CAPACITY PRE-LOG ACHIEVING SCHEMES FOR STATIONARY FREQUENCY-FLAT MULTI-ANTENNA CHANNELS

Umer SALIM, Dirk SLOCK

Eurecom Institute , 2229 route des Crêtes, B.P. 193,
06904 Sophia Antipolis Cedex, FRANCE
e-mail:{umer.salim,dirk.slock}@eurecom.fr

ABSTRACT

We analyze the Mutual Information of stationary Frequency-Flat MIMO Channels, that are hence characterized by a Doppler spectrum. Absence of Channel State Information at Transmitter or Receiver (no CSIT/CSIR) is assumed. For peak-power limited SISO frequency-flat channels with stationary Gaussian fading, it has been shown by Lapidoth [1] that at high SNR, the capacity is determined by a pre-log factor that is equal to the bandwidth of frequencies where the channel Doppler spectrum is zero (the complementary part of the Doppler bandwidth).

In this paper, we give simple upper and lower bounds for the capacity of MIMO channels. These bounds are very revealing about the multiplexing gain (pre-log factor) of the system. Then we extend Lapidoth's result to MIMO channels with the help of these bounds. In a general (block) stationary setting, the absence of CSIR decreases the pre-log with a factor equal to 1 minus the average number of parameters per symbol period that parameterize the channel. This reduction term is proportional to the Doppler bandwidth and the number of transmit antennas. We introduce channel parameterizations that induce a split in the transmitted symbols between "learning" symbols (that carry loglog(SNR) information) and "data" symbols (that carry log(SNR) information). The pre-log factor is the proportion of "data" symbols. This decomposition shows the optimality of certain training schemes for practical SNR values. The optimal pre-log requires optimization w.r.t. the number of active transmitting antennas, as a function of Doppler bandwidth.

The Eurecom Institute's research is partially supported by its industrial partners: BMW, Bouygues Telecom, Cisco Systems, France Télécom, Hitachi Europe, SFR, Sharp, ST Microelectronics, Swisscom, Thales. The research work leading to this paper has also been partially supported by the European Commission under the IST 6FP research network of excellence NEWCOM and by the French RNRT projects Lao Tseu and OPUS.

1. INTRODUCTION

Information theoretic bounds for different types of channels have got utmost importance since the explosion of research in MIMO promised new dimensions for data communication. Such capacity bounds are very important in the sense that they give the theoretical limits and motivate researchers to achieve them in practical systems literally or asymptotically. The area of capacity analysis for non-coherent (no CSIR and no CSIT) fading channels has received considerable attention in recent years since the usual assumption of perfect CSIR is not true in practical systems and channel realizations need to be estimated for correct decoding of data.

Usually block fading models are assumed for obtaining capacity bounds in the no CSIR (non-coherent) case. In the standard version of this model [2], the fading remains constant over blocks consisting of T symbol periods, and changes independently from block to block. Capacity bounds are obtained by introducing training segments in an ad hoc fashion. For the standard block fading model, the capacity is shown [2], [3] to grow logarithmically with SNR. Later Liang and Veeravalli [4] allowed the fading to vary inside the block with a certain correlation matrix characterized by the rank Q and showed for SISO channels that the capacity pre-log is $(1-Q/T)$.

Non-coherent capacity has also been analyzed with the channel fading process being symbol-by-symbol stationary. In this model, fading is not independent but time selective without any block structure. Surprisingly, this model leads to very different capacity results: contrary to log(SNR) capacity growth in block fading channels, here the capacity grows only double logarithmically with SNR at high SNR [5], [6], [7] when the fading process is non-bandlimited, i.e. the channel prediction error is non-zero even if infinite past is known.

For symbol-by-symbol stationary Gaussian fading channels, if the Doppler spectrum is band-limited (limited support), then the fading process is called non-regular and the prediction error given the infinite past is zero. Lapidoth [1] studied the SISO case for this kind of fading processes showing that capacity grows logarithmically with SNR and capacity pre-log is the Lebesgue measure of the frequencies where the spectral density of the fading process (Doppler spectrum) has nulls.

Chen and Veeravalli [8] introduce a block-stationary channel model that can encompass both the per-symbol stationary and block fading models. They obtain the SISO capacity pre-log for both cases. They argue that the log(SNR) regime results from the rank deficiency of the correlation matrix of the fading process (though bandlimited fading only leads to rank deficiency over a block as the block length goes to infinity).

