MUTUAL INFORMATION WITHOUT CHANNEL KNOWLEDG AT THE RECEIVER

Abdelkader Medles, Dirk T.M. Slock

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

ABSTRACT

In this paper, we analyze the semiblind mutual information (MI) between the input and output of a MIMO channel. To that end we consider the popular block fading model. We assume that some training/pilot symbols get inserted at the beginning of each burst. We show that the average MI over a transmission burst can be decomposed into symbol position dependent contributions. The MI component at a certain symbol position optimally combines semiblind information up to that symbol position (with perfect input recovery up to that position) with blind information from the rest of the burst. We also analyze the asymptotic regime for which we can formulate optimal channel estimates and evaluate the capacity loss with respect to the known channel case. Asymptotically, the decrease in MI involves Fisher Information matrices for certain channel estimation problems. We also suggest to exploit correlations in the channel model to improve estimation performance and minimize capacity loss.

1. INTRODUCTION

We consider single user spatial multiplexing systems over flat fading MIMO channels. We furthermore consider the usual block fading model in which data gets transmitted over a number of bursts such that the channel is constant over a burst but fading independently between bursts. The transmitter is assumed to have no channel knowledge. The formidable capacity increase realizable with such dual antenna array systems has been shown [5],[2] to be proportional to the minimum of the antenna array dimensions for channel with i.i.d. fading entries. At least, this is the case when the receiver has perfect channel knowledge. This condition is fairly straightforward to approach in SISO systems by inserting pilot/training data in the transmission, with acceptable capacity decrease [1]. For MIMO systems of large dimensions though, the training overhead for good channel estimation quality becomes far from negligible though, especially for higher Doppler speeds such as in mobile communications. The effect of channel estimation errors on the MI has been analyzed in [4], whereas optimal design of training based channel estimation has been addressed in [6]. The true capacity of the system is higher though than the MI obtained by training based channel estimates [2],[3]. In this paper, we attempt to approach the true capacity by optimal semiblind channel estimation that is suggested by a decomposition of the MI. Related background material appeared in [7] where a variety of semiblind MIMO channel estimation techniques were introduced.

Eurecom's research is partially supported by its industrial partners: Ascom, Swisscom, Thales Communications, ST Microelectronics, CEