IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Gaussian Codes and the Scaled Nearest Neighbor Decoding in Fading Multi-Antenna Channels, Part 2: Optimal Training Sequences for the Piece-Wise Constant Channel

Hanan Weingarten, Yossef Steinberg, and Shlomo Shamai (Shitz)

CCIT Report #369
January 2002

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

המרכז לטכנולוגיות תקשורת ומידע
הפקולטה להנדסת חשמל
הטכניון - מכון טכנולוגי לישראל, חיפה 32000, ישראל

Gaussian Codes and the Scaled Nearest Neighbor Decoding in Fading Multi-Antenna Channels, Part 2: Optimal training Sequences for the Piece-Wise Constant Channel

Hanan Weingarten, Yossef Steinberg, and Shlomo Shamai (Shitz)

Abstract

We investigate training schemes that are used for block fading multi-antenna channels. The Generalized Mutual Information is used as a criterion for optimizing the training sequence and information symbol structure. We give new results concerning the training sequences and optimal antenna usage for the block fading channel. We show that the number of antennas which should be used for training and information sequences depends on the SNR and the energy allotted for training. We also show that an optimal training scheme is one where a "pilot" is sent through each of the antennas, one antenna at a time.

Index Terms

Block fading, Fading channels, Gaussian codebooks, Multi-antenna, Generalized Mutual Information, Piece-wise constant channels, Scaled nearest neighbor decoder, Training

---

This research (no. 050-054) was supported by THE ISRAEL SCIENCE FOUNDATION.

H. Weingarten is with the Department of Electrical Engineering, Technion-IIT, Haifa, Israel 32000. Phone: +972 4 8294704, e-mail: whanan@tx.technion.ac.il

Y. Steinberg is with the Department of Electrical Engineering, Technion-IIT, Haifa, Israel 32000. Phone: +972 4 8294656, fax: +972 4 8323041, e-mail: ysteinbe@ee.technion.ac.il

S. Shamai is with the Department of Electrical Engineering, Technion-IIT, Haifa, Israel 32000. Phone: +972 4 8294713, fax: +972 4 8323041, e-mail: sshlomo@ee.technion.ac.il

I. Introduction

Block fading (piece-wise constant) channels are usually used to emulate slowly varying fading channels. The fading is assumed to be constant during a time period equal to the channel coherence length and changing between coherence periods. The coherence period is referred to as the block length. Frequently, when the receiver has no prior knowledge of the fading in the channel, part of the block is used to transmit training sequences which are used by the receiver to gain an estimate of the fading. The rest of the block is used to transmit information sequences. The estimate is then used by the decoder in the decoding process as if it was a perfect channel estimation. Such decoders which use an externally supplied channel estimations will be referred to as coherent decoders as opposed to non-coherent decoders which decode the data without any channel side information(CSI).

Justification for the use of training sequences in multi-antenna Gaussian channels is given in [8] and here as well for the case where the channel coherence period is long compared to the number of transmit antennas. Moreover, in [12], [13], the authors show that for high SNRs, the rate achieved by the training based scheme comes within a constant of the channel's capacity and that constant does not depend on the SNR.

The accuracy of the channel estimation affects the decoding process and the achievable information rates. Even in single antenna channels, it may be hard to obtain good channel fading estimations as the channel may change too rapidly. The question of estimation accuracy is even more acute in multi-antenna channels, where longer training sequences are needed to enable the estimation of a large number of fading receiver-transmitter paths. In this paper, we investigate the training sequences which optimize the achievable rates in block fading multi-antenna channels.

It should be noted that there are some limitations on the effectiveness of the block

fading model. The results in [12], [13] show that for block fading channels and high SNRs, the capacity grows as log(SNR). However, Lapidoth and Moser [6] show that as long as the channel fading is not perfectly constant during the coherence interval, the capacity at very high SNRs grows only as log(log(SNR)). As the block fading model is only an approximation of the fading channel, the true growth rate of the channel capacity is the one stated by Lapidoth and Moser. Consequently, during our discussion we will assume that our channel estimation error is much higher than the error incurred by the channel's variations during the coherence interval, such that the block fading model still gives good indication of how the "real" channel behaves.

Previous investigations of the training sequences using an information theoretic approach were done for the case of the multi-antenna block fading channel [4] and the case of the slowly varying frequency selective channel [1], [9]. The investigation in [4] (and for a large extent in [1] and [9] as well) is based on a lower bound on the channel capacity. The channel estimate is obtained using the MMSE estimator and the lower bound on the capacity is maximized by optimizing the length of the training sequence and the allocation of power to the training and information sequences.

A different approach was examined in [2], [3] where training based communication schemes were examined for the case where the maximum likelihood (ML) estimator is used to estimate the fading of a multi-antenna channel. The criteria for building good training sequences was to minimize the overall error of the ML channel estimator. In [3] it was shown that when the nearest neighbor decoder is used along with the ML channel estimator, minimizing the overall estimation error, minimizes the decoding metric of the correct codeword. It should be noted that this only indicates the benefit of minimizing the overall estimation error when the ML estimator is used together with the nearest neighbor decoder but does

not indicate that using the ML estimator is beneficial.

Our approach to the problem is different and is based on previous results that were presented in the first part of our work [11]. We will investigate the performance of a given communication system which uses a random Gaussian coding scheme and a modified scaled nearest neighbor decoder. Our goal is to optimize the training sequences and the Gaussian code in order to maximize the achievable rates under this scheme. This approach was previously investigated in [7] for the single antenna case and gives a good indication of how, many of the familiar coding techniques, will behave under imperfect channel estimation. The modification made to the scaled nearest neighbor decoder serves to improve the performance of the familiar decoder. Moreover, we will rely on the optimality of the MMSE channel estimator, as was proved in [11], and our investigation will be limited only to the choice of that estimator.

The optimal training sequence found in [4], was optimized for the special case where the transmit power is equally distributed between all transmit antennas and for this case, the achievable rate in our scheme and the lower bound on the capacity in [4], coincide. However, we will investigate the general case where the only limitation on the distribution of the Gaussian codebook is its overall transmit power, and the optimization is done over the choice of the training and information sequences such that the achievable rate is maximized. We will present somewhat surprising results regarding the optimal training sequence and transmit antenna usage. We will not repeat the optimization of the allocation of power to training and information sequences. The interested reader is referred to [4].

A. Channel Model

In our model, the communication system contains t transmit and r receive antennas. We assume that the channel between the transmitters and receivers is a block fading channel

and that the fading coherence duration is T seconds. A new symbol is sent through the channel every T_s seconds. Consequently, we shall assume that the following expression can be used to describe our channel:

$$Y_k = H_{\lceil kT_s/T \rceil} X_k + Z_k \tag{1}$$

Where

- k is the symbol's index.
- X_k denotes the k'th input symbol. This is a column vector which has t variables.
- Z_k denotes an additive noise vector. This is a column vector with r elements. We assume that Z_k is a circularly symmetric complex Gaussian vector with IID elements in space and time. All elements in Z_k have zero mean and variance N (i.e. $E[Z_k Z_k^\dagger] = NI_r$).
- $H_{\lceil kT_s/T \rceil}$ is the channel fading matrix of block $\#\lceil kT_s/T \rceil$. This is an $r \times t$ complex matrix where element $(H_{\lceil kT_s/T \rceil})_{ij}$ is the fading between transmitter j and receiver i. The elements of $H_{\lceil kT_s/T \rceil}$ are IID circularly symmetric complex Gaussian random variables with zero mean and variance σ_H^2. We assume that the noise process, fading process and the input message are independent. Furthermore, we assume that the block fading matrix, H_i, is an ergodic process (where i is the block number).
- Y_k is the channel output at time index k. This is a column vector with r elements.

We assume that a random coding scheme is used to encode data and that the code symbols are generated by a circularly symmetric Gaussian random source. The code-book is revealed both to encoder and decoder. Each code-word contains n symbols where n is assumed to be arbitrarily large in order to decrease the probability of decoding error. Each symbol contains t elements and is statistically independent of symbols generated for other codewords or for other time-slots. We will assume that the random source has a covariance

matrix, $E[X_k X_k^\dagger] = Q$, and the average energy per information symbol is given by:

$$tr\{Q\} = T_s P_d \ , \tag{2}$$

where P_d is the average transmit power of the information sequences.

As mentioned, part of the coherence interval is used to transmit a training sequence. We use T_τ to denote the time allotted for training and $T_d = T - T_\tau$ to denote the time allotted for information transmission (see Figure 1). We shall assume that the number of symbols transmitted during the information sequence period, T_d, and during the training sequence period, T_τ, is an integer and will be denoted by n_d and n_τ respectively, such that,

$$\begin{aligned} n_d &= T_d/T_s \\ n_\tau &= T_\tau/T_s \end{aligned} \tag{3}$$

The overall number of symbols in a coherence interval will be denoted by n_T and we can write,

$$n_T = n_d + n_\tau = T/T_s. \tag{4}$$

We will focus on the case where the transmit power during training sequences differs from the average transmit power of the information sequences and we will use P_τ to denote the transmit power of the training sequences. In addition, X_τ and Y_τ will be used to denote the training sequence and the channel output during the transmission of the training sequence. We assume that the same training sequence, X_τ, is used for all blocks. As a result of our channel model we may write:

$$Y_\tau = H_1 X_\tau + Z_1^{n_\tau} \ ,$$

where X_τ and Y_τ are $t \times n_\tau$ and $r \times n_\tau$ matrices respectively. The channel output, Y_τ, is used by the receiver to calculate the channel fading estimate, $\widehat{H}_{\lceil kT_s/T \rceil}$, where $\widehat{H}_{\lceil kT_s/T \rceil}$ is the channel estimate for block number $\lceil kT_s/T \rceil$.

The overall average energy used during a coherence interval is fixed and is given by:

$$T_\tau P_\tau + T_d P_d = TP \ , \tag{5}$$

where P is the average transmit power and is usually fixed by design requirements. The SNR in the channel can now be defined by:

$$SNR = \frac{\sigma_H^2 P}{N} \tag{6}$$

The decoder is a modified version of the scaled nearest neighbor decoder. This decoder was introduced and studied in [11]. We use the familiar nearest neighbor decoder and add a scaling factor which is a function of the fading estimation, $\widehat{H}_{\lceil kT_s/T \rceil}$. The scaling improves the performance of the decoder by attenuating the metrics of "noisy" blocks and amplifying the metrics of good blocks. The decoding procedure for codewords of length n is described by the following equation:

$$\hat{m} = \arg\min_m \left(\frac{1}{n} \sum_{k=1}^{n} \| \widetilde{K}(\widehat{H}_{\lceil kT_s/T \rceil})(Y_k - \widehat{H}_{\lceil kT_s/T \rceil} \cdot X(m)_k) \|^2 \right) \tag{7}$$

Where

- $\| \cdot \|^2$ denotes the squared norm of a vector.
- $\hat{m}$ is the decoded message ID.
- $\widehat{H}_{\lceil kT_s/T \rceil}$ is the channel fading estimation of block $\#\lceil kT_s/T \rceil$. It is an r by t $(r \times t)$ complex matrix.
- $X_k(m)$ is the vector of code-word m at time slot k.
- n is the codeword length, which may span many coherence intervals, T.
- and $\widetilde{K}(\widehat{H}_{\lceil kT_s/T \rceil})$ is the scaling factor. This is a square, $r \times r$, matrix which we assume to be a function of the fading estimation, and thus is constant within each coherence interval, T.

We declare that an error has occurred when $\hat{m} \neq m$ or when there is more than one message for which the scaled distance is the minimum distance.

During the rest of this paper we will omit the block number, $\lceil kT_s/T \rceil$, from the channel fading and channel estimation matrices for simplicity of presentation, but we will continue to assume that new fading matrices and channel estimations are generated for each block. In addition, we will use a short hand notation for the function of the scaling factor and write $\widetilde{K}$ instead of $\widetilde{K}(\widehat{H}_{\lceil kT_s/T \rceil})$.

II. Achievable Rates - Generalized Mutual Information (Previous Results)

The achievable rates of the communication scheme described above are also referred to as the Generalized Mutual Information, I_{GMI}. In [11] we have calculated I_{GMI} for a more general setting, where the channel is a general additive noise, fading multi-antenna channel, and the fading estimation is based on some side information which is not strictly specified.

The assumptions we made in [11] are fulfilled by the communication system described above. This results from the fact that H, $\widehat{H}$, X_k and Z_k meet the ergodic requirements in [11] and the fading estimation is a strictly causal function of the channel output. Therefore, we can use the results given in [11] in our discussion here.

The main result in [11] states that the achievable rate, I_{GMI}, is given by the following expression:

$$I_{GMI} = E\Big[\log\det\big(I_\tau + P_{\widetilde{N}}^{-1}\widehat{H}Q\widehat{H}^\dagger\big)\Big] \tag{8}$$

where

$$P_{\widetilde{N}} = P_{\widetilde{N}}(\widehat{H}) = E\Big[\widetilde{H}Q\widetilde{H}^\dagger + Z_k Z_k^\dagger \Big| \widehat{H}\Big] \tag{9}$$

and where $\widetilde{H} = H - \widehat{H}$ is the channel estimation error and the expectation in (8) is over the distribution of $\widehat{H}$. It can be seen that I_{GMI} is equal to the capacity of a coherent fading

channel with additive white (in time) Gaussian noise where the fading is equal to $\widehat{H}$ and is known at the receiver, and the noise power, $P_{\widetilde{N}}$, is a function of the channel fading, $\widehat{H}$, and is given in (9). The decoder must use the optimal scaling factor, $\widetilde{K}$, to achieve this rate. It is shown in [11] that the optimal choice of $\widetilde{K}$ is

$$\widetilde{K}(\widehat{H}) = (P_{\widetilde{N}})^{-\frac{1}{2}} \tag{10}$$

Note that the rate in (8), is given in nats per symbol.

The matrix $P_{\widetilde{N}}$ can be interpreted as the "overall noise" which is caused by both the channel noise and the channel estimation error. Consequently, we can think of the role of the scaling factor as an equalizer, which equalizes the "overall noise" over all space and time dimensions.