We should emphasize that the fading processes of interest to us in this paper are stationary and strictly bandlimited, the ones for which Lapidoth [1] established the capacity pre-log in SISO case. In section 2, we give the system model. Section 3 comes with our simple lower and upper bounds of non-coherent capacity. In section 4, we give two capacity pre-log achieving schemes for SISO systems and section 5 is about MIMO extension of these two schemes. In section 6, there is characterization of optimal number of active transmitting antennas in terms of Doppler bandwidth. Section 7 gives the concluding remarks and references are given in the end.

2. SYSTEM MODEL

We consider a MIMO fading channel whose time-k output $Y_k \in C^N$ is given by

$$Y_k = H_k X_k + Z_k \tag{1}$$

where $X_k \in C^M$ denotes the time-k channel input vector and the fading matrix $H_k \in C^{N \times M}$ represents the time-k fading matrix and $Z_k \in C^N$ denotes the additive gaussian noise vector. Here C denotes the complex field and, M and N represent the number of transmit and receive antennas respectively. We assume that the zero-mean circularly complex Gaussian noise is spatiotemporally white with spatial covariance matrix I_N, which represents the $N \times N$ identity matrix. The channel fading process $\{H_k\}$ is assumed to be stationary, ergodic and with finite second order moment, i.e. $E[||H_k||^2] < \infty$. We take the fading process to be strictly bandlimited, so it is a non-regular stochastic process with limited Doppler spectrum support. Moreover we impose the restriction that the support is of size $1/D$ for each channel entry, where D is an integer (extensions to a rational D are possible).

If we are working with SISO systems, our channel model is

$$y_k = h_k x_k + z_k \tag{2}$$

where everything is now complex scalar but channel fading and noise have the same temporal properties as in MIMO case.

Sometimes we will be working over a block of D symbol times. In that case the joint description of (1) over D symbol periods becomes

$$\underline{Y}_k = \underline{H}_k \underline{X}_k + \underline{Z}_k \quad . \tag{3}$$

We adopt the convention of representing the variables for D-symbol block as underlined letters. Here $\underline{Y}_k$ and and $\underline{Z}_k$ have lengths $N \times D$, $\underline{X}_k$ is of length $M \times D$ and $\underline{H}_k = blockdiag(H_{kD+1}, H_{kD+2}, ...H_{kD+D})$, where each H_{kD+i} represents the usual $N \times M$ channel matrix.

For the input power constraint, we typically choose to work under the peak power constraint as normally communication systems are peak-power limited in practice. Thus power at all transmitting antennas can never exceed SNR, the peak power, thus

$$X_k^{\dagger} X_k \ \leq \ SNR \quad . \tag{4}$$

Throughout this paper, $(.)^T$ and $(.)^{\dagger}$ will denote transpose and Hermitian transpose operators respectively.

The capacity pre-log is normally defined as

$$PreLog = \lim_{SNR \to \infty} \frac{C(SNR)}{log(SNR)} \quad . \tag{5}$$

We define a new capacity parameter which may help us better understand the asymptotic capacity reduction when CSIR is not available. It is called Asymptotic Capacity Reduction Factor (ACRF) and is defined as

$$ACRF = \lim_{SNR \to \infty} \frac{C_{NO-CSIR}(SNR)}{C_{CSIR}(SNR)} \quad . \tag{6}$$

Thus the ACRF is the ratio of non-coherent capacity to coherent capacity at very high values of SNR.

3. NON-COHERENT CAPACITY BOUNDS

For our MIMO system in equation (1), the capacity can be calculated from the well-known expression

$$C = \lim_{n \to \infty} \frac{1}{n} \sup_{p_{X^n}} I(X^{1:n}; Y^{1:n}) \tag{7}$$

where the maximization is done over all input distributions which satisfy the power constraint. The mutual information in the above expression can be decomposed as follows

$$I(X^{1:n}; Y^{1:n}) = I(X_d^{1:n}, X_t^{1:n}; Y_d^{1:n}, Y_t^{1:n})$$
$$= \underbrace{I(X_t^{1:n}; Y_d^{1:n}, Y_t^{1:n})}_{I_1} + \underbrace{I(X_d^{1:n}; Y_d^{1:n}, Y_t^{1:n} | X_t^{1:n})}_{I_2} \tag{8}$$

The subscripts t and d denote "training" and "data" respectively, and superscript $1 : n$ shows that the length of the sequence ranges from 1 to n. Training and data here can be time multiplexed (in which case also the outputs get time multiplexed) or superimposed or a combination of both. Or training and data can more generally live in two complementary subspaces. The term "training" here may be misleading. Indeed, also the "training" symbols carry information. Nevertheless, apart from data transmission they also allow the channel to be estimated, with channel estimates that serve as a basis for the complementary "data" symbols. To diminish the confusion, we shall instead call these "training" symbols "learning" symbols. In suboptimal approaches, these learning symbols may get replaced by classical training symbols.

