\max_{p_x(.),\,\mathbb{E}[\|\tilde{\mathbf{x}}_n\|^2]=\mathcal{P}_{dt}} \mathrm{I}\,(\,\tilde{\mathbf{y}}_n^{(dt)};\,\tilde{\mathbf{x}}_n\,|\,\widehat{\tilde{\boldsymbol{\mathcal{H}}}}_n\,)\right] \\
&\quad + \frac{1}{N}\sum_{n=N_t}^{N-1}\mathbb{E}\left[\max_{p_s(.),\,\mathbb{E}[\|\tilde{\mathbf{s}}_n\|^2]=\mathcal{P}_d} \mathrm{I}\,(\,\tilde{\mathbf{r}}_n;\,\tilde{\mathbf{s}}_n\,|\,\widehat{\tilde{\boldsymbol{\mathcal{H}}}}_n\,)\right]
\end{aligned}
$$

bits/channel use

$$
\begin{aligned}
&\geq \frac{1}{N}\times\frac{M}{P}\sum_{n=0}^{N_t-1}\log\det\,(\mathbf{I}_{KM} + \mathcal{P}_{dt}\,\sigma_{z^{(dt)}}^2(n)\,\widehat{\tilde{\boldsymbol{\mathcal{H}}}}_n^{\mathsf{H}}\,\widehat{\tilde{\boldsymbol{\mathcal{H}}}}_n) \\
&\quad + \frac{1}{N}\sum_{n=N_t}^{N-1}\log\det\,(\mathbf{I}_{KP} + \mathcal{P}_d\,\sigma_{z^{(d)}}^2(n)\,\widehat{\tilde{\boldsymbol{\mathcal{H}}}}_n^{\mathsf{H}}\,\widehat{\tilde{\boldsymbol{\mathcal{H}}}}_n). \quad (49)
\end{aligned}
$$

*Proof.* See Appendix 1.6. □

## 6 Simulation results

In our simulation, we selected $K = 2$, $P = 32$, $L = 3$, and $M = 24$ (since $M = P - V$ and $V$ is set to $K(L+1)$). The training vectors are generated according to (32). We also set $\mathcal{P} = 1$, so that the SNR is defined as: $SNR \triangleq -10\log_{10}\sigma_z^2$. In designing the optimal training vector and in generating Gaussian information symbols over each of the $K$ transmit antennas, their variances have been appropriately scaled such that the total power constraint on the overall system is satisfied. We selected the Rayleigh channel variance to be $\sigma_h^2 = 1/(L+1)$. Thus, the Rayleigh channel adopted is an uncorrelated uniform scattering model. Moreover, we averaged the results over 500 randomly generated MIMO channel vectors. Given the fact that the channel capacity lower bound given by (49) is quite involved, we do not attempt to provide analytical results for the optimal power allocation and the optimal number of blocks out of $N$ that carry superimposed training symbols. Consequently, we resort to numerical optimization to determine optimal $\mathcal{P}_{dt}$, $\mathcal{P}_t$, $\mathcal{P}_d$, and $N_t$.

## 6.1 Performance evaluation of optimal training designs over non-time-varying wireless channels

In this section, we have generated the Rayleigh channels such that there is no correlation between successive block indices. In other words, each MIMO channel vector of any index is assumed to be independent of the MIMO channel vector of any other index. Moreover, we consider each block to contain training and information symbols such that channel tracking is not performed. Thus each block is represented by (20) and (21) alone. This also implies that $\mathcal{P}_{dt} + \mathcal{P}_t = 1$ and $\mathcal{P}_d = 0$ are assumed.

### 6.1.1 Comparison of the MIMO channel estimator and BCRB

Figure 1 provides a comparison of the Bayesian CRBs with the corresponding variances of channel estimators for varying training powers. In this figure, we normalized the Bayesian CRB and the channel variance values by the number of MIMO channel coefficients, i.e., $K^2(L + 1)$. From Figure 1, we notice that when a large fraction of total power is allocated for training symbols, the lower bound of the channel estimator progressively decreases. When the training symbols carry a small fraction of the total power, the under-performance of the MMSE channel estimator that is evaluated based solely on (20) w.r.t the lower bound is evident. On the other hand, as the power of training symbols increases, the difference between the MMSE channel estimator variance and the Bayesian lower bound achievable is negligible. In other words, the role played by the term, $\sigma_x^4 \, \mathbf{\Xi}(\mathbf{Q})$ in (11) is progressively minimized.

### 6.1.2 MMSE estimation of information symbols

Figure 2 describes the performance of an MMSE equalizer for estimating the information symbols using (21). We provide the MMSE variance characteristics for the case when the true channel was used in (21) as well as the case where estimated channels were used. Unsurprisingly, the
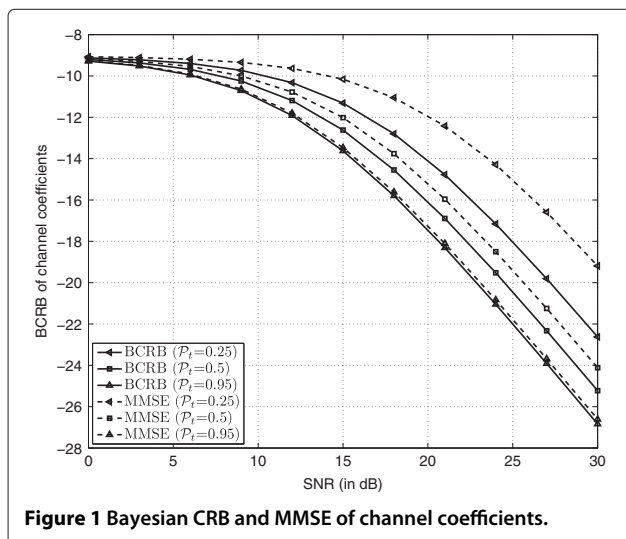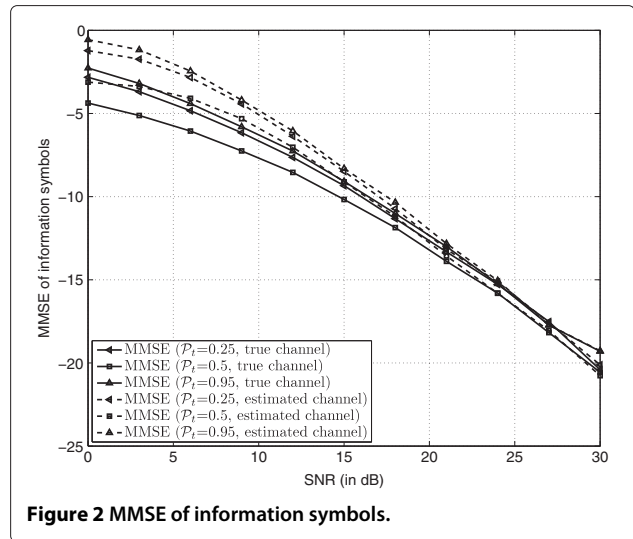
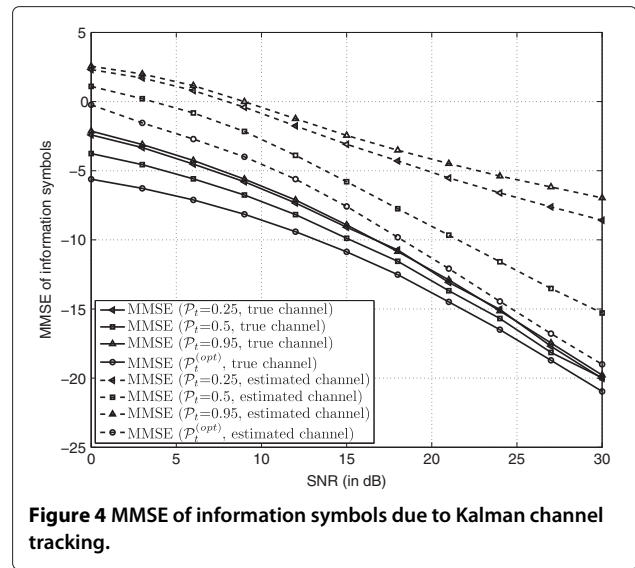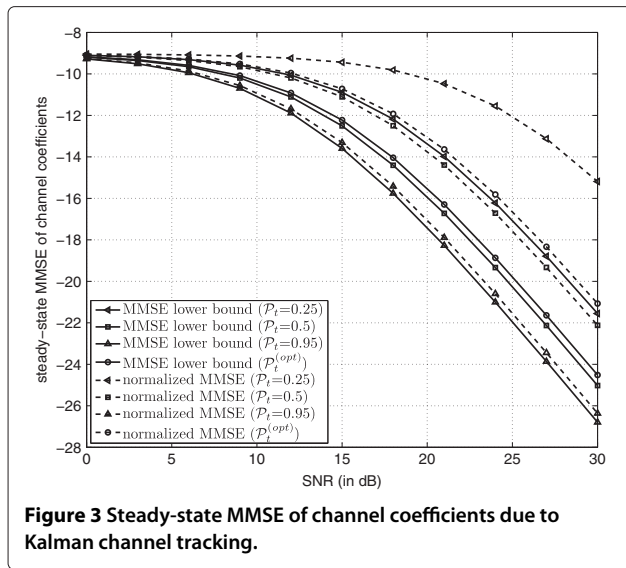**Figure 2 MMSE of information symbols.**