An important result in [11] states that the optimal choice of the channel estimation is the MMSE estimator. That is, to optimize the achievable rate, we must choose: $\widehat{H} = E[H|SI]$ (where SI is the side information). Therefore, we will limit our discussion to this type of estimator. In our case, the side information is the training sequence, X_τ, and the channel output during training, Y_τ. Our model is a Gaussian model and as a result, $\widehat{H}$ is actually a linear function of Y_τ and is given by:

$$\widehat{H} = E[H|X_\tau, Y_\tau] = \sigma_H^2 Y_\tau X_\tau^\dagger (N I_t + \sigma_H^2 X_\tau X_\tau^\dagger)^{-1} \tag{11}$$

A consequence of the MMSE estimator in a Gaussian regime is the fact that the estimation error is independent of $\widehat{H}$ (actually independent of any function of Y_τ). Furthermore, the additive noise, at time slots used for information transmission, is independent of the fading estimation due to the fact that the noise is white. Therefore, $P_{\widetilde{N}}$, in our case, is actually constant and independent of the fading estimation and we may rewrite expression (9) as follows:

$$P_{\widetilde{N}} = E\left[\widetilde{H} Q \widetilde{H}^\dagger\right] + N I_r \tag{12}$$

Clearly, the scaling factor, K, is also constant and has the same value for all blocks. The estimation error depends on the choice of the training sequence and therefore, the value of the scaling factor is determined by the choice of the training sequence and matrix Q.

III. Optimal Training Sequences

In the work by Hassibi and Hochwald [4], the transmit power was assumed to be equally distributed between the transmit antennas such that, $Q = \frac{T_s P_d}{t} I_t$. This power distribution was shown to be optimal [10] for a channel model similar to the one here but where the receiver has perfect channel state information (i.e. there is no estimation error). However, when the channel estimation is no longer perfect, the transmit power distribution used in [4] is no longer optimal.

In our investigation we will show that the choices of the optimal training sequence and structure of the input signal covariance matrix, Q, depend on the SNR. More to the point, we will show that the number of useful antennas depends (amongst other things) on the SNR. This indicates that knowledge of the SNR at the transmitter is crucial for good communications as it is in any communication system, otherwise the transmitter can not know at what rate to transmit. Actual implications of SNR on the structure of the optimal training sequence will also be discussed for specific cases.

An important feature of the fading matrix, in our channel model, is its isotropic distribution. The distribution of the fading is invariant to any right or left multiplication by a unitary matrix. This attribute is a result of the Gaussian distribution and independence of the elements of the fading matrix [10]. Consequently, the derivation here will be specific to this problem and does not directly apply to the selective fading problem [1], [9].

Before stating a theorem on the optimal achievable rate, we shall rewrite X_τ and Q as

follows:

$$X_\tau = U_\tau \Lambda_\tau V_\tau \qquad Q = U_Q \Lambda_Q U_Q^\dagger \tag{13}$$

where U_τ and V_τ are $t \times t$ and $n_\tau \times n_\tau$ unitary matrices and Λ_τ is a $t \times n_\tau$ diagonal matrix such that $tr\{\Lambda_\tau \Lambda_\tau^\dagger\} = tr\{X_\tau X_\tau^\dagger\} = T_\tau P_\tau$. Similarly, U_Q is a unitary matrix and Λ_Q is a diagonal matrix such that $tr\{Q\} = T_s P_d$.

The following theorem will be useful in unveiling the structure of the optimal choice of the training sequence and Q:

Theorem 1: The choice of X_τ and Q which optimizes the rate, I_{GMI}, is such that $U_\tau = U_Q$ and the eigenvalues of $X_\tau X_\tau^\dagger$ and Q (i.e. $\Lambda_\tau \Lambda_\tau^\dagger$ and Λ_Q) are the solution to the following optimization problem:

$$I_{GMI}^{OPT} = \sup_{\substack{\Lambda_Q \geq 0 \\ tr\{\Lambda_\tau \Lambda_\tau^\dagger\} \leq T_\tau P_\tau \\ tr\{\Lambda_Q\} \leq T_s P_d}} \frac{T - T_\tau}{T} E\left[\log\det\left(I_r + \frac{\sigma_H^2}{N} \cdot \frac{H \Lambda_Q \Lambda_\tau \Lambda_\tau^\dagger (N I_t + \sigma_H^2 \Lambda_\tau \Lambda_\tau^\dagger)^{-1} H^\dagger)}{tr\{\sigma_H^2 \Lambda_Q (N I_t + \sigma_H^2 \Lambda_\tau \Lambda_\tau^\dagger)^{-1}\} + 1}\right)\right] \tag{14}$$

Furthermore, the unitary matrices, U_Q and V_τ can be chosen arbitrarily and the choice has no effect on the achievable rate.

The proof is deferred to the appendix.

It is clear from Theorem 1 that it is enough to find the optimal eigenvalues of $X_\tau X_\tau^\dagger$ and Q. Moreover, we can associate each eigenvalue of $X_\tau X_\tau^\dagger$ with an eigenvalue of Q. The number of non-zero eigenvalues of Q ($\text{rank}(Q)$) may be interpreted as the number of effective transmit antennas during data communication. Even though, we may use all antennas during transmission, if $\text{rank}(Q)$ is smaller than the number of transmit antennas, the data on $t - \text{rank}(Q)$ of the antennas will be a linear combination of the data transmitted through the other antennas. Similarly, $\text{rank}(X_\tau X_\tau^\dagger)$ is interpreted as the number of effective antennas during training. From (14) we conclude that it is pointless to have $\text{rank}(Q) > \text{rank}(X_\tau X_\tau^\dagger)$ as some of the eigenvalues of Q will be associated with zero eigenvalues of $X_\tau X_\tau^\dagger$ and therefore,

their effect will be nullified in expression (14).

However, we must solve (14) to find the optimal eigenvalues and that can be quite elaborate in general. Therefore, we will only address two extreme ends of this problem. We will examine the solution to the above problem when the training energy ($T_\tau P_\tau$) is either very high or very low. During the rest of our discussion we will denote the i'th eigenvalue of Λ_τ by $\lambda_{\tau,i}$ and the i'th eigenvalue of Λ_Q by $\lambda_{Q,i}$.

A. Optimization for high training energy, $\frac{\sigma_H^2 T_\tau P_\tau}{\min\{t,n_\tau\}\cdot N} \gg 1$

By high training power we imply, $\frac{\sigma_H^2 T_\tau P_\tau}{\min\{t,n_\tau\}\cdot N} \gg 1$. That is, the effective training energy per degree of freedom is much higher than the noise power at each receiver. We will also assume that the training energy is much higher than the average transmit energy per information symbol (i.e. $T_\tau P_\tau \gg T_s P_d$). This will occur, for example, when the channel coherence interval (T) is much larger than the training duration (T_τ) such that any large increase in the training energy has a negligible impact on the information sequences average power (5). Note that this does not necessarily mean that more energy is allotted for training, $T_\tau P_\tau$, than is allotted for information symbols, $T_d P_d$.