3.1. Capacity Lower Bound

For the lower bound on the capacity, we can consider $X_t^{1:n}$ as pure training sequence, so information I_1 in equation (8) goes to zero. But this known training sequence allows channel estimation with finite estimation error covariance so that the effect of this channel estimation error, when the channel estimate gets used in the data part (and no other information gets used for the estimation of the channel), is at worst a finite increase of the effective noise power. Hence the difference of I_2 in equation (8) from the full CSIR case is at most some finite constant. Thus the pre-log of I_2 is that of the full CSIR case, but there is a reduction factor due to the presence of training which leads to

$$ACRF = 1 - \frac{training \quad size}{training \quad size + data \quad size}$$

where the "size" should be interpreted as the number of dimensions of the corresponding subspace. To achieve the full CSIR pre-log in data part, the training length should be sufficient to allow deterministic identifiability of the channel, meaning that if the channel is considered as a deterministic signal, it should be identifiable with zero error in the absence of noise. Hence we get for the overall capacity:

$$ACRF \;\geq\; 1 - \frac{learning \quad size}{learning \quad size + data \quad size}$$

3.2. Capacity Upper Bound

For the upper bound on the channel capacity, we cannot ignore the mutual information associated to learning part $X_t^{1:n}$. Now, since X_t and Y_t live in corresponding subspaces whereas Y_d lives in an orthognal subspace, $I_1 = I(X_t^{1:n}; Y_d^{1:n}, Y_t^{1:n}) = I(X_t^{1:n}; Y_t^{1:n})$. Now, as long as the the size of the learning part is not more than the smallest possible size that allows channel identifiability, we are in the "regular" case of Lapidoth [5] and the capacity I_1 grows with SNR at most as $(learning\ size) \times \log\log(SNR)$, hence its pre-log is zero. For an upper bound on I_2, we can just take the full CSIR assumption leading to a capacity growth with a pre-log equal to the data size. As a result we get for the overall capacity:

$$ACRF \;\leq\; 1 - \frac{learning \quad size}{learning \quad size + data \quad size}.$$

Combining lower and upper bounds, we get equality for ACRF.

Finer Analysis of I_2: a decomposition leads to

$$I_2 = \sum_{i=1}^{n} I(X_d^i; Y_d^{1:n}, Y_t^{1:n} | X_t^{1:n}, X_d^{1:i-1}) \qquad (9)$$

(this decomposition is not necessarily in time, it can also be along a subspace basis). This sum term indicates that for the detection of each of the data symbols in SISO or vector X_d in MIMO case, we can use the channel estimate from the learning part and all previously detected input symbols. Furthermore the presence of all output symbols asserts the need to do blind channel estimation to fully exploit the information present in that term as discussed in [9]. This indicates that to get the actual capacity, and in particular the proper constant term at high SNR, one needs to perform semi-blind channel estimation within the data subspace, since based on past inputs and outputs and future outputs. The future output may give important channel information, especially in the multiple receive antenna case. In any case, whether the channel estimate is based on the "learning" input and output only or whether it is based on full semiblind information does not change the pre-log factor, but only an additive constant in the asymptotic capacity. So, to summarize, in the no CSIR case, the input can be split into a "learning" subspace and its orthogonal complement, the "data" subspace. The "learning" subspace is of minimal dimensions just to allow deterministic identifiability of the channel and hence corresponds to the "regular" case in Lapidoth's terminology and carries information of the order of log[log(SNR)]. The "data" subspace is the main subspace for transmission of data and its reduced dimension represents the reduced pre-log factor.

4. SISO CAPACITY PRE-LOG ACHIEVING SCHEMES

In this section, we give two schemes which show us the capacity limit in high SNR regime and even enable us to achieve the capacity pre-log.