curves corresponding to true channel values suggest that the MMSE variance of the estimated information symbols is lower than those that result when estimated channel vectors are used. However, a more interesting observation is that the performance is impaired both when $\mathcal{P}_t$ is too low or too high. Specifically, when $\mathcal{P}_t = 0.25$, a small fraction of the total power is employed to gather channel estimates. Since, there are bound to be numerous errors in this scenario, data estimation suffers. Conversely, when $\mathcal{P}_t = 0.95$, only a small fraction of total power is expended for information symbols and hence the data estimation suffers. On the other hand, when $\mathcal{P}_t = 0.5$, the performance appears to be better. However, we refrain from computing the optimal power allocation by considering a capacity lower bound similar to (49) for the non-time-varying wireless channel scenario and reserve such an analysis for the next section.

## 6.2 Kalman tracking of time-varying wireless channels

In this section, we selected the MIMO channel vectors such that they are correlated. The excitation noise is generated with the appropriate variance so that the channel vectors are WSS. Throughout this section, we set $N = 10$. We then performed a numerical optimization as mentioned above and determined that the following values result maximize the lower bound obtained in (49): $\mathcal{P}_{dt} = 0.32$, $\mathcal{P}_t = 0.41$, $\mathcal{P}_d = 0.27$ and $N_t = 4$. When a non-optimal value of $\mathcal{P}_{dt}$ is allocated to training symbols, the division of the remaining power to information symbols in the training phase and the data transmission phase is arbitrarily chosen.

### 6.2.1 Steady-state MMSE of the channel estimator

In Figure 3, we provide the steady-state MMSE characteristics of the channel estimator when a Kalman filter is used to track the channel. We set $a = 0.95$ and fixed $N_t = 4$ in

**Figure 1 Bayesian CRB and MMSE of channel coefficients.**

**Figure 3 Steady-state MMSE of channel coefficients due to Kalman channel tracking.**



**Figure 4 MMSE of information symbols due to Kalman channel tracking.**

order to generate these characteristics. We also analyzed the characteristics for $\mathcal{P}_t = 0.25, 0.5$ and $0.95$ in addition to the optimal value. The MMSE lower bound shown in Figure 3 corresponds to (36) whereas the normalized MMSE corresponds to averaging (38b) and (38c) every $N$ blocks. We notice that for small $\mathcal{P}_t$ the errors committed due to channel predictions in the data transmission phase cause significant deviation of the normalized steady-state MMSE from the lower bound. Only at high values of $\mathcal{P}_t$, these errors become insignificant. Of particular importance is the fact that even at an SNR of 30 dB and with optimal training power allocation, the deterioration suffered due to prediction errors w.r.t the lower bound is close to 3 dB.

#### 6.2.2 MMSE estimation of information symbols due to Kalman channel tracking

Figure 4 shows the resulting MMSE estimation error variance characteristics of information symbols. The solid curves rely on true channel estimates whereas the dashed curves depend on not only the estimated channel states, but also on Kalman predictions. Similar to Figure 2, we see that when $\mathcal{P}_t$ is too small or too large the error variance of information symbol estimation suffers greatly. Even at high SNR, the non-judicious power allocation combined with Kalman predictions during the data transmission phase leads to numerous errors. On the other hand, optimal power allocation between training and information symbols leads to the lowest possible information symbol estimation error variance.

#### 6.2.3 Capacity bounds

The final simulation example that we will consider is the capacity upper and lower bound characteristics. While the upper bound characteristics for varying levels of $\mathcal{P}_t$

exhibit a gradual improvement toward the theoretical upper bound, the lower bound characteristics are more abrupt. This can be attributed to the prediction errors that occur in the Kalman prediction stage during the data transmission phase. When $\mathcal{P}_t = 0.95$, the fraction of the total power available for information symbols in the training phase and the data transmission phase is minuscule and hence the achievable capacity lower bound is small. In contrast, this value can be improved by more than 15 bits/channel use with an optimal allocation of training and data powers (Figure 5).

### 7 Conclusion

In this article, we have shown that similar to a SISO case, an OFDM linear precoder with an FDM training sequence satisfies the orthogonality condition and results in decoupled channel estimation and symbol detection.



**Figure 5 Capacity upper and lower bounds.**
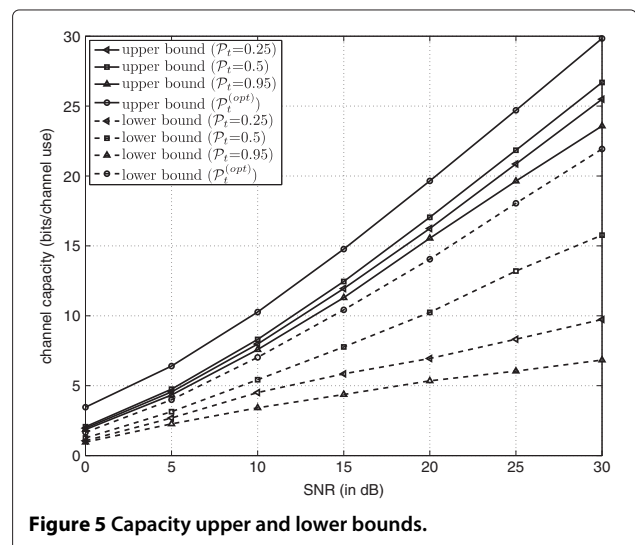
Furthermore, we have derived optimal training sequences such that the FDM training sequences between different antennas are phase-shift orthogonal to each other. Based on the structure of the training matrices, the Kalman filter recursion was simplified to a scalar recursion. Eventually, the upper and lower bounds on the channel capacity were obtained by utilizing the Kalman filter's MMSE expressions to account for imperfect channel estimates. We showed that the Kalman filter predictions affect the capacity calculations substantially. Taking this degradation into account, we numerically determined the optimal training power allocation and optimal number of training blocks to achieve the best possible capacity lower bound. Finally, we provided numerical results to support the theoretical results.

## 8 Appendix 1
### 1.1 Modeling time-variations in low-mobility wireless channels
While the complex random variable description of a wireless channel as *Rayleigh*, *Rician*, etc., forms one aspect of characterization, another one involves taking the time-variations of the channel filter taps into consideration. A common assumption on the random process that drive the time-variations of the channel filter taps is its wide-sense stationarity. In other words, the mean and the auto-correlation functions of each filter tap are assumed to be independent of time, with the latter being a function of the time-difference alone. Further, each tap at a given time instant is assumed to be independent of every other tap at any time instant. Together these two assumptions give rise to the wide-sense stationary, uncorrelated scattering (WSSUS) model.

*Autoregressive model*: A widely used approach to model time variations of a WSSUS channel is by a general $P$th order AR random process. By considering (7a), the AR model that helps us to specify the correlation between the current state of the system and the past states is as shown below:

$$\mathbf{h}_n = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{h}_{n-p} + \mathbf{B}\,\mathbf{u}_n. \tag{A.1}$$

In (A.1), each element in $\{\mathbf{A}_p\}$ is termed as an AR coefficient matrix or a state-transition matrix and $\mathbf{u}_n$ as the excitation or driving noise vector. The eigen values of each element in $\{\mathbf{A}_p\}$ are assumed to be less than 1 in magnitude and the driving noise is assumed to be i.i.d and complex Gaussian distributed with zero mean. The AR model admits the following Yule–Walker equations to describe the covariance function of the process [37].

$$\mathbf{R_h}[\,n\,] = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{R_h}[\,n-p\,] + \sigma_u^2 \mathbf{B}\,\mathbf{B}^{\mathsf{H}}. \tag{A.2}$$

Assuming that $\mathbf{R_h}[\,0\,]$ and $\{\mathbf{A}_p\}$ are known, we can apply the fact that $\mathbf{R_h}[\,n\,] = \mathbf{R_h}[\,-n\,]$, and recursively find $\{\mathbf{R_h}[\,n\,]\}$ for $n = 1, 2, \ldots, P$. We can also find a non-unique $\mathbf{B}$ by computing the square-root of (A.2) for $n = 0$ ([38], p. 358).