Corollary 1: If $\frac{\sigma_H^2 T_\tau P_\tau}{\min\{t,n_\tau\}\cdot N} \gg 1$, and the training energy is much higher than the average energy per information symbol, $T_\tau P_\tau \gg T_s P_d$, then, the eigenvalues bellow optimize the achievable rate, I_{GMI}.

$$\lambda_{Q,i} = \begin{cases} \frac{T_s P_d}{\min\{t,n_\tau\}} & i \leq \min\{t,n_\tau\} \\ 0 & \text{otherwise} \end{cases}$$
$$|\lambda_{\tau,i}|^2 = \begin{cases} \frac{T_\tau P_\tau}{\min\{t,n_\tau\}} & i \leq \min\{t,n_\tau\} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Proof: Under the assumption that $\frac{\sigma_H^2 T_\tau P_\tau}{\min\{t,n_\tau\}\cdot N} \gg 1$ and for the eigenvalues given in (15), we can write,

$$\frac{\sigma_H^2|\lambda_{\tau,i}|^2}{N+\sigma_H^2|\lambda_{\tau,i}|^2} = \frac{\frac{\sigma_H^2 T_\tau P_\tau}{\min\{t,n_\tau\}\cdot N}}{1+\frac{\sigma_H^2 T_\tau P_\tau}{\min\{t,n_\tau\}\cdot N}} \approx 1.$$

Similarly, under the assumption that $T_\tau P_\tau \gg T_s P_d$, we can write,

$$tr\{\sigma_H^2\Lambda_Q(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}\}+1 = \frac{\sigma_H^2 T_s P_d}{\min\{t,n_\tau\}\cdot N+\sigma_H^2 T_\tau P_\tau}+1 \approx 1\,,$$

where the second equality is a direct consequence of assigning (15) into the trace term. Therefore, by (14), we can write the following expression for the rate achieved by the eigenvalues given in (15):

$$I_{GMI}^{\text{high}} \approx \frac{T-T_\tau}{T} E\Big[\log\det\Big(I_r+\frac{T_s P_d}{N\min\{t,n_\tau\}} H I_t^{\min\{t,n_\tau\}} H^\dagger\Big)\Big]\,, \tag{16}$$

where $I_t^{\min\{t,n_\tau\}}$ is a diagonal matrix with either ones or zeros on the diagonal such that

$$tr\{I_t^{\min\{t,n_\tau\}}\} = \min\{t,n_\tau\}.$$

The element, $I_t^{\min\{t,n_\tau\}}$, appears in the expression above because the number of eigenvalues in $\Lambda_\tau\Lambda_\tau^\dagger$ is limited by $\min\{t,n_\tau\}$ and therefore, only $\min\{t,n_\tau\}$ eigenvalues of Λ_Q come into play in expression (14). We can use the convexity ($\bigcap$) property of the log det function and the isotropic distribution of H [10] to write:

$$I_{GMI}^{\text{high}} \geq \frac{T-T_\tau}{T} E\Big[\log\det\Big(I_r+\frac{1}{N} H\Lambda_Q I_t^{\min\{t,n_\tau\}} H^\dagger\Big)\Big]\,, \quad \forall\Lambda_Q \in \Big\{\Lambda_Q \Big| tr\{\Lambda_Q\} = T_s P_d\Big\} \tag{17}$$

and by the fact that the log det is an increasing monotone function and that:

$$\frac{1}{N} H\Lambda_Q I(\Lambda_\tau\Lambda_\tau^\dagger)_t^{\min\{t,n_\tau\}} H^\dagger \geq \frac{\sigma_H^2}{N}\cdot\frac{H\Lambda_Q\Lambda_\tau\Lambda_\tau^\dagger(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}H^\dagger)}{tr\{\sigma_H^2\Lambda_Q(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}\}+1}\,,$$

where

$$\left(I(\Lambda_\tau\Lambda_\tau^\dagger)_t^{\min\{t,n_\tau\}}\right)_{ij}=\begin{cases}1 & i=j\ ,\ (\Lambda_\tau\Lambda_\tau^\dagger)_{ii}>0\\ 0 & \text{otherwise}\end{cases},$$

we can write:

$$I_{GMI}^{\text{high}}\geq\frac{T-T_\tau}{T}E\Big[\log\det\Big(I_r+\frac{\sigma_H^2}{N}\cdot\frac{H\Lambda_Q\Lambda_\tau\Lambda_\tau^\dagger(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}H^\dagger)}{tr\{\sigma_H^2\Lambda_Q(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}\}+1}\Big)\Big]$$
$$\forall(\Lambda_Q,\Lambda_\tau)\in\left\{\Lambda_Q,\Lambda_\tau\middle|tr\{\Lambda_\tau\Lambda_\tau^\dagger\}\leq T_\tau P_\tau\ ,\ tr\{\Lambda_Q\}\leq T_sP_d\right\} \tag{18}$$

which is the r.h.s of (14). Therefore, for high training power, we have $I_{GMI}^{\text{high}}\approx I_{GMI}^{OPT}$. ■

The result given in expression (16) is appropriate when our channel estimation is such that the portion of the "overall noise" caused by the channel estimation error is negligible compared to the channel noise power, NI_τ.

B. Optimization for low training energy, $r\cdot\frac{\sigma_H^2T_\tau P_\tau}{N}\ll 1$.

By low training energy we imply that $r\cdot\frac{\sigma_H^2T_\tau P_\tau}{N}\ll 1$. That is, the effective training energy accumulated over all receive antennas is much lower than the noise variance per antenna. Clearly, the optimal choice of the eigenvalues of $X_\tau X_\tau^\dagger$ and Q is different from the solution given in corollary 1 and is stated by the following corollary:

Corollary 2: when $r\cdot\frac{\sigma_H^2T_\tau P_\tau}{N}\ll 1$, the eigenvalues bellow optimize the achievable rate, I_{GMI}.

$$\lambda_{Q,i}=\begin{cases}T_sP_d & i=1\\ 0 & \text{otherwise}\end{cases}$$
$$|\lambda_{\tau,i}|^2=\begin{cases}T_\tau P_\tau & i=1\\ 0 & \text{otherwise}\end{cases} \tag{19}$$

Proof: By the condition that $r \cdot \frac{\sigma_H^2 T_\tau P_\tau}{N} \ll 1$ we can conclude that $\frac{1}{N}(N + \sigma_H^2 |\lambda_{\tau,i}|^2) \approx 1$. Equally, we can write, $tr\{\sigma_H^2 \Lambda_Q (N I_t + \sigma_H^2 \Lambda_\tau \Lambda_\tau^\dagger)^{-1}\} \approx \frac{\sigma_H^2}{N} T_s P_d$. Therefore, we can replace the problem given in expression (14) with the following one:

$$I_{GMI}^{\text{low}} \approx \sup_{\substack{\Lambda_Q \geq 0 \\ tr\{\Lambda_\tau \Lambda_\tau^\dagger\} \leq T_\tau P_\tau \\ tr\{\Lambda_Q\} \leq T_s P_d}} \frac{T - T_\tau}{T} E\Big[\log\det\Big(I_r + \frac{\sigma_H^2}{N} \cdot \frac{H \Lambda_Q \Lambda_\tau \Lambda_\tau^\dagger H^\dagger)}{\sigma_H^2 T_s P_d + N}\Big)\Big] \tag{20}$$

Using the equality, $\det(I + AB) = \det(I + BA)$ and by the fact that the cumulative product of the eigenvalues of a non-negative hermitian matrix, R, is always smaller than the cumulative product of the values on its diagonal ($\prod \lambda_{R,i} \leq \prod R_{ii}$), we can upper-bound expression (20) and write:

$$I_{GMI}^{\text{low}} \leq \sup_{\substack{\Lambda_Q \geq 0 \\ tr\{\Lambda_\tau \Lambda_\tau^\dagger\} \leq T_\tau P_\tau \\ tr\{\Lambda_Q\} \leq T_s P_d}} \frac{T - T_\tau}{T} \sum_{i=1}^{t} E\Big[\log\Big(1 + \frac{\sigma_H^2}{N} \frac{H_{.i}^\dagger H_{.i} \lambda_{Q,i} |\lambda_{\tau,i}|^2}{\sigma_H^2 T_s P_d + N}\Big)\Big], \tag{21}$$

where $H_{.i}$ is the i'th column of H. We can approximate the above equation by relying on the fact that for small x (such that $0 < x \ll 1$), $\log(1 + x) \approx x$, and because of the low training energy,

$$\frac{\sigma_H^2}{N} \frac{H_{.i}^\dagger H_{.i} \lambda_{Q,i} |\lambda_{\tau,i}|^2}{\sigma_H^2 T_s P_d + N} = \frac{H_{.i}^\dagger H_{.i} \lambda_{Q,i}}{\sigma_H^2 T_s P_d + N} \cdot \frac{\sigma_H^2 |\lambda_{\tau,i}|^2}{N} \ll 1$$

for most probable instances of H. Therefore, we can write

$$I_{GMI}^{\text{low}} \lesssim \sup_{\substack{\Lambda_Q \geq 0 \\ tr\{\Lambda_\tau \Lambda_\tau^\dagger\} \leq T_\tau P_\tau \\ tr\{\Lambda_Q\} \leq T_s P_d}} \frac{T - T_\tau}{T} \sum_{i=1}^{t} \Big(\frac{\sigma_H^2}{N} E\Big[H_{.i}^\dagger H_{.i}\Big] \frac{\lambda_{Q,i} |\lambda_{\tau,i}|^2}{\sigma_H^2 T_s P_d + N}\Big) \tag{22}$$