4.1. Sub Sampling Approach

As Doppler spectrum is band-limited to 1/D, so we can downsample with the integer downsampling factor D according to Nyquist's theorem. Thus we get a grid as shown in the figure 1. Over the downsampled instants, we transmit learning

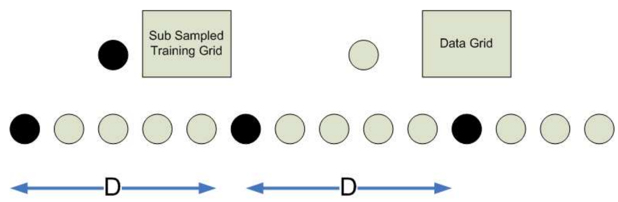

Fig. 1. Subsampling Grid.

symbols (either low rate or known to the receiver) and rest are the data symbols. So there is one learning symbol after each (D-1) data symbols. Thus over a block of D symbols, we have D prediction problems, (D-1) of which are singular, i.e. the prediction error will go to zero in the absence of noise and D-th prediction error is a white noise at sub-sampled rate (1/D).

$$\sigma^2 = \exp^{\int_{-1/2D}^{1/2D} lnS_{hh}(f)df} \quad (10)$$

The above is the classical result for the prediction error variance of a process in terms of its spectral description, where $S_{hh}(f)$ represents the power spectral density of the discrete time fading process and is equal to the fourier transform of the autocorrelation function of the channel fading process. Originally spectrum is bandlimited but when we downsample with the factor D (the learning grid), the spectrum becomes non-bandlimited over the learning grid giving non-zero prediction error.

Channel estimates on learning grid may be obtained by causal linear prediction and for the data grid they can be obtained by non-causal LMMSE Wiener filtering. Because of the presence of additive noise, prediction will not be perfect. The error in channel estimation has its worst effect when it is white, so in this case it gets added up with the white noise already present. This reduces the effective SNR at the receiver and causes a shift in the curve of capacity versus SNR but the slope of this curve remains unchanged corresponding to capacity pre-log.

This subsampling approach makes causal estimates over the learning grid and for data grid, channel estimate corresponding to each data symbol is obtained by non-causal estimates over the learning grid and causal estimates over the previously detected symbols.

4.1.1. *Capacity in Learning Grid*

About these learning symbols in the grid, we don't specify them to be perfectly known to receiver before transmission. They may be learning symbols in the true sense that they are known to the receiver before transmission or they may be coming from a low rate stream which allows decoding even in the absence of CSIR. If they are pure training, mutual information over this learning grid is zero, but if they are data symbols, communication over the learning grid becomes like communication over a non-bandlimited channel so there is no growth with log(SNR) over this grid. Capacity growth with log[log(SNR)] for non-bandlimited case has been shown in [5] and fading number (the constant term accompanying log[log(SNR)] has been calculated in [10], [5] and [6].

At very high SNR, noise at the receiver can be neglected and input output relation over this learning grid can be represented as

$$y = hx \quad (11)$$

where we have decided to drop the indices. This system can be divided into 'norm system' and 'direction system'.

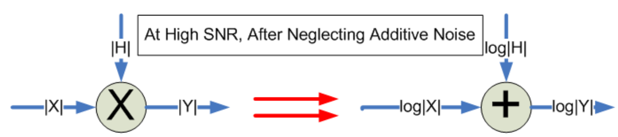

Fig. 2. Multiplicative channel to Additive channel

If we represent $y = |y|\exp(j\theta_y)$ where $|y|$ denotes magnitude of complex output y and θ_y denotes its direction(phase) and same holds true for x and h, then

$$y = hx \Rightarrow \begin{cases} \text{norm system:} & |y| = |h|\,|x| \\ \text{direction system:} & \theta_y = \theta_h + \theta_x \end{cases}$$

The norm system after taking logarithm converts into an additive channel whose capacity is well known.

$$\log|y| = \log|h| + \log|x| \quad (12)$$

The mutual information of the direction system is finite whatever is the distribution of channel phase and is strictly zero when channel phase is uniformly distributed from $-\pi$ to $+\pi$. Even in this case, it helps estimating the channel and pass this channel information to data grid. On the other hand norm system is responsible for the log[log(SNR)] growth of channel capacity over the learning grid. Because log(SNR) is the high snr capacity of additive noise channel but in the norm system we already have logarithm of the input which causes the capacity to grow double logarithmically. So at very high SNR, per symbol capacity in learning part is

$$C_L = \log\log(SNR) + \chi(\{h\}) + o(SNR) \quad (13)$$

$\chi(\{h\})$ is termed fading number in [5] and $o(SNR)$ terms goes to zero as SNR goes to ∞.