### 1.2 Proof of Theorem 1
From [39], we know that the complex FIM is given by the equation,

$$
\begin{aligned}
&\mathbb{E}\left\{ \left( \frac{\partial \ln p(\mathbf{y}_n; \mathbf{h}_n)}{\partial \mathbf{h}_n^*} \right) \left( \frac{\partial \ln p(\mathbf{y}_n; \mathbf{h}_n)}{\partial \mathbf{h}_n^*} \right)^{\mathsf{H}} \right\} \\
&= \mathbb{E}\left\{ \mathbb{E}\left[ \left( \frac{\partial \ln p(\mathbf{y}_n|\mathbf{h}_n)}{\partial \mathbf{h}_n^*} \right) \left( \frac{\partial \ln p(\mathbf{y}_n|\mathbf{h}_n)}{\partial \mathbf{h}_n^*} \right)^{\mathsf{H}} | \mathbf{h}_n \right] \right\} \\
&+ \mathbb{E}\left\{ \left( \frac{\partial \ln p(\mathbf{h}_n)}{\partial \mathbf{h}_n^*} \right) \left( \frac{\partial \ln p(\mathbf{h}_n)}{\partial \mathbf{h}_n^*} \right)^{\mathsf{H}} \right\}. \tag{B.1}
\end{aligned}
$$

In the second equality of the above equation, the inner expectation in the first term is w.r.t $\mathbf{y}_n$, whereas the outer expectation is w.r.t $\mathbf{h}_n$. The log-likelihood function of the probability density function, $p(\mathbf{y}_n|\mathbf{h}_n)$ in (B.1) and its derivative are as follows:

$$\ln p(\mathbf{y}_n|\mathbf{h}_n) = constant - \ln\left(|\mathbf{R}_{\mathbf{y}_n}|\right) - \mathbf{u}^{\mathsf{H}} \mathbf{R}_{\mathbf{y}_n}^{-1}\, \mathbf{u}, \tag{B.2a}$$

$$\frac{\partial \ln p(\mathbf{y}_n|\mathbf{h}_n)}{\partial \mathbf{h}_n^*} = - \frac{\partial \ln |\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}|}{\partial \mathbf{h}_n^*} - \frac{\partial \mathbf{u}^{\mathsf{H}} \mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1} \mathbf{u}}{\partial \mathbf{h}_n^*}, \tag{B.2b}$$

where, $\mathbf{u} \triangleq (\mathbf{y}_n - (\mathbf{I}_K \otimes \mathbf{T})\,\mathbf{h}_n)$ and

$$
\begin{aligned}
\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n} &= \sigma_x^2\, \mathcal{H}_n\, \mathbf{Q}\, \mathbf{Q}^{\mathsf{H}} \mathcal{H}_n^{\mathsf{H}} + \sigma_z^2 \mathbf{I}_{KP} \\
&= \sigma_x^2 \sum_{j=0}^{KM-1} \mathcal{Q}_j \mathbf{h}_n \mathbf{h}_n^{\mathsf{H}} \mathcal{Q}_j^{\mathsf{H}} + \sigma_z^2 \mathbf{I}_{KP}, \tag{B.3a}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1} &= \sigma_z^{-2} \mathbf{I}_{KP} - \sigma_z^{-4} \sigma_x^2 \mathcal{H}_n\, \mathbf{Q} \\
&\quad \times \left( \mathbf{I}_{KM} + \sigma_z^{-2} \sigma_x^2 \mathbf{Q}^{\mathsf{H}} \mathcal{H}_n^{\mathsf{H}} \mathcal{H}_n \mathbf{Q} \right)^{-1} \mathbf{Q}^{\mathsf{H}} \mathcal{H}_n^{\mathsf{H}} \\
&= \sigma_z^{-2} \mathbf{I}_{KP} - \sigma_z^{-4} \sigma_x^2 \mathcal{H}_n\, \mathbf{Q}\, \mathbf{G}\, \mathbf{Q}^{\mathsf{H}} \mathcal{H}_n^{\mathsf{H}}. \tag{B.3b}
\end{aligned}
$$

and $\mathcal{Q}_j \triangleq \mathbf{I}_K \otimes [\, \mathcal{Q}_{j,1}\ \mathcal{Q}_{j,2}\ \ldots\ \mathcal{Q}_{j,K} ]$ are obtained from each of the $KM$ columns of $\mathbf{Q}$. The matrices $\{\mathcal{Q}_{j,k}\}$ are a result of applying the commutativity property of convolution. It should be noted that we have not utilized the block-diagonal structure of $\mathbf{Q}$ in obtaining $\{\mathcal{Q}_{j,k}\}$. In other words, the matrices $\{\mathcal{Q}_{j,k}\}$ are constructed without explicit consideration of the fact that $(K-1)P$ out of $KP$ elements in each column of $\mathbf{Q}$ are zeros. In addition, we have utilized the matrix inversion lemma in obtaining (B.3b) where $\mathbf{G} \triangleq \left( \mathbf{I}_{KM} + \sigma_z^{-2} \sigma_x^2 \mathbf{Q}^{\mathsf{H}} \mathcal{H}_n^{\mathsf{H}} \mathcal{H}_n \mathbf{Q} \right)^{-1}$.

We now evaluate the two partial derivatives in (B.2b) separately.

$\frac{\partial \ln |\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}|}{\partial \mathbf{h}_n^*}$ Using ([40], (9)), we note that

$$
\begin{aligned}
\frac{\partial \ln |\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}|}{\partial \mathbf{h}_n^*} &= \mathcal{D}_{\mathbf{h}_n^*} \ln |\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}| \\
&= (\mathcal{D}_{\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}} \ln |\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}|)\, \mathcal{D}_{\mathbf{h}_n^*} \mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n} \\
&\quad + (\mathcal{D}_{\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^*} \ln |\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}|)\, \mathcal{D}_{\mathbf{h}_n^*} \mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^*.
\end{aligned}
\tag{B.4}
$$

Here, $\mathcal{D}_{\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}} \ln |\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}| = \mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-\top}$ and $\mathcal{D}_{\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^*} \ln |\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}| = 0$ ([41], Table II). Moreover, from (B.3a), we see that (cf. ([40], (1)),

$$
\begin{aligned}
d\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n} &= \sigma_x^2 \sum_{j=0}^{KM-1} \mathcal{Q}_j \mathbf{h}_n\, d\mathbf{h}_n^{\mathsf{H}} \mathcal{Q}_j^{\mathsf{H}} \\
&\quad + \sigma_x^2 \sum_{j=0}^{KM-1} \mathcal{Q}_j\, d\mathbf{h}_n \mathbf{h}_n^{\mathsf{H}} \mathcal{Q}_j^{\mathsf{H}},
\end{aligned}
\tag{B.5a}
$$

$$
\begin{aligned}
d\mathrm{vec}\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n} &= \sigma_x^2 \sum_{j=0}^{KM-1} (\mathcal{Q}_j^* \otimes \mathcal{Q}_n\mathbf{h}_n)\, d\,\mathrm{vec}(\mathbf{h}_n^*) \\
&\quad + \sigma_x^2 \sum_{j=0}^{KM-1} (\mathcal{Q}_j^*\mathbf{h}_n^* \otimes \mathcal{Q}_n)\, d\,\mathrm{vec}(\mathbf{h}_n).
\end{aligned}
\tag{B.5b}
$$

From the above equations and ([40], Table III), we notice that, $\mathcal{D}_{\mathbf{h}_n^*} \mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n} = \sigma_x^2 \sum_{j=0}^{KM-1} (\mathcal{Q}_j^* \otimes \mathcal{Q}_n\mathbf{h}_n)$. It should be noted that the definition of the partial derivative for the case of a scalar function w.r.t a column vector adopted by Hjørungnes and Gesbert result in a row vector ([40], Table III, 2nd row). We consider this definition to lead to transposed derivative and perform a transpose operation of the results obtained based on ([40], (9)) in order to obtain the FIM with appropriate dimensions. Consequently,

$$
\begin{aligned}
&\frac{\partial \ln |\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}|}{\partial \mathbf{h}_n^*} \\
&= (\mathcal{D}_{\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}} \ln |\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}|)\, \mathcal{D}_{\mathbf{h}_n^*} \mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n} \\
&= \left( \mathrm{vec}^{\top} \left[ \frac{\partial \ln |\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}|}{\partial \mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}} \right] \left[ \frac{\partial \mathrm{vec}\,\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}}{\partial^{\top}\mathbf{h}_n^*} \right] \right)^{\top} \\
&= \sigma_x^2 \sum_{j=0}^{KM-1} (\mathcal{Q}_j^{\mathsf{H}} \otimes \mathbf{h}_n^{\top}\mathcal{Q}_j^{\top})\, \mathrm{vec}(\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-\top}).
\end{aligned}
\tag{B.6}
$$