It can be easily shown that the approximation made in (22) is optimized by the choice of the eigenvalues given in (19). However, for this choice of Λ_Q and Λ_τ, the upper bound in (21) can be replaced with equality. Therefore, this choice is the optimal choice for very low training energy. ■

First, note that the optimal distribution of the power and training covariance matrices varied significantly between the low and high training energy cases. The major difference is in the number of antennas which are effectively used. When the training energy is very low, we effectively use only a single antenna during information transmission. Obviously, we may choose $U_\tau = U_Q$ arbitrarily and thus distribute the transmit power between all antennas, but this will only cause a linear dependence between the antennas during the transmission of the information sequences. Conversely, when the training energy is very high, all antennas are used simultaneously and effectively. The information sent at each of the antennas may be independent of the information sent through the other antennas.

The above results have some resemblance to the water pouring solution which divides the transmit power between channels depending on the SNR in each channel and the total transmit power. When the total transmit power is very low, under the water pouring solution, only the channels with the highest SNRs will be used. We can see similar behavior in our case.

As mentioned, as the coherence interval grows and becomes considerably larger than the training interval ($T \gg T_\tau$), the training power can be easily made very high without significant loss to the information sequences power. Therefore, the solution in Corollary 1 is relevant and expression (16) is used to approximate I_{GMI}. It is easy to see that if $T_\tau \geq t$, expression (16) is equal to the capacity of the coherent channel [10] and we can conclude that as the coherence interval becomes larger, the capacity of the non-coherent channel [8], approaches the coherent channel capacity. This was previously shown in [5], [8] for the case of $t = r = 1$ and was extended to any t and r using heuristic considerations.

By adopting the training scheme of Corollary 1 and using expression (14), we can see that for $T_\tau = t$, the achievable rate, I_{GMI}, will coincide with the lower bound on the capacity

calculated in [4], [13]. Zheng and Tse [13], [12] showed that at high SNR's, this scheme comes within a constant of the non-coherent channel capacity.

By choosing the unitary matrices U_τ and U_Q to be the identity matrix, I_t, it is clear from corollary 1 and 2 that for both cases the information sequence power is equally distributed between all useful (a single antenna for the low training energy case and $\min\{t, n_\tau\}$ antennas for high training power case) transmit antennas. The same distribution of transmit power is assumed in [4]. Therefore, the results for the optimal allocation of power and duration of the training and information sequences in [4] are also applicable to the above two cases.

Last, at very low SNR's we expect the training energy to be low as well and as a result, the optimal training scheme will effectively use only a single transmit antenna. This resembles the result in [13], [12] which states that at low SNRs, the non-coherent channel capacity is achieved using only a single transmit antenna. Furthermore, we can conclude from expression (20) that at very low SNRs , the achievable rate is proportional to $\Lambda_Q \Lambda_\tau \Lambda_\tau^\dagger$. Since both Λ_Q and $\Lambda_\tau \Lambda_\tau^\dagger$ are proportional to the SNR, we can see that the achievable rate is proportional to the square of the SNR at very low SNRs. This result was also shown in [4] and indicates that our communication system performs much worse than the channel capacity at low SNRs [7].

C. Maximum training duration

In [4] the authors show that if the powers of the training sequence, P_τ, and the information sequence, P_d, are allowed to differ (i.e. $P_\tau \neq P_d$), then the optimal choice of the training duration is $n_\tau = t$. However, their results are applicable to the achievable rate, I_{GMI}, only when $\Lambda_\tau \Lambda_\tau^\tau$ and Λ_Q are a scalar multiple of I_t. We will now state a corollary regarding the optimal training duration for a general choice of Q and X_τ.

Corollary 3: If the powers of the training sequence, P_τ, and the information sequence,

P_d, are allowed to vary, the training sequence, X_τ, which maximizes I_{GMI} is as long as its rank (i.e. for the optimal training sequence, $\text{rank}(X_\tau X_\tau^\dagger) = n_\tau$).

Proof: If there is no requirement for equal training and data power, we can choose U_τ, V_τ and U_Q to be identity matrices (Theorem 1). For such a choice, if $n_\tau > \text{rank}(X_\tau X_\tau^\dagger)$, the last $n_\tau - \text{rank}(X_\tau X_\tau^\dagger)$ columns of X_τ will be zero. However, there is no point of having time slots at which nothing is sent (zero power for training). It makes better sense to use these spare time-slots to transmit information symbols. Nevertheless, we must remember that the total energy used for information symbols is limited by $T_d P_d = TP - T_\tau P_\tau$ (by (5)). If we increase the time allotted for information (increase T_d), we will, respectively, decrease the energy per information symbol ($T_s P_d$). If T_d' is the "new" time-period allotted for information and $\gamma = \frac{T_d'}{T_d}$, then the energy allotted per information symbol in the "new" scheme is $T_s P_d/\gamma$.

We will prove the corollary in the following manner: Assume that Λ_τ and Λ_Q are optimized for the original training duration and power allocation. Then we may write expression (14) as follows:

$$\begin{aligned} I_{GMI}^{OPT}(T_d) &= \frac{T_d}{T} E\Big[\log\det\Big(I_r + \frac{\sigma_H^2}{N}\cdot\frac{H\Lambda_Q\Lambda_\tau\Lambda_\tau^\dagger(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}H^\dagger)}{tr\{\sigma_H^2\Lambda_Q(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}\}+1}\Big)\Big] \\ &= \frac{T_d'}{T}\sum_{i=1}^{T_d'}\frac{e_i}{T_d'} E\Big[\log\det\Big(I_r + \frac{\sigma_H^2}{N}\cdot\frac{H\Lambda_Q\Lambda_\tau\Lambda_\tau^\dagger(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}H^\dagger)}{tr\{\sigma_H^2\Lambda_Q(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}\}+1}\Big)\Big] \end{aligned} \tag{23}$$

where

$$e_i = \begin{cases} 1 & \forall i \le T_d \\ 0 & \text{otherwise} \end{cases}$$