4.1.2. *Capacity in Data Grid*

We know that learning grid may have its capacity growth like log[log(SNR)], but once detected, these symbols act as training symbols for the data grid. Either way, whether learning is pure training or low rate stream, it will give capacity pre-log of zero. But channel estimates formed with the help of learning grid make possible the coherent detection of symbols at data grid. Although there will be a penalty over the effective SNR, but growth rate will be with log(SNR) over the data grid. So per symbol capacity over the data grid is

$$C_D = \log(SNR) + \alpha \tag{14}$$

where α is some constant which doesn't depend upon SNR.

4.1.3. *Capacity For Sub Sampling Scheme*

Thus in a straightforward manner, over each grid of D symbol times, (D-1) form data part and have capacity growth with log(SNR) and 1 symbol forms learning grid. So capacity for this scheme can be characterized to be

$$C_{SISO} = (1 - \frac{1}{D})[\log(SNR) + \alpha] + \frac{1}{D}[\log\log(SNR) + \chi(\{h\})] + o(SNR) \tag{15}$$

Thus for SISO systems

$$PreLog = ACRF^{SISO} = (1 - 1/D) \tag{16}$$

4.2. Learning and Data Subspaces

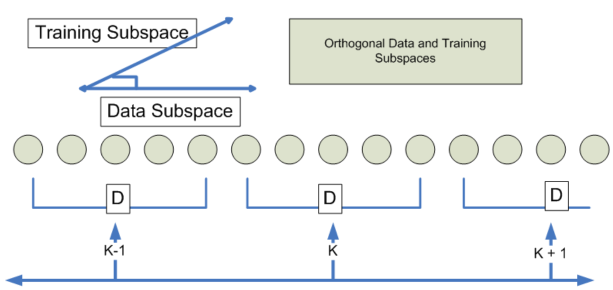

Fig. 3. Learning and Data Subspaces

We can vectorize our channel with D elements in each vector. Corresponding to this vector channel, input vector X may have D dimensioinal subspace.

$$\underline{x}_k = [x_{kD+1} \;\; x_{kD+2} \;\; \ldots \;\; x_{kD+D}]^T$$

Similarly $\underline{y}_k$ and $\underline{z}_k$ represent vectorized output and noise samples corresponding to the k-th block. But fading values for this block are represented as the diagonal elements of the matrix with zero off diagonal values

$$\underline{h}_k = Diag[h_{kD+1} \;\; h_{kD+2} \;\; \ldots \;\; h_{kD+D}]^T$$

We could make an arrangement so as to use one dimension for learning and rest of (D-1) dimensions as data. The one dimensional subspace used by the learning should have its projection orthogonal to the projection of (D-1) dimensional subspace used by data. If we put power constraint over input vector $\underline{x}$, then we need to optimize the power between data and learning part but we put the constraint separately over both so $\underline{x}$ is also peak power contrained.

$$\underline{x} = [A_t \quad A_d] \begin{bmatrix} \underline{x}_t \\ \underline{x}_d \end{bmatrix} \tag{17}$$

where A_t and A_d are special matrices such that their projections are orthogonal, i.e. $[A_t \quad A_d]$ is unitary.

$$P_{A_t} = P^{\perp}_{A_d} \tag{18}$$

where $P_A = A(A^{\dagger}A)^{\#}A^{\dagger}$, and $(.)^{\#}$ denotes Moore-Penrose pseudo-inverse. So the received signal is

$$\underline{y} = \underline{h}A_t\underline{x}_t + \underline{h}A_d\underline{x}_d + \underline{z} \tag{19}$$

Receiver can recover learning symbols as the subspaces spanned by learning and data are orthogonal. Here again, the learning sub-space may have true training symbols or they may be coming from a low rate stream. But the situation is same as it was in the subsampled scheme. Continuously we have $1/D$ resource usage as a training or low rate data transmission which gives us capacity pre-log of zero, but for the rest $1 - 1/D$ resource, channel estimates are available from the learning sub-space so communication becomes coherent for this resource and we get capacity increase of log(SNR), and hence capacity pre-log and ACRF both are equal to $1 - 1/D$. There is again reduction in effective SNR because of noisy channel estimation causing shift in the capacity curve but leaving the capacity pre-log unharmed.

For this scheme, channel estimation is purely causal. Channel estimate at each symbol instant is obtained by previously received learning symbols and previously detected data symbols. Thus it differs from subsampling approach due to its causal functionality.

5. MIMO CAPACITY PRE-LOG ACHIEVING SCHEMES

In this section, we give the MIMO extensions to the schemes we proposed for SISO case in the previous section.

5.1. Sub Sampling Approach

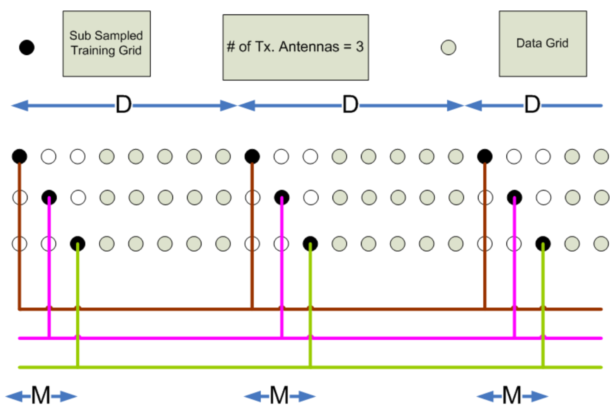

Fig. 4. MIMO Subsampling Grid.

With the same reasoning as in the SISO case, as Doppler spectrum is band-limited to 1/D for each channel entry, we downsample with the integer downsampling factor D according to Nyquist's theorem. But as it was shown by Hassibi [11], to properly estimate the MIMO channel matrix, we need learning length in symbol periods equal to the number of transmit antennas. In fact in one symbol period, even with perfectly known input data, only the projection of the channel matrix on the input can be estimated, i.e. only one column of the channel matrix. Thus to estimate the whole channel matrix in a group of D transmissions, we need to transmit learning for M symbol times (corresponding to M columns in channel matrix). This points to a very important fact for MIMO channels that MIMO case stays regular (non-zero prediction error with infinite past) as long as $M \times DopplerBW \geq 1$.

Thus in the very beginning, there is a multiplicative reduction factor of $(1 - M/D)$ with the capacity as in a group of D transmissions, M carry only the learning symbols. But these transmissions allow us to make channel estimates. Channel estimates on this learning grid may be obtained by causal linear prediction and for the data grid they can be obtained by non-causal LMMSE Wiener filtering. In fact capacity decomposition done in section 3 shows us that this channel estimation should be based upon all learning symbols and previously detected data symbols.

5.1.1. *Capacity in Learning Grid*

As explained above, communication over the learning grid behaves like communication over a non-bandlimited channel so there is no growth with log(SNR) over this grid. Capacity growth with log[log(SNR)] has been shown in [5] and fading number (the constant term accompanying log[log(SNR)] has been calculated in [10], [5] and [6].

At very high SNR, noise at the receiver can be neglected and input output relation over this learning grid can be represented as

$$Y = HX \tag{20}$$

Like we did for SISO systems, this system can be divided into 'norm system' and 'direction system'.

$Y = \overline{Y}||Y||\,,\ X = \overline{X}||X||$.

where we represent $\overline{F}$ as the unit norm vector of vector F.

$$Y = HX \Rightarrow \begin{cases} \text{norm system:} & ||Y|| = ||H\overline{X}||\,||X|| \\ \text{direction system:} & \overline{Y} = H\overline{X}/||H\overline{X}|| \end{cases}$$

Again multiplicative channel of 'norm system' can be converted into an additive noise channel by taking the logarithm of both sides of the above 'norm system' equation which gives us

$$\log ||Y|| = \log ||H\overline{X}|| + \log ||X|| \tag{21}$$

Like in SISO systems, this 'norm system' is responsible for double logarithmic growth of capacity. All degrees of freedom collapse in this case over learning grid and coefficient of log[log(SNR)] is '1' even for MIMO [5]. On the other hand, 'direction system' has zero capacity if the discrete time channel coefficients are spatially i.i.d. (independent and identically distributed). Even if they are not i.i.d., capacity of direction system is finite and has no scaling with SNR. Hence per symbols capacity in learning grid is

$$C_L = \log\log(SNR) + \chi(\{H\}) + o(SNR) \tag{22}$$

$\chi(\{H\})$ is the fading number for this matrix valued fading process [5] and $o(SNR)$ terms goes to zero as SNR goes to ∞.

5.1.2. *Capacity in Data Grid*

Learning grid may carry pure training or low rate stream, it will give capacity pre-log of zero. But channel estimates formed with the help of learning grid make possible the coherent detection of symbols at data grid. Thus full multiplexing gain of $min(M, N)$ can be exploited in the data part. Although there will be a penalty over the effective SNR, but growth rate will be with $min(M, N)\log(SNR)$ over the data grid. So per symbol capacity over the data grid is

$$C_D = min(M, N)\log(SNR) + \alpha \tag{23}$$

where α is some constant which doesn't depend upon SNR.

5.1.3. Capacity For MIMO Sub Sampling Scheme

Combining the mutual information obtainable in both the learning part and data part, where learning part prevails M symbol times and data part for (D-M) symbol times over each group of D symbols,

$$C_{MIMO} = (1 - \frac{M}{D})[min(M,N)\log(SNR) + \alpha] + \frac{M}{D}[\log\log(SNR) + \chi(H)] + o(SNR) \tag{24}$$

Thus capacity pre-log for this MIMO system turns out to be $min(M,N) * (1 - M/D)$ and $ACRF^{MIMO} = (1 - M/D)$. ACRF clearly shows that Doppler bandwidth gets multiplied by the number of antennas, causing greater capacity reduction as compared to SISO systems.

5.2. Learning and Data Subspaces

For this approach, we consider a block of D symbol periods. Hence system model is the one represented in equation (3). If we break the input in 'learning' and 'data' then

$$\underline{X} = [A_t \quad A_d] \begin{bmatrix} \underline{X}_t \\ \underline{X}_d \end{bmatrix} \tag{25}$$

where underlined letters show that we are working over blocks of D-symbol period and A_t and A_d are special matrices such that their projections are orthogonal, i.e. $[A_t \quad A_d]$ is unitary. The received signal can be expressed as

$$\underline{Y} = \underline{H}A_t\underline{X}_t + \underline{H}A_d\underline{X}_d + \underline{Z} \tag{26}$$

The matrices A_t and A_d can further be expressed as

$$A_t = \underbrace{[I_M \quad 0]^T}_{B_t} \otimes I_M \qquad A_d = \underbrace{[0 \quad I_{D-M}]^T}_{B_d} \otimes I_M \tag{27}$$

Furthermore these can be expressed as

$$C_t = B_t \otimes I_N \qquad C_d = B_d \otimes I_N \tag{28}$$

Then at the receiver side, because of orthogonal projections learning and data dimensions can be separated.

$$\underline{Y}_t = C_t^\dagger \underline{Y} = C_t^\dagger \underline{H}A_t\underline{X}_t + C_t^\dagger \underline{Z} \tag{29}$$

$$\underline{Y}_d = C_d^\dagger \underline{Y} = C_d^\dagger \underline{H}A_d\underline{X}_d + C_d^\dagger \underline{Z} \tag{30}$$

Input vector $\underline{X}$ for D symbol times will have length $M \times D$ and will span $M \times D$ dimensional subspace as we mentioned in our D-symbol model in equation (3). Now out of these $M \times D$ dimensions available, we use $M \times M$ dimensional subspace for learning which will enable us to have sufficiently accurate channel estimation because at receiver from $\underline{Y}_t$ channel estimates can be obtained and rest of $M \times (D - M)$ dimensional subspace will carry data.

If we represent the channel coefficients for D-symbol block in one long vector represented as $\underline{h_{k'}}$, hence having dimension $M \times N \times D$, it can be represented as MIMO auto-regressive (AR) model of infinite order

$$\underline{h_{k'}} = \sum_{i=1}^{\infty} U_i \underline{h_{k'-i}} + \underline{\tilde{h}_{k'}} \tag{31}$$

The covariance matrix R of prediction error $\underline{\tilde{h}_{k'}}$ will have its rank $M \times M$.

The $M \times M$ dimensional subspace used by learning will be orthogonal to $M \times (D - M)$ dimensional subspace used by data. In case of no noise, it will allow perfect estimation of channel, but noise presence gives error in estimation. But transmission over the data subspace behaves like coherent communication. Thus learning subspace will introduce a capcity reduction factor $ACRF^{MIMO} = (1 - M/D)$ and we will get the capacity pre-log of $min(M,N) * (1 - M/D)$ from the data subspace.

We have seen in MIMO case that $ACRF^{MIMO} = (1 - M/D)$. Thus we remark that Doppler bandwidth gets multiplied with the number of transmitting antennas causing more capcity reduction.

6. OPTIMAL NUMBER OF TX. ANTENNAS

We have shown in the previous sections of this paper that optimal pre-log for multiple-transmit and multiple-receive antenna system is given by the expression

$$MIMO\ PreLog = min(M,N) * (1 - M/D) \tag{32}$$

This pre-log expression indicates that sometimes it might be sub-optimal to use all of the available transmitting antennas. Because learning length over which we obtain capacity scaling like log[log(SNR)] is proportional to product of number of trasmit antennas (M) and Doppler's bandwidth (1/D).

If Doppler's bandwidth (1/D) is very small i.e. D is sufficiently larger than M and N, then pre-log $min(M,N) * (1 - M/D)$ shows that optimal number of transmit antennas should be $min(M,N)$ because it will reduce M/D loss factor to $min(M,N)/D$.

But if Doppler's bandwidth is not very small i.e. D is comparable to M and N and assuming $M < N$ pre-log factor is $M' * (1 - M'/D)$, where M' denotes the number of active transmitting antennas and $M' \leq M$. Now we can find the optimal number of transmitting antennas to achieve this capacity pre-log by differentiating the pre-log w.r.t. M', the number of active transmitting antennas.

$$\frac{d}{dM'}M' * (1 - M'/D) = 1 - \frac{2M'}{D} \tag{33}$$

Just equating the above expression to zero and solving for M' gives us the result of $M' = \frac{D}{2}$. So combining with the fact that $M' = min(M, N)$ also holds, we get the following result

$$Optimal\ Tx.\ Antennas = M' = min(M, N, \frac{D}{2}) \tag{34}$$

Remark 1: Intuitively the optimal input for a stationary channel should be stationary. Apparently our approach with learning and data grids seems block-stationary or cyclostationary. But this input can be stationarized by introducing a uniform time offset over D-symbol block.

Remark 2: We took Doppler bandwidth to be $1/D$. This can be generalized to the rational case of p/q. Vectorizing the stationary channel over p will lead to a stationary vector process with matrix spectrum of rank q for SISO and infinite order prediction error covariance matrix will also be of rank q.

7. CONCLUDING REMARKS

We have shown the capcity pre-log for SISO and MIMO channels. Moreover our presented two schemes help achieve this pre-log in practical systems.

A striking observation is that the systems with M transmit antennas should be varying M times more slowly as compared to the systems with only one transmit antenna in order to have comparable capacity reduction in case of no CSIR.

We also characterize the optimal number of active trasmitting antennas in terms of Doppler bandwidth of the channel fading process to achieve optimal pre-log at high values of SNR.

8. REFERENCES

[1] A. Lapidoth, "On the asymptotic capacity of stationary gaussian fading channels," *IEEE Trans. on Information Theory*, vol. 51, pp. 437–446, February 2005.

[2] T. Marzetta and B. Hochwald, "Capacity of a mobile multiple-antenna communications link in rayleigh flat fading," *IEEE Trans. on Information Theory*, vol. 45, pp. 139–157, January 1999.

[3] L. Zheng and D. N. C. Tse, "Communication on the grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. on Information Theory*, vol. 48, pp. 359–383, February 2002.

[4] Y. Liang and V. V. Veeravalli, "Capacity of noncoherent time-selective rayleigh-fading channels,," *IEEE Trans. on Information Theory*, vol. 50, pp. 3095–3110, December 2004.

[5] A. Lapidoth and S. Moser, "Capacity bounds via duality with applications to multi-antenna systems on flat fading channels," *IEEE Trans. on Information Theory*, vol. 49, pp. 2426–2467, October 2003.

[6] T. Koch and A. Lapidoth, "The fading number and degrees of freedom in non-coherent mimo fading channels: a peace pipe," *IEEE International Symposium on Information Theory*, vol. 52, pp. 437–453, September 2005.

[7] A. Lapidoth and S. Moser, "The fading number of single-input multiple-output fading channels with memory," *IEEE Trans. on Information Theory*, vol. 52, pp. 437–453, February 2006.

[8] J. Chen and V. V. Veeravalli, "Capacity results for block-stationary gaussian fading channels,," *IEEE International Symposium on Information Theory*, vol. 50, pp. 3095–3110, July 2006.

[9] M. Abdelkader and D. Slock, "Mutual information without channel knowledge at the receiver," *SPAWC 2003, 4th IEEE Workshop on Signal processing advances in wireless communications, June 15-18, 2003 - Rome, Italy*.

[10] S. Moser, "The fading number of multiple-input multiple-output fading channels with memory," *IEEE International Symposium on Information Theory*, pp. 521–525, June 2007.

[11] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?,," *IEEE Transactions on Information Theory*, vol. 49, pp. 2058–2080, April 2003.