$\frac{\partial\, \mathbf{u}^{\mathsf{H}}\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1}\mathbf{u}}{\partial \mathbf{h}_n^*}$ Using ([40], (9)), we can similarly show that

$$
\begin{aligned}
\frac{\partial\, \mathbf{u}^{\mathsf{H}}\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1}\mathbf{u}}{\partial \mathbf{h}_n^*} &= -\sigma_x^2 \sum_{j=0}^{KM-1} (\mathcal{Q}_j^{\mathsf{H}} \otimes \mathbf{h}_n^{\top}\mathcal{Q}_j^{\top})(\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1} \otimes \mathbf{R}_{\mathbf{y}_n}^{-\top})(\mathbf{u} \otimes \mathbf{u}^*) \\
&\quad - (\mathbf{I}_K \otimes \mathbf{T})^{\mathsf{H}}\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1}\mathbf{u}.
\end{aligned}
\tag{B.7}
$$

Hence, from (B.2b),

$$
\begin{aligned}
\frac{\partial \ln p(\mathbf{y}|\mathbf{h}_n)}{\partial \mathbf{h}_n^*} &= \sigma_x^2 \sum_{j=0}^{KM-1} (\mathcal{Q}_j^{\mathsf{H}} \otimes \mathbf{h}_n^{\top}\mathcal{Q}_j^{\top}) \\
&\quad \times [\, (\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1} \otimes \mathbf{R}_{\mathbf{y}_n}^{-\top})(\mathbf{u} \otimes \mathbf{u}^*) - \mathrm{vec}(\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-\top})] \\
&\quad + (\mathbf{I}_K \otimes \mathbf{T})^{\mathsf{H}}\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1}\mathbf{u}.
\end{aligned}
\tag{B.8}
$$

Before we evaluate the inner expectation in the first term of (B.1), we recall that,

$$
\begin{aligned}
\mathbb{E}[\mathbf{u} \otimes \mathbf{u}^*] &= \mathbb{E}\left[\mathrm{vec}(\mathbf{u}^*\mathbf{u}^{\top})\right] = \mathrm{vec}(\mathbb{E}\left[\mathbf{u}^*\mathbf{u}^{\top}\right]) \\
&= \mathrm{vec}(\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{\top}).
\end{aligned}
\tag{B.9}
$$

Incidentally, by utilizing the above result, we can see that $\mathbb{E}\left[\frac{\partial \ln p(\mathbf{y}_n|\mathbf{h}_n)}{\partial \mathbf{h}_n^*}\right] = 0$ indicating that the regularity condition is satisfied. Employing (B.8) and (B.9), we can show that

$$
\begin{aligned}
&\mathbb{E}\left[ \left(\frac{\partial \ln p(\mathbf{y}|\mathbf{h}_n)}{\partial \mathbf{h}_n^*}\right) \left(\frac{\partial \ln p(\mathbf{y}|\mathbf{h}_n)}{\partial \mathbf{h}_n^*}\right)^{\mathsf{H}} |\mathbf{h}_n \right] \\
&= (\mathbf{I}_K \otimes \mathbf{T})^{\mathsf{H}}\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1}(\mathbf{I}_K \otimes \mathbf{T}) \\
&\quad + \sigma_x^4 \sum_{j=0}^{KM-1} (\mathcal{Q}_j^{\mathsf{H}} \otimes \mathbf{h}_n^{\top}\mathcal{Q}_j^{\top})(\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1} \otimes \mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-\top}) \\
&\quad \times \sum_{n=0}^{KM-1} (\mathcal{Q}_j^* \otimes \mathcal{Q}_n\mathbf{h}_n).
\end{aligned}
\tag{B.10}
$$

Moreover, we observe that $\mathbb{E}\left\{\left(\frac{\partial \ln p(\mathbf{h}_n)}{\partial \mathbf{h}_n^*}\right) \left(\frac{\partial \ln p(\mathbf{h}_n)}{\partial \mathbf{h}_n^*}\right)^{\mathsf{H}}\right\} = \sigma_h^{-2}\mathbf{I}_{K^2(L+1)}$. Substituting this result along with (B.10) and (B.3b) in (B.1) gives:

$$
\begin{aligned}
\mathcal{I}(\mathbf{h}_n) &= \mathbb{E}_{\mathbf{h}}\left\{ (\mathbf{I}_K \otimes \mathbf{T})^{\mathsf{H}}\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1}(\mathbf{I}_K \otimes \mathbf{T}) \right\} \\
&\quad + \sigma_x^4\, \Xi(\mathbf{Q}) + \sigma_h^{-2}\mathbf{I}_{K^2(L+1)} \\
&= \sigma_z^{-2}(\mathbf{I}_K \otimes \mathbf{T})^{\mathsf{H}}\,(\mathbf{I}_K \otimes \mathbf{T}) \\
&\quad - \sigma_z^{-4}\sigma_x^2\, \mathbb{E}_{\mathbf{h}}\left\{ (\mathbf{I}_K \otimes \mathbf{T})^{\mathsf{H}}\, \mathcal{H}_n\, \mathbf{Q}\, \mathbf{G}\, \mathbf{Q}^{\mathsf{H}}\mathcal{H}_n^{\mathsf{H}}(\mathbf{I}_K \otimes \mathbf{T}) \right\} \\
&\quad + \sigma_x^4\, \Xi(\mathbf{Q}) + \sigma_h^{-2}\mathbf{I}_{K^2(L+1)} \\
&= \sigma_z^{-2}(\mathbf{I}_K \otimes \mathbf{T}^{\mathsf{H}}\, \mathbf{T}) - \sigma_z^{-4}\sigma_x^2\, \Xi(\mathbf{t}, \mathbf{Q}) + \sigma_x^4\, \Xi(\mathbf{Q}) \\
&\quad + \sigma_h^{-2}\mathbf{I}_{K^2(L+1)}.
\end{aligned}
\tag{B.11}
$$

### 1.3 Proof of Theorem 2

From (12c), we see that $\mathbf{G}$ is the inverse of sum of two full-rank positive-definite matrices. This is because, $\mathcal{H}_n$ is a Rayleigh-fading channel matrix of full rank (with probability 1) due to (A1) and (C1) stipulates that $\mathbf{Q}$ be a full column-rank matrix. Hence, $\mathbf{Q}^{\mathsf{H}}\mathcal{H}_n^{\mathsf{H}}\mathcal{H}_n\mathbf{Q}$ is a matrix with strictly-positive eigenvalues. Together with the fact that $\mathbf{I}_{KM}$ is also a matrix with strictly-positive eigen-values, we arrive at the result that $\mathbf{G} \succ \mathbf{0}$. By making a similar argument, we can show that $\mathbf{R}_{\mathbf{y}_n|\mathbf{h}_n}^{-1} \succ \mathbf{0}$. As a result of the above statements, we can claim that $\boldsymbol{\Xi}(\mathbf{t}, \mathbf{Q}) \succeq \mathbf{0}$ and $\boldsymbol{\Xi}(\mathbf{t}) \succeq \mathbf{0}$. Combining the above results with (C1) leads us to conclude that $\mathcal{I}(\mathbf{h}_n) \succ \mathbf{0}$. Now, based on a previous observation that only $\boldsymbol{\Xi}(\mathbf{t}, \mathbf{Q})$ is the term under the designer's control, we see that $\mathcal{I}^{(opt)}(\mathbf{h}_n) \succeq \mathcal{I}(\mathbf{h}_n)$ where the optimal Bayesian FIM for a training-based channel estimator is as follows:

$$\mathcal{I}^{(opt)}(\mathbf{h}_n) = \sigma_z^{-2}(\mathbf{I}_K \otimes \mathbf{T})^{\mathsf{H}}(\mathbf{I}_K \otimes \mathbf{T}) + \sigma_x^4\,\boldsymbol{\Xi}(\mathbf{Q})$$
$$+ \sigma_h^{-2}\mathbf{I}_{K^2(L+1)}. \tag{C.1}$$