Using the convexity of the log det function, we can write:

$$\begin{aligned} I_{GMI}^{OPT}(T_d) &\le \frac{T_d'}{T} E\Big[\log\det\Big(I_r + \frac{\sigma_H^2}{N}\cdot\frac{H\sum_{i=1}^{T_d'}(\frac{e_i}{T_d'}\Lambda_Q)\Lambda_\tau\Lambda_\tau^\dagger(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}H^\dagger)}{tr\{\sigma_H^2\Lambda_Q(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}\}+1}\Big)\Big] \\ &= \frac{T_d'}{T} E\Big[\log\det\Big(I_r + \frac{\sigma_H^2}{N}\cdot\frac{H\frac{\Lambda_Q}{\gamma}\Lambda_\tau\Lambda_\tau^\dagger(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}H^\dagger)}{tr\{\sigma_H^2\Lambda_Q(NI_t+\sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}\}+1}\Big)\Big] \end{aligned} \tag{24}$$

By decreasing the trace term in the above expression we will make the entire expression larger. Therefore, we can write:

$$I_{GMI}^{OPT}(T_d) \leq \frac{T_d'}{T} E\Big[\log\det\Big(I_\tau + \frac{\sigma_H^2}{N} \cdot \frac{H\frac{\Lambda_Q}{\gamma}\Lambda_\tau\Lambda_\tau^\dagger(NI_t + \sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}H^\dagger)}{tr\{\sigma_H^2\frac{\Lambda_Q}{\gamma}(NI_t + \sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}\} + 1}\Big)\Big] \tag{25}$$
$$\leq I_{GMI}^{OPT}(T_d')$$

The last inequality proves that using more time-slots at the expense of the power of the information symbols is worthwhile and therefore, we may write

$$n_\tau \leq \text{rank}(X_\tau X_\tau^\dagger) \leq \min\{n_\tau, t\}\ ,$$

where the second inequality is due to the fact that X_τ is a matrix of size $t \times n_\tau$. The above equation is true only if $n_\tau = \text{rank}(X_\tau X_\tau^\dagger)$. ■

We can use the fact that $\text{rank}(X_\tau X_\tau^\dagger) \leq t$ and say that in general, the optimal time allotted for training can not be greater than the number of transmit antennas.

The above corollary and Theorem 1 lead us to conclude that when no constraint is put on the power allocation between training and information sequences, an optimal training sequence sends individual "pilots" one antenna at a time through each of the useful antennas. The matching code-book sends signals independently through each of the trained antennas.

We will not discuss the case where the training and information sequences are required to have equal power apart from saying that in that case, this may be done by choosing V_τ such that it equally distributes the training sequence power over all training time slots.

D. Plots of I_{GMI} for different training schemes

We have plotted in Figure 2 curves of I_{GMI} vs. $\frac{\sigma_H^2 T_s P}{N}$ for a channel with 8 receive antennas ($r = 8$) and communication schemes that use 1, 4 and 8 transmit antennas ($t = 1, 4, 8$). The coherence length is assumed to be $n_T = 40$ symbols long. The information and

training sequences power distribution are similar to those in [4] such that $\Lambda_Q = \frac{T_s P_d}{t} I_t$ and $X_\tau X_\tau^\dagger = T_s P_\tau I_t$. The allocation of power between the training and information sequences was optimized for each $\frac{\sigma_H^2 T_s P}{N}$ and was done according to the results in [4] (which apply only to these types of Q and X_τ). The achievable rates, I_{GMI} were calculated by using the expression obtained in [10] for the coherent multi-antenna fading channel. We can see from Figure 2 that bellow 5.2dB the 4 transmit antennas scheme outperforms the 8 antennas scheme. Similarly, bellow -2.9dB, the single transmit antenna scheme outperforms the 4 antennas scheme.

IV. Summary and Conclusions

By Theorem 1 we can see that not only does the training energy and duration have an effect on the achievable rate, but also the structure of the training sequence impacts performance. It is also clear that the covariance matrix, Q, has to be adapted to the chosen training sequence or viceversa.

Corollaries 1 and 2 indicate that the optimal number of useful antennas varies significantly depending on the training energy and SNR. We may conclude that if there is some sort of negotiation between the transmitter and receiver prior to information exchange, such that the transmitter may learn the SNR, the system designer would do wisely to allow the transmitter to decide on the number of antennas it will use during the information exchange, based on the SNR. Such that as the SNR becomes higher, more antennas (of the available ones) will be used for transmission.

Finally, we have shown that in the case where the training and information powers are allowed to vary, an optimal training scheme is one which transmits a pilot through each of the antennas, one at a time. By Theorem 1 we can see that the equal power requirement may be fulfilled by choosing the unitary matrix, V_τ, appropriately, such as it distributes the

training energy over time. Similarly, $U_\tau = U_Q$ may be used to distribute the transmit power over antennas.

APPENDIX

I. PROOF OF THEOREM 1

Proof: Using (11) and the representation of the training sequence in (13), we can write the fading estimation, $\widehat{H}$, and fading estimation error, $\widetilde{H}$, as follows:

$$\begin{aligned}\widehat{H} &= \sigma_H^2(\breve{H}\Lambda_\tau + \breve{Z}_1^{n_\tau})\Lambda_\tau^\dagger(NI_t + \sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}U_\tau^\dagger \\ \widetilde{H} &= \left(\breve{H} - \sigma_H^2(\breve{H}\Lambda_\tau + \breve{Z}_1^{n_\tau})\Lambda_\tau^\dagger(NI_t + \sigma_H^2\Lambda_\tau\Lambda_\tau^\dagger)^{-1}\right)U_\tau^\dagger\end{aligned} \tag{26}$$

where $\breve{H} = HU_\tau$ and $\breve{Z}_1^{n_\tau} = Z_1^{n_\tau}V_\tau$. The elements of the matrices H and $Z_1^{n_\tau}$, in our channel model, are circularly symmetric Gaussian RVs and independent of each other. Therefore, both matrices are isotropic [10] (the distribution of the matrices is invariant to either right or left multiplication by a unitary matrix). Consequently, the distributions of $\breve{H}$ and $\breve{Z}_1^{n_\tau}$ are identical to those of H and $Z_1^{n_\tau}$. From (26) we can make the following conclusions:

1. The achievable rate (8) depends only on the distribution of $\widehat{H}$ and $\widetilde{H}$ (through (12)). Since in both cases the distribution is not affected by V_τ, the choice of V_τ (13) has no effect on the achievable rate.
2. The joint distribution of $\widehat{H}$ and $\widetilde{H}$ is the same as the joint distribution of $\breve{\widehat{H}}U_\tau^\dagger$ and $\breve{\widetilde{H}}U_\tau^\dagger$ where $\breve{\widehat{H}}$ and $\breve{\widetilde{H}}$ are the estimated fading and estimation error when X_τ is chosen to be diagonal.
3. From the isotropic distributions of the noise and the fading we can see that the distribution of $\widehat{H}$ and $\widetilde{H}$ are invariant to any left multiplication by a unitary matrix.
4. $P_{\widetilde{N}}$ is a scalar multiple of the identity matrix. This can be seen by rewriting (12) as

follows:

$$P_{\tilde{N}} = E\Big[\tilde{H}Q\tilde{H}^{\dagger}\Big|\widehat{H}\Big] + NI_r = E\Big[U\tilde{H}Q\tilde{H}^{\dagger}U^{\dagger}\Big|\widehat{H}\Big] + NI_r$$

for all unitary matrices U. The second equality follows from conclusion (3). Therefore, $P_{\tilde{N}} = UP_{\tilde{N}}U^{\dagger}$ for all unitary matrices, U. This can be so only if $P_{\tilde{N}} = \alpha I_r$ for some α. Where α can be found by using the equality:

$$\begin{aligned} r\alpha &= tr\{P_{\tilde{N}}\} \\ &= rN + tr\Big\{E\Big[Q\tilde{H}^{\dagger}\tilde{H}\Big|\widehat{H}\Big]\Big\} \\ &= rN + tr\Big\{QU_{\tau}E\Big[\breve{\tilde{H}}^{\dagger}\breve{\tilde{H}}\Big|\widehat{H}\Big]U_{\tau}^{\dagger}\Big\} \end{aligned}$$

The rows of $\breve{\tilde{H}}$ are independent. Therefore, it is possible to show that for the case of the MMSE estimator:

$$E\Big[\breve{\tilde{H}}^{\dagger}\breve{\tilde{H}}\Big|\widehat{H}\Big] = r\sigma_H^2 N(NI_t + \sigma_H^2\Lambda_{\tau}\Lambda_{\tau}^{\dagger})^{-1}.$$

Therefore, we can write:

$$\alpha = tr\Big\{\sigma_H^2 N(NI_t + \sigma_H^2\Lambda_{\tau}\Lambda_{\tau}^{\dagger})^{-1}\breve{Q}\Big\} + N \tag{27}$$

where $\breve{Q} = U_{\tau}^{\dagger}QU_{\tau}$.

5. Consequently, we can write I_{GMI} in the following manner:

$$I_{GMI} = E_{\widehat{H}}\left[\log\det\left(I_r + \alpha^{-1}\widehat{H}Q\widehat{H}^{\dagger}\right)\right] \tag{28}$$

where α is dependent only on the diagonal elements of $\breve{Q}$ and $\Lambda_{\tau}\Lambda_{\tau}^{\dagger}$.

As mentioned before, $\widehat{H} \sim \breve{\widehat{H}}U_{\tau}^{\dagger}$ (i.e $\widehat{H}$ and $\breve{\widehat{H}}U_{\tau}^{\dagger}$ have the same probability law). The rows of $\breve{\widehat{H}}$ are IID and have the following covariance matrix:

$$E\big[\breve{\widehat{H}}_i^{\dagger}\breve{\widehat{H}}_i\big] = (\sigma_H^2)^2\Lambda_{\tau}\Lambda_{\tau}^{\dagger}(NI_t + \sigma_H^2\Lambda_{\tau}\Lambda_{\tau}^{\dagger})^{-1} = (D_{\tau})^2 \quad \forall i = 1..r$$

where $\breve{\tilde{H}}_i$ is the i'th row of $\breve{\tilde{H}}$ and D_τ is a non-negative $t \times t$ diagonal matrix. Therefore, $\breve{\tilde{H}} \sim \frac{1}{\sigma_H} H D_\tau$, and thus $\widehat{H} \sim \frac{1}{\sigma_H} H D_\tau U_\tau^\dagger$. This allows us to rewrite (28) as follows:

$$I_{GMI} = E_H \left[\log \det \left(I_r + \frac{\alpha^{-1}}{\sigma_H^2} H D_\tau \breve{Q} D_\tau H^\dagger \right) \right] \tag{29}$$

where $\breve{Q}$ is as defined in (27).

However, we are interested in optimizing the choice of Q and X_τ. We will use a similar scheme to that used in [10] to show that $\breve{Q}$ must be diagonal for an optimal solution. For that purpose we will define a unitary matrix U_i' which is similar to the identity matrix except for the i'th element which is -1. Below is an example

$$U_i' = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & -1 & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix}$$

First, note that if $\breve{Q}$ is multiplied both from the right and left by U_i', both the i'th row and column are negated except for the i'th element on the diagonal which remains unchanged. We will denote this "new" $\breve{Q}$ by $\breve{Q}' = U_i' \breve{Q} U_i'$ which is also a non-negative Hermitian matrix. We will also denote the achievable rate using $\breve{Q}$ by $I_{GMI}(\breve{Q})$ and the achievable rate using $\breve{Q}'$ by $I_{GMI}(\breve{Q}')$. From conclusion (5) (above) we can see that we get the same α for $\breve{Q}$ and $\breve{Q}'$. Furthermore, it is easily shown that $H D_\tau \breve{Q} D_\tau H^\dagger$ and $H D_\tau \breve{Q}' D_\tau H^\dagger$ have the same probability law. Therefore, $I_{GMI}(\breve{Q}) = I_{GMI}(\breve{Q}')$.

Second, because log det is a convex ($\cap$) function, we can write:

$$I_{GMI}(\breve{Q}) = \frac{1}{2}(I_{GMI}(\breve{Q}') + I_{GMI}(\breve{Q})) \leq I_{GMI}\left(\frac{1}{2}(\breve{Q}' + \breve{Q})\right) = I_{GMI}(\breve{Q}'')$$

where $\breve{Q}''$ has the same values on the diagonal as $\breve{Q}$ but the other elements on the i'th row and column are zero.

Therefore, from the above two notes we may conclude that of all possible $\breve{Q}$ with a given set of values on the diagonal, the one which gives the best rate is that which is diagonal. Consequently, the optimal choice of U_τ must be that which causes $\breve{Q}$ to be diagonal. If we write Q as $U_Q \Lambda_Q U_Q^\dagger$, the optimal choice of U_τ for the optimal choice of Q would be:

$$U_\tau = U_Q \tag{30}$$

From (27), (29) and (30) we can see that:

$$\begin{aligned}
&\sup_{\substack{X_\tau\,,\,Q\geq 0\\ tr\{X_\tau X_\tau^\dagger\}\leq T_\tau P_\tau\\ tr\{Q\}\leq T_s P_d}} I_{GMI} = \\
&\sup_{\substack{\Lambda_Q\geq 0\,,\,\Lambda_\tau\\ tr\{\Lambda_\tau \Lambda_\tau^\dagger\}\leq T_\tau P_\tau\\ tr\{\Lambda_Q\}\leq T_s P_d}} E\Big[\log\det(I_\tau + \frac{1}{\alpha} H\sigma_H^2 \Lambda_Q \Lambda_\tau \Lambda_\tau^\dagger (NI_t + \sigma_H^2 \Lambda_\tau \Lambda_\tau^\dagger)^{-1} H^\dagger)\Big] = \\
&\sup_{\substack{\Lambda_Q\geq 0\\ tr\{\Lambda_\tau \Lambda_\tau^\dagger\}\leq T_\tau P_\tau\\ tr\{\Lambda_Q\}\leq T_s P_d}} E\Big[\log\det\Big(I_\tau + \frac{\sigma_H^2}{N}\cdot\frac{H\Lambda_Q \Lambda_\tau \Lambda_\tau^\dagger (NI_t + \sigma_H^2 \Lambda_\tau \Lambda_\tau^\dagger)^{-1} H^\dagger)}{tr\{\sigma_H^2 \Lambda_Q (NI_t + \sigma_H^2 \Lambda_\tau \Lambda_\tau^\dagger)^{-1}\} + 1}\Big)\Big]
\end{aligned} \tag{31}$$

The last equality in (31) is dependent only on the eigenvalues of the optimal $X_\tau X_\tau^\dagger$ and Q. Therefore, we may find the optimal training sequence by finding the optimal eigenvalues for X_τ and Q and by choosing $U_Q = U_\tau$. We can now take into account the fact that we only use $\frac{T-T_\tau}{T}$ of the block length to send information symbols and multiply the result above by this constant to get the effective information rate per symbol. ■

Acknowledgment

We thank Amos Lapidoth for his insights in the stimulating discussions we had.

REFERENCES

[1] Srihari Adireddy, Lang Tong, and Harish Viswanathan, "Optimal placement of training for unknown channels," *IEEE Transactions on Information Theory*, to appear.

[2] Jaiganesh Balakrishnan, Markus Rupp, and Harish Viswanathan, "Optimal channel training for multiple antenna systems," *Multiaccess, Mobility and Teletraffic for Wireless Communications Workshop (Duck Key, Florida, USA)*, Dec. 2000.

[3] C.-S. Chou and D.W. Lin, "Signal design and receiver dimensioning space-time Viterbi equalisation," *IEE Proc. Commun.*, , no. 3, pp. 132–138, June 2001.

[4] Babak Hassibi and Bertrand M. Hochwald, "How much training is needed in multiple-antenna wireless links?," *IEEE Transactions on Information Theory*, to appear.

[5] Bertrand M. Hochwald and Thomas L. Marzetta, "Unitary space-time modulation for multiple-antenna communications in rayleigh flat fading," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 543–563, Mar. 2000.

[6] Amos Lapidoth and Stefan M. Moser, "Convex-programming bounds on the capacity of flat-fading channels," *Proc. IEEE International Symposium on Information Theory*, p. 52, June 2001.

[7] Amos Lapidoth and Shlomo Shamai (Shitz), "Fading channels: How perfect need "perfect side-information" be?," *IEEE Transactions on Information Theory*, to appear.

[8] Thomas L. Marzetta and Bertrand M. Hochwald, "Capacity of mobile multiple-antenna communication link in rayleigh flat fading," *IEEE Transactions on Information Theory*, vol. 45, pp. 139–157, 1999.

[9] Shuichi Ohno and Geogios B. Giannakis, "Capacity maximizing pilots for wireless OFDM over rapidly fading channels," *IEEE Transactions on Information Theory*, to appear.

[10] I. E. Telatar, "Capacity of multi-antenna gaussian channels," *Eur. Trans. Telecom.*, vol. 10, pp. 585–595, Nov. 1999.

[11] Hanan Weingarten, Yossi Steinberg, and Shlomo Shamai (Shitz), "Gaussian codes and the scaled nearest neighbor decoding in fading multi-antenna channels, part 1: Information theoretic perspective," *submitted to IEEE Transactions on Information Theory*, 2001.

[12] L. Zheng and D. Tse, "Packing spheres into the grassman manifold: A geometric approach to noncoherent multi-antenna channels," *Proc. 37'th Annual Allerton Conference on Communication, Control and Computing*, Sept. 1999.

[13] Lizhong Zheng and David N.C. Tse, "Communication on the grassman manifold: A geometric approach to the non-coherent multiple antenna channel," *IEEE Transactions on Information Theory*, to appear.

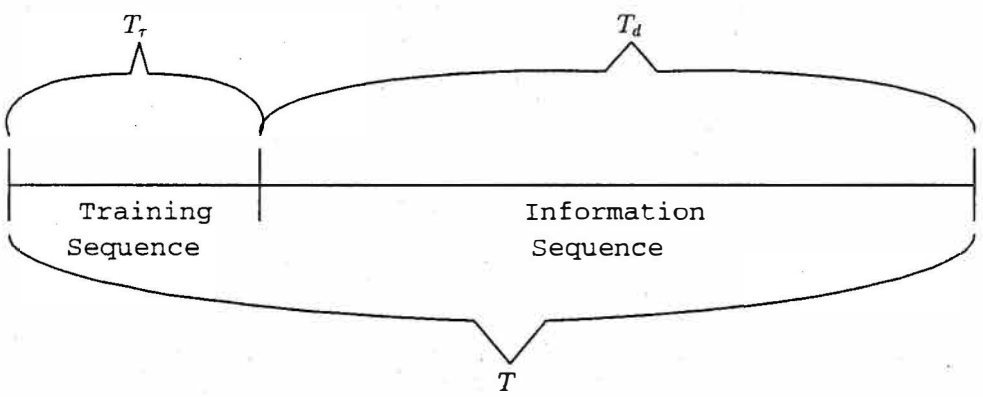

Fig. 1. Training scheme. The fading coherence period is T. Part of the coherence interval (T_τ) is used to send a training sequence while the rest of the block (T_d) is used for transmitting information sequences.

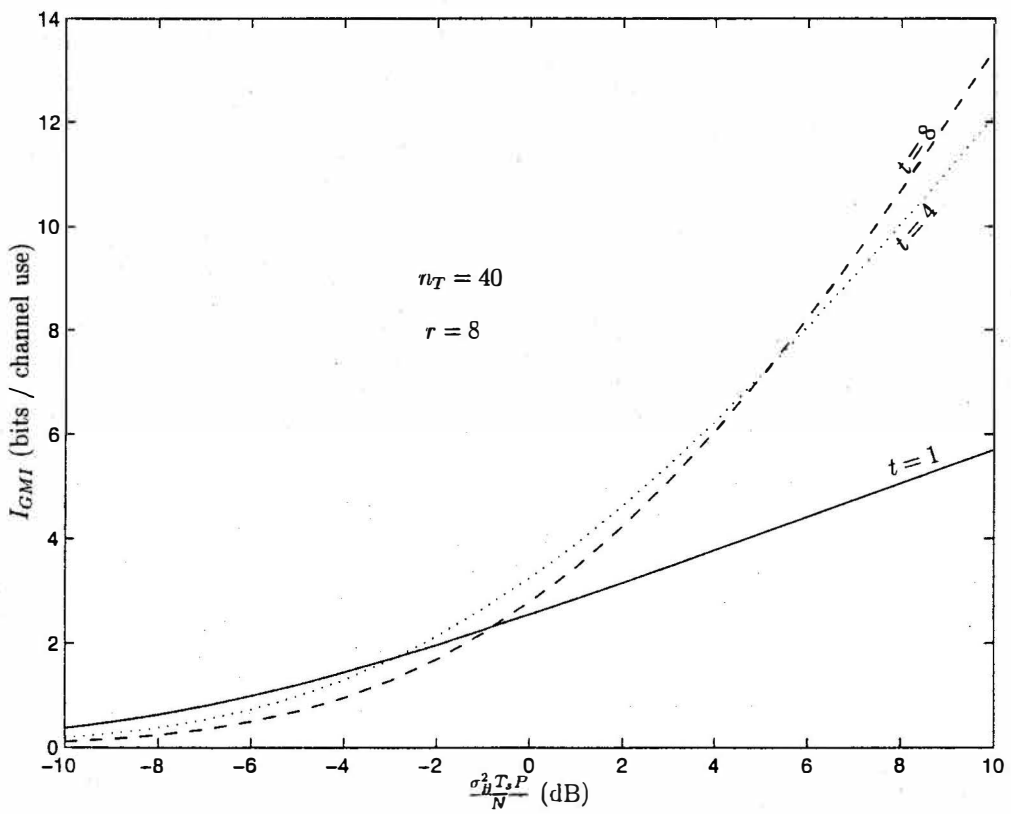

Fig. 2. Achievable rate, I_{GMI}, as a function of $\frac{\sigma_H^2 T_S P}{N}$ for 1, 4 and 8 transmit antennas and 8 receive antennas. The coherence interval is $n_T = 40$ symbols long and the number of symbols used during training is equal to the number of transmit antennas, $n_\tau = t$. The training scheme is suboptimal and therefore, the curves cross.

The *Center for Communication and Information Technologies* (CCIT)

is managed by the Department of Electrical Engineering.

This Technical Report is listed also as

EE PUB #1303, January 2002.