It should be recalled that for any $\mathbf{A}, \mathbf{B} \succ \mathbf{0}$ such that $\mathbf{A} \succeq \mathbf{B}$, we have $\mathbf{B}^{-1} \preceq \mathbf{A}^{-1}$ and therefore, $\mathrm{tr}\,(\mathbf{B}^{-1}) \preceq \mathrm{tr}\,(\mathbf{A}^{-1})$. Hence, finding the conditions under which $\mathcal{I}^{(opt)}(\mathbf{h}_n) \succeq \mathcal{I}(\mathbf{h}_n)$ is equivalent to finding the conditions under which $\mathrm{tr}\,(\mathcal{I}^{(opt)}(\mathbf{h}_n)^{-1}) \preceq \mathrm{tr}\,(\mathcal{I}(\mathbf{h}_n)^{-1})$. where $\mathrm{tr}\,(\mathcal{I}^{(opt)}(\mathbf{h}_n)^{-1})$ is the Bayesian CRB of a non-decision-aided channel estimator for the system model described in (8). We can now see that $\mathcal{I}^{(opt)}(\mathbf{h}_n)$ is obtained by making $\boldsymbol{\Xi}(\mathbf{t}, \mathbf{Q}) = 0$ which in turn is possible by enforcing the condition:

$$(\mathbf{I}_K \otimes \mathbf{T})^{\mathsf{H}}\mathcal{H}_n\mathbf{Q}$$
$$= \begin{bmatrix} \mathbf{T}^{\mathsf{H}}\mathbf{H}_{1,1}^{(n)}\mathbf{Q}_1 & \mathbf{T}^{\mathsf{H}}\mathbf{H}_{1,2}^{(n)}\mathbf{Q}_2 & \cdots & \mathbf{T}^{\mathsf{H}}\mathbf{H}_{1,K}^{(n)}\mathbf{Q}_K \\ \mathbf{T}^{\mathsf{H}}\mathbf{H}_{2,1}^{(n)}\mathbf{Q}_1 & \mathbf{T}^{\mathsf{H}}\mathbf{H}_{2,2}^{(n)}\mathbf{Q}_2 & \cdots & \mathbf{T}^{\mathsf{H}}\mathbf{H}_{2,K}^{(n)}\mathbf{Q}_K \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{T}^{\mathsf{H}}\mathbf{H}_{K,1}^{(n)}\mathbf{Q}_1 & \mathbf{T}^{\mathsf{H}}\mathbf{H}_{K,2}^{(n)}\mathbf{Q}_2 & \cdots & \mathbf{T}^{\mathsf{H}}\mathbf{H}_{K,K}^{(n)}\mathbf{Q}_K \end{bmatrix} \tag{C.2}$$
$$= \mathbf{0}.$$

In other words,

$$\mathbf{T}^{\mathsf{H}}\mathbf{H}_{i,j}^{(n)}\mathbf{Q}_j = \mathbf{0} \qquad 1 \le i, j \le K. \tag{C.3}$$

We now utilize the commutativity property of convolution and in a manner similar in the construction of the matrices, $\{\mathbf{T}_k\}$, we see that $\mathbf{H}_{i,j}^{(n)}\mathbf{Q}_j = [\mathbf{Q}_{j,1}\mathbf{h}_{i,j}^{(n)}$ $\mathbf{Q}_{j,2}\mathbf{h}_{i,j}^{(n)} \;\;\ldots\;\; \mathbf{Q}_{j,M-1}\mathbf{h}_{i,j}^{(n)}]$ where the circulant matrices, $\{\mathbf{Q}_{j,m}\}$ are constructed such that $[\,q_{j,m}[0]\;\ldots\;q_{j,m}[P-1]\,]^{\mathsf{T}}$ is the first column and $[\,q_{j,m}[0]\;\;t_k[P-L]\;\ldots\;q_{j,m}[P-1]\,]$ is the first row. The column vector $q_{j,m}$ is the $m$th column of the $j$th linear precoder, $\mathbf{Q}_j$. Hence,

$$\mathbf{T}_i^{\mathsf{H}}\mathbf{Q}_{j,m} = \mathbf{0} \qquad 1 \le i, j \le K \qquad 0 \le m \le M-1. \tag{C.4}$$

### 1.4 Proof of Lemma 1

First, it is easy to see that the optimal MMSE estimator coincides with the linear MMSE estimator for the system under consideration i.e., (20), due to the joint Gaussian nature of the unknown parameter and the observation vectors. The optimal MMSE channel estimator, $\hat{\mathbf{h}}_n$ is now ([31], (11.33) and (11.35)):

$$\hat{\mathbf{h}}_n = \sigma_z^{-2}[\,\sigma_{\mathbf{h}}^{-2}\mathbf{I}_{K^2(L+1)} + \sigma_z^{-2}\,(\mathbf{I}_K \otimes \tilde{\mathbf{T}}^{\mathsf{H}}\tilde{\mathbf{T}})\,]^{-1}$$
$$\times (\mathbf{I}_K \otimes \tilde{\mathbf{T}}^{\mathsf{H}})\,\tilde{\mathbf{y}}_n, \tag{D.1a}$$
$$\mathbf{C}_{\Delta\mathbf{h}_n} = [\,\sigma_{\mathbf{h}}^{-2}\mathbf{I}_{K^2(L+1)} + \sigma_z^{-2}\,(\mathbf{I}_K \otimes \tilde{\mathbf{T}}^{\mathsf{H}}\tilde{\mathbf{T}})\,]^{-1}. \tag{D.1b}$$

where, $\Delta\mathbf{h}_n = \mathbf{h}_n - \hat{\mathbf{h}}_n$. The resulting channel estimator error variance is,

$$\sigma_{\Delta\mathbf{h}_n}^2 = \mathrm{trace}\,\{\mathbf{C}_{\Delta\mathbf{h}_n}\}$$
$$= \mathrm{trace}\,\{[\,\sigma_{\mathbf{h}}^{-2}\mathbf{I}_{K^2(L+1)} + \sigma_z^{-2}\,(\mathbf{I}_K \otimes \tilde{\mathbf{T}}^{\mathsf{H}}\tilde{\mathbf{T}})\,]^{-1}\}. \tag{D.2}$$

The optimal $\tilde{\mathbf{T}}$, $\tilde{\mathbf{T}}^{(opt)}$ needs to minimize $\sigma_{\Delta\mathbf{h}_n}^2$ subject to (27). An equivalent representation of this pilot power constraint that will be useful for finding $\tilde{\mathbf{T}}^{(opt)}$ is as follows:

$$\mathrm{trace}\,(\tilde{\mathbf{T}}_k^{\mathsf{H}}\tilde{\mathbf{T}}_k) = \mathcal{P}_t/K, \tag{D.3a}$$
$$\mathrm{trace}\,(\tilde{\mathbf{T}}^{\mathsf{H}}\tilde{\mathbf{T}}) = (L+1)\mathcal{P}_t. \tag{D.3b}$$

As a result of (D.3b), we have

$$\tilde{\mathbf{T}}^{(opt)} = \arg_{\tilde{\mathbf{T}}}\;\min_{\mathrm{trace}\,(\tilde{\mathbf{T}}^{\mathsf{H}}\tilde{\mathbf{T}})=(L+1)\mathcal{P}_t}\;\sigma_{\Delta\mathbf{h}_n}^2. \tag{D.4}$$

From ([15], Appendix I), we see that for any $M \times M$ dimensional positive-definite matrix $\mathbf{A}$,

$$\mathrm{trace}(\mathbf{A}^{-1}) \ge \sum_{m=0}^{M-1}\frac{1}{a_{m,m}}, \tag{D.5}$$

where the equality is attained if and only if $\mathbf{A}$ is diagonal. Therefore, if $\tilde{\mathbf{T}}^{(opt)}$ is employed to perform the MIMO–OFDM channel estimation, the resulting variance of the MMSE channel estimator is as follows:

$$\sigma_{\Delta\mathbf{h}_n}^2 = \mathrm{trace}\,\{[\,\sigma_h^{-2}\mathbf{I}_{K^2(L+1)} + \sigma_z^{-2}\,(\mathbf{I}_K \otimes \tilde{\mathbf{T}}^{\mathsf{H}}\tilde{\mathbf{T}})\,]^{-1}\}$$
$$\ge \sum_{l=0}^{K^2(L+1)-1}\frac{1}{[\,\sigma_h^{-2}\mathbf{I}_{K^2(L+1)} + \sigma_z^{-2}\,(\mathbf{I}_K \otimes \tilde{\mathbf{T}}^{\mathsf{H}}\tilde{\mathbf{T}})\,]_{l,l}}$$
$$= \frac{K^3(L+1)\,\sigma_z^2\,\sigma_h^2}{K\sigma_z^2 + \sigma_h^2\,\mathcal{P}_t}, \tag{D.6}$$

and equality in the above equation is attained when $\tilde{\mathbf{T}}^{\mathsf{H}}\tilde{\mathbf{T}}$ is diagonal.

#### 1.4.1 *Optimal training design*

We now design the optimal training design that achieves the minimum MSE variance shown in (D.6). We will see that in order to attain this bound, the pilot sequences of each transmit antenna as well as their relationship with

the training sequences emitted from every other transmit antenna need to satisfy certain specific properties. These properties are a direct consequence of (D.3b) and (D.5).

A closer observation of $\tilde{\mathbf{T}}^{\mathsf{H}} \tilde{\mathbf{T}}$ reveals the following:

$$\tilde{\mathbf{T}}^{\mathsf{H}} \tilde{\mathbf{T}} = \begin{bmatrix} \mathbf{R}_{1,1} & \dots & \mathbf{R}_{1,K} \\ \mathbf{R}_{2,1} & \dots & \mathbf{R}_{2,K} \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{K,1} & \dots & \mathbf{R}_{K,K} \end{bmatrix}, \qquad (D.7)$$

where the $(L+1) \times (L+1)$ dimensional submatrix $\mathbf{R}_{k_1,k_2}$ is defined based on (19) as:

$$\mathbf{R}_{k_1,k_2} \triangleq \hat{\mathcal{W}}_{0:L}^{\mathsf{H}} \tilde{\mathbf{T}}_{k_1}^{\mathsf{H}} \tilde{\mathbf{T}}_{k_2} \hat{\mathcal{W}}_{0:L}. \qquad (D.8)$$

The minimum variance as shown in (D.6) is therefore attained when

$$\mathbf{R}_{k_1,k_2} = \hat{\mathcal{W}}_{0:L}^{\mathsf{H}} \tilde{\mathbf{T}}_{k_1}^{\mathsf{H}} \tilde{\mathbf{T}}_{k_2} \hat{\mathcal{W}}_{0:L} = \frac{\bar{\mathcal{P}}}{K} \mathbf{I}_{L+1} \delta(k_1 - k_2). \quad (D.9)$$

**Case - $k_1 = k_2$** In order to understand the conditions that need to be imposed on the structure of $\tilde{\mathbf{T}}_k$, we examine an arbitrary element of $\mathbf{R}_{k,k}$. From (D.9), we notice that

$$\begin{aligned}[\mathbf{R}_{k,k}]_{l_1,l_2} &= \sum_{\nu=0}^{V-1} |\tilde{\mathbf{T}}_k[\nu, \nu]|^2 \exp\{-j2\pi l_\nu (l_2 - l_1)/P\} \\ &= \frac{\mathcal{P}_t}{K} \delta(l_1 - l_2). \end{aligned} \qquad (D.10)$$

It can be verified that the above expression is true under three conditions:

(C6) $P = VS$ where $S \in \mathbb{Z}^+$.
(C7) $\{l_\nu\}$, the index set of subcarriers that carry pilot symbols are chosen such that $l_\nu = l_s + \nu S$ for $l_s \in [0, S-1]$ and $0 \le \nu \le V - 1$.
(C8) The pilot tones are all equally powered so that $|\tilde{\mathbf{T}}_k[\nu, \nu]|^2 = \frac{\mathcal{P}_t}{KV}$.

The above conditions indicate that the pilot symbols used for channel estimation must be *equispaced* in the subcarrier domain and *equipowered*. Due to (C8), we see that $\tilde{\mathbf{T}}_k^* \tilde{\mathbf{T}}_k = \frac{\mathcal{P}_t}{KV} \mathbf{I}_V$. Combined with the fact that $\hat{\mathcal{W}}_{0:L}^{\mathsf{H}} \hat{\mathcal{W}}_{0:L} = V \mathbf{I}_{L+1}$, we see that $\mathbf{R}_{k,k} = \frac{\mathcal{P}_t}{K} \mathbf{I}_{L+1}$. We now see that when $K = 1$, the following pilot sequence design:

$$\begin{aligned} \tilde{\mathbf{T}}_k &= \boldsymbol{\phi}(V) \\ &= \sqrt{\frac{\mathcal{P}_t}{KV}} [ \exp(j\phi_0) \quad \exp(j\phi_1) \ \dots \ \exp(j\phi_{V-1})]^{\mathsf{T}}, \end{aligned} \qquad (D.11)$$

meets conditions (C6), (C7), and (C8) that mandate the usage of equipowered, equispaced pilots at each transmit antenna. In (D.11), $\{\phi_\nu\}$ are any arbitrary values in $[-\pi, \pi]$.

**Case - $k_1 \ne k_2$** We now incorporate the consequences of imposing the condition, $\mathbf{R}_{k_1,k_2} = 0$ when $k_1 \ne k_2$, in (D.11). We again utilize (D.9) and apply (C7). We see that

$$\begin{aligned}[\mathbf{R}_{k_1,k_2}]_{l_1,l_2} &= \exp\{-j2\pi l_s (l_1 - l_2)/P\} \\ &\times \sum_{\nu=0}^{V-1} \tilde{\mathbf{T}}_{k_1}^* [\nu, \nu] \tilde{\mathbf{T}}_{k_2} [\nu, \nu] \\ &\times \exp\{-j2\pi \nu S(l_1 - l_2)/P\}, \end{aligned} \qquad (D.12)$$

which equals zero when

$$\sum_{\nu=0}^{V-1} \tilde{\mathbf{T}}_{k_1}^* [\nu, \nu] \tilde{\mathbf{T}}_{k_2} [\nu, \nu] \exp\{-j2\pi \nu S(l_1 - l_2)/P\} = 0$$

$$\Leftrightarrow \sum_{\nu=0}^{V-1} \tilde{\mathbf{T}}_{k_1}^* [\nu, \nu] \tilde{\mathbf{T}}_{k_2} [\nu, \nu] \exp\{-j2\pi \nu(l_1 - l_2)/V\} = 0, \qquad (D.13)$$

$\forall 1 \le k_1, k_2 \le K$ and $\forall (l_1 - l_2) \in [0, \pm 1 \ \dots \ \pm L]$. This condition clearly suggests that in the frequency-domain, phase-shift orthogonality is required between the pilot sequences of different transmit antennas over the range, $(l_1 - l_2) \in [0, \pm 1 \ \dots \ \pm L]$. Equivalently, this translates to circular shift orthogonality in the time-domain. We now define a diagonal $V \times V$ dimensional phase-shifting matrix, $\mathcal{E}_k$ as shown below:

$$[\mathcal{E}_k]_{\nu,\nu} = \exp\left[\frac{j2\pi \nu f_k}{V}\right], \qquad (D.14)$$

$\forall \ 0 \le \nu \le V - 1, \quad 1 \le k \le K$ and design the pilot sequence such that

$$\tilde{\mathbf{T}}_k = \mathcal{E}_k \boldsymbol{\phi}(V). \qquad (D.15)$$

As a result of substituting (D.14) in (D.13), we see that

$$\sum_{\nu=0}^{V-1} \exp\{-j2\pi \nu(f_{k_1} - f_{k_2} + l_1 - l_2)/V\} = 0. \quad (D.16)$$

We selected $\mathcal{E}_k$ as shown in (D.14) so that we can exploit the property of summation of the roots of unity. In order to do so, we require that the term, $(f_{k_1} - f_{k_2} + l_1 - l_2)$ be a non integer-multiple of $V$. So, we choose $f_k = k(V-L-1)$. In conclusion, the training design shown in (32) meets not only conditions (C6), (C7), and (C8) but also (D.13) so that phase-shift orthogonality is maintained between the pilot sequences of any pair of transmit antennas.

### 1.5 Proof of Theorem 4

By denoting the entropy using $\mathsf{H}(.)$ and applying the definition of mutual information, we can write the following expression,

$$\mathsf{I}(\tilde{\mathbf{y}}_n^{(dt)}; \tilde{\mathbf{x}}_n | \tilde{\mathcal{H}}_n) = \mathsf{H}(\tilde{\mathbf{x}}_n | \tilde{\mathcal{H}}_n) - \mathsf{H}(\tilde{\mathbf{x}}_n | \tilde{\mathcal{H}}_n, \tilde{\mathbf{y}}_n^{(dt)}), \quad (E.1)$$

for the system model in the training phase. In the above equation, $\mathsf{H}(\tilde{\mathbf{x}}_n|\widetilde{\mathcal{H}}_n)$ is maximized when $\tilde{\mathbf{x}}_n$ is Gaussian. Hence, with $\mathbf{R}_{\tilde{\mathbf{x}}_n} = \mathcal{P}_{dt}\,\mathbf{I}_{KM}$,

$$\mathsf{H}(\tilde{\mathbf{x}}_n|\widetilde{\mathcal{H}}_n) = \log\det(\pi\,e\,\mathcal{P}_{dt}\,\mathbf{I}_{KM}). \qquad (E.2)$$

On the other hand,

$$\mathsf{H}(\tilde{\mathbf{x}}_n|\widetilde{\mathcal{H}}_n,\,\tilde{\mathbf{y}}_n^{(dt)}) = \log\det(\pi\,e\,\mathbf{R}_{\tilde{\mathbf{x}}_n|\tilde{\mathbf{y}}_n^{(dt)},\widetilde{\mathcal{H}}_n}), \quad (E.3)$$

where $\mathbf{R}_{\tilde{\mathbf{x}}_n|\tilde{\mathbf{y}}_n^{(dt)},\widetilde{\mathcal{H}}_n} = ((1/\mathcal{P}_{dt})\,\mathbf{I}_{KM} + (1/\sigma_z^2)\,\widetilde{\mathcal{H}}_n^{\mathsf{H}}\,\widetilde{\mathcal{H}}_n)^{-1}$. In writing the covariance matrix of the random vector, $\tilde{\mathbf{x}}_n$ conditioned on $\tilde{\mathbf{y}}_n^{(dt)}$ and $\widetilde{\mathcal{H}}_n$, we utilize the fact that $\tilde{\mathbf{z}}_n$ is Gaussian distributed. Consequently,

$$\mathsf{I}(\tilde{\mathbf{y}}_n^{(dt)};\tilde{\mathbf{x}}_n|\widetilde{\mathcal{H}}_n) = \log\det\left(\mathbf{I}_{KM} + \frac{\mathcal{P}_{dt}}{\sigma_z^2}\,\widetilde{\mathcal{H}}_n\,\widetilde{\mathcal{H}}_n^{\mathsf{H}}\right). \quad (E.4)$$

By a similar approach, we can show that

$$\mathsf{I}(\tilde{\mathbf{r}}_n;\tilde{\mathbf{s}}_n|\overline{\mathcal{H}}_n) = \log\det\left(\mathbf{I}_{KP} + \frac{\mathcal{P}_d}{\sigma_z^2}\,\overline{\mathcal{H}}_n\,\overline{\mathcal{H}}_n^{\mathsf{H}}\right), \quad (E.5)$$

in the data transmission phase. Due to the fact that the mutual information between the transmitted and estimated data vectors is independent of the block index, we represent the channel capacity upper bound by choosing the mutual information with respect to an arbitrary block indices as shown in (41). It can also be observed that we have included the appropriate normalization factor since only $M$ out of $P$ subcarriers in each OFDM symbol in the training phase, carry information symbols.

### 1.6 Proof of Theorem 5

We again apply the definition of mutual information and write the expression,

$$\mathsf{I}(\tilde{\mathbf{y}}_n^{(dt)};\tilde{\mathbf{x}}_n|\widehat{\widetilde{\mathcal{H}}}_n) = \mathsf{H}(\tilde{\mathbf{x}}_n|\widehat{\widetilde{\mathcal{H}}}_n) - \mathsf{H}(\tilde{\mathbf{x}}_n|\widehat{\widetilde{\mathcal{H}}}_n,\tilde{\mathbf{y}}_n^{(dt)}), \quad (F.1)$$

for the system model in the training phase. Similar to (E.2), we see that

$$\mathsf{H}(\tilde{\mathbf{x}}_n|\widehat{\widetilde{\mathcal{H}}}_n) = \log\det(\pi\,e\,\mathcal{P}_{dt}\,\mathbf{I}_{KM}), \qquad (F.2)$$

whereas,

$$\mathsf{H}(\tilde{\mathbf{x}}_n|\widehat{\widetilde{\mathcal{H}}}_n,\tilde{\mathbf{y}}_n^{(dt)}) \leq \log\det(\pi\,e\,\mathbf{R}_{\tilde{\mathbf{x}}_n|\tilde{\mathbf{y}}_n^{(dt)},\widehat{\widetilde{\mathcal{H}}}_n}), \quad (F.3)$$

with equality if and only if $\tilde{\mathbf{x}}_n$ given $\tilde{\mathbf{y}}_n^{(dt)}$ and $\widehat{\widetilde{\mathcal{H}}}_n$ is drawn from a Gaussian distribution with the covariance matrix, $\mathbf{R}_{\tilde{\mathbf{x}}_n|\tilde{\mathbf{y}}_n^{(dt)},\widehat{\widetilde{\mathcal{H}}}_n}$. We can now formulate an expression for $\mathbf{R}_{\tilde{\mathbf{x}}_n|\tilde{\mathbf{y}}_n^{(dt)},\widehat{\widetilde{\mathcal{H}}}_n}$ by assuming that LMMSE estimator[b] has been used to estimate $\tilde{\mathbf{x}}_n$:

$$\mathbf{R}_{\tilde{\mathbf{x}}_n|\tilde{\mathbf{y}}_n^{(dt)},\widehat{\widetilde{\mathcal{H}}}_n} = \mathbf{R}_{\tilde{\mathbf{x}}_n} - \mathbf{R}_{\tilde{\mathbf{x}}_n,\tilde{\mathbf{y}}_n^{(dt)}}\,\mathbf{R}_{\tilde{\mathbf{y}}_n^{(dt)}}\,\mathbf{R}_{\tilde{\mathbf{y}}_n^{(dt)},\tilde{\mathbf{x}}_n}, \quad (F.4)$$

where

$$\mathbf{R}_{\tilde{\mathbf{x}}_n,\tilde{\mathbf{y}}_n^{(dt)}} \triangleq \mathbb{E}\{\tilde{\mathbf{x}}_n\,\tilde{\mathbf{y}}_n^{(dt)\mathsf{H}}|\widehat{\widetilde{\mathcal{H}}}_n\} = \mathcal{P}_{dt}\,\widehat{\widetilde{\mathcal{H}}}_n^{\mathsf{H}}, \quad (F.5a)$$

$$\begin{aligned}\mathbf{R}_{\tilde{\mathbf{y}}_n^{(dt)}} &\triangleq \mathbb{E}\{\tilde{\mathbf{y}}_n^{(dt)}\,\tilde{\mathbf{y}}_n^{(dt)\mathsf{H}}|\widehat{\widetilde{\mathcal{H}}}_n\}\\ &= \mathcal{P}_{dt}\,\widehat{\widetilde{\mathcal{H}}}_n\,\widehat{\widetilde{\mathcal{H}}}_n^{\mathsf{H}} + \sigma_{z^{(dt)}}^2(n)\,\mathbf{I}_{KM}. \quad (F.5b)\end{aligned}$$

In (F.5a) and (F.5b), we have utilized the orthogonality property of LMMSE estimation and eliminated the covariance terms between the information symbol vector, $\tilde{\mathbf{x}}_n$ and the effective noise vector, $\tilde{\mathbf{z}}_n^{(dt)}$. By substituting (F.5a) and (F.5b) in (F.4), we see that

$$\begin{aligned}\mathbf{R}_{\tilde{\mathbf{x}}_n|\tilde{\mathbf{y}}_n^{(dt)},\widehat{\widetilde{\mathcal{H}}}_n} &= \mathcal{P}_{dt}\,\mathbf{I}_{KM} - \mathcal{P}_{dt}^2\,\widehat{\widetilde{\mathcal{H}}}_n^{\mathsf{H}}\\ &\quad\times(\mathcal{P}_{dt}\,\widehat{\widetilde{\mathcal{H}}}_n\,\widehat{\widetilde{\mathcal{H}}}_n^{\mathsf{H}} + \sigma_{z^{(dt)}}^2(n)\,\mathbf{I}_{KM})^{-1}\,\widehat{\widetilde{\mathcal{H}}}_n\\ &= ((1/\mathcal{P}_{dt})\,\mathbf{I}_{KM} + \sigma_{z^{(dt)}}^2(n)\,\widehat{\widetilde{\mathcal{H}}}_n^{\mathsf{H}}\,\widehat{\widetilde{\mathcal{H}}}_n)^{-1}. \quad (F.6)\end{aligned}$$

Finally, from (F.2), (F.3) and (F.6), we have,

$$\mathsf{I}(\tilde{\mathbf{y}}_n^{(dt)};\tilde{\mathbf{x}}_n|\widehat{\widetilde{\mathcal{H}}}_n) \geq \log\det(\mathbf{I}_{KM} + \mathcal{P}_{dt}\,\sigma_{z^{(dt)}}^2(n)\,\widehat{\widetilde{\mathcal{H}}}_n^{\mathsf{H}}\,\widehat{\widetilde{\mathcal{H}}}_n). \quad (F.7)$$

By a similar approach, we can show that

$$\mathsf{I}(\tilde{\mathbf{r}}_n;\tilde{\mathbf{s}}_n|\widehat{\overline{\mathcal{H}}}_n) \geq \log\det(\mathbf{I}_{KP} + \mathcal{P}_d\,\sigma_{z^{(d)}}^2(n)\,\widehat{\overline{\mathcal{H}}}_n^{\mathsf{H}}\,\widehat{\overline{\mathcal{H}}}_n), \quad (F.8)$$

in the data transmission phase. Therefore, the channel capacity lower bound is obtained by combining (F.7) and (F.8) as shown in (49).

## 2 Endnotes

[a]$(\mathbf{A} + \mathbf{B}\,\mathbf{C}\,\mathbf{D})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\,\mathbf{A}^{-1}\,\mathbf{B})^{-1}\mathbf{D}\,\mathbf{A}^{-1}$.
[b]Notice that when $\tilde{\mathbf{z}}_n^{(dt)}$ is Gaussian, the LMMSE estimator coincides with an MMSE estimator.

**Author details**
[1]Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, Worcester, MA 01609, USA. [2]Department of Electrical Engineering, United States Naval Academy, Annapolis, MD 21402, USA.

**References**
1. GJ Foschini, Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. Bell Labs Tech. J. **1**(2), 41–59 (1996)
2. N Merhav, G Kaplan, A Lapidoth, S Shamai Shitz, On information rates for mismatched decoders. IEEE Trans. Inf. Theory. **40**, 1953–1967 (1994)
3. A Lapidoth, P Narayan, Reliable communication under channel uncertainty. IEEE Trans. Inf. Theory. **44**, 2148–2177 (1998)

4. M Medard, The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel. IEEE Trans. Inf. Theory. **46**, 933–946 (2000)

5. A Lapidoth, S Shamai, Fading channels: how perfect need "perfect side information" be? Theory. **48**, 1118–1134 (2002)

6. J Yang, S Roy, On joint transmitter and receiver optimization for multiple-input-multiple-output (MIMO) transmission systems. IEEE Trans. Commun. **42**, 3221–3231 (1994)

7. A Scaglione, G Giannakis, S Barbarossa, Redundant filterbank precoders and equalizers. Parts I & II. IEEE Trans. Signal Process. **47**, 1988–2006 (1999)

8. D Palomar, M Lagunas, J Cioffi, Optimum linear joint transmit-receive processing for MIMO channels with QoS constraints. IEEE Trans. Signal Process. **52**, 1179–1197 (2004)

9. L Tong, B Sadler, M Dong, Pilot-assisted wireless transmissions: general model, design criteria, and signal processing. IEEE Signal Process Mag. **21**, 12–25 (2004)

10. R Negi, J Cioffi, Pilot tone selection for channel estimation in a mobile OFDM system. IEEE Trans. Consum. Electron. **44**, 1122–1128 (1998)

11. S Adireddy, L Tong, H Viswanathan, Optimal placement of training for frequency-selective block-fading channels. IEEE Trans. Inf. Theory. **48**, 2338–2353 (2002)

12. M Dong, L Tong, Optimal design and placement of pilot symbols for channel estimation. IEEE TransSignal Process. **50**, 3055–3069 (2002)

13. S Ohno, G Giannakis, Optimal training and redundant precoding for block transmissions with application to wireless OFDM. IEEE Trans. Commun. **50**, 2113–2123 (2002)

14. J Manton, I Mareels, Y Hua, in *Proceedings of the 25th IEEE International COnference on Acoustics Speech, and Signal Processing, 2000 (ICASSP '00)*, vol. 5. Affine precoders for reliable communications, (2000), pp. 2749–2752

15. S Ohno, G Giannakis, Capacity maximizing mmse-optimal pilots for wireless OFDM over frequency-selective block Rayleigh-fading channels. IEEE Trans. Inf Theory. **50**, 2138–2145 (2004)

16. X Ma, L Yang, G Giannakis, Optimal training for MIMO frequency-selective fading channels. IEEE Trans. Wirel. Commun. **4**, 453–466 (2005)

17. J Baltersee, G Fock, H Meyr, Achievable rate of MIMO channels with data-aided channel estimation and perfect interleaving. IEEE J. Sel. Areas Commun. **19**, 2358–2368 (2001)

18. B Hassibi, B Hochwald, How much training is needed in multiple-antenna wireless links? IEEE Trans. Inf. Theory. **49**, 951–963 (2003)

19. D Samardzija, N Mandayam, Pilot-assisted estimation of MIMO fading channel response and achievable data rates. IEEE Trans. Signal Process. **51**, 2882–2890 (2003)

20. A Vosoughi, A Scaglione, On the effect of receiver estimation error upon channel mutual information. IEEE Trans. Signal Process. **54**, 459–472 (2006)

21. M Coldrey, P Bohlin, Training-based MIMO systems–Part, I: performance comparison. IEEE Trans. Signal Process. **55**, 5464–5476 (2007)

22. M Ding, S Blostein, Maximum mutual information design for MIMO systems with imperfect channel knowledge. IEEE Trans. Inf. Theory. **56**, 4793–4801 (2010)

23. A Vosoughi, A Scaglione, Everything you always wanted to know about training: guidelines derived using the afine precoding framework and the CRB. IEEE Trans. Signal Process. **54**, 940–954 (2006)

24. I Barhumi, G Leus, M Moonen, Optimal training design for MIMO OFDM systems in mobile wireless channels. IEEE Trans. Signal Process. **51**, 1615–1624 (2003)

25. M Dong, L Tong, B Sadler, in *IEEE Military Communications Conference 2002*, vol. 1. Optimal pilot placement for channel tracking in OFDM, (2002), pp. 602–606

26. G Giannakis, Filterbanks for blind channel identification and equalization. IEEE Signal Process. Lett. **4**, 184–187 (1997)

27. M Stojanovic, J Proakis, J Catipovic, Analysis of the impact of channel estimation errors on the performance of a decision-feedback equalizer in fading multipath channels. IEEE Trans. Commun. **43**, 877–886 (1995)

28. C Komninakis, C Fragouli, A Sayed, R Wesel, Multi-input multi-output fading channel tracking and equalization using Kalman estimation. IEEE Trans. Signal Process. **50**, 1065–1076 (2002)

29. IE Telatar, Capacity of multi-antenna gaussian channels. Eur. Trans. Telecommun. **10**, 585–595 (1999)

30. H Schulze, C Lueders, *Theory and Applications of OFDM and CDMA: Wideband Wireless Communications*. (John Hoboken, 2005)

31. SM Kay, *Fundamentals of Statistical Signal Processing, Vol. I-Estimation Theory*. (Prentice Hall, Englewood Cliffs, NJ, 1993)

32. D Vaughan, A nonrecursive algebraic solution for the discrete Riccati equation. IEEE Trans. Autom. Control. **15**, 597–599 (1970)

33. A Laub, A Schur method for solving algebraic Riccati equations. IEEE Trans. Autom. Control. **24**, 913–921 (1979)

34. R Leland, An alternate calculation of the discrete-time Kalman filter gain and Riccati equation solution. IEEE Trans. Autom. Control. **41**, 1817–1819 (1996)

35. H Minn, N Al-Dhahir, Optimal training signals for MIMO OFDM channel estimation. IEEE Trans. Wirel. Commun. **5**, 1158–1168 (2006)

36. M Dong, L Tong, B Sadler, Optimal insertion of pilot symbols for transmissions over time-varying flat fading channels. IEEE Trans. Signal Process. **52**, 1403–1418 (2004)

37. S Kay, *Modern Spectral Estimation: Theory and Application*. (Prentice-Hall, Upper Saddle River, 1999)

38. P Stoica, RL Moses, *Spectral Analysis of Signals*. (Prentice-Hall, Upper Saddle River, 2005)

39. HL Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. (Wiley-Interscience, USA, 2001)

40. A Hjørungnes, D Gesbert, Complex-valued matrix differentiation: techniques and key results. IEEE Trans. Signal Process. **55**, 2740–2746 (2007)

41. A Hjørungnes, D Gesbert, D Palomar, in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3. Unified theory of complex-valued matrix differentiation, (2007), pp. 345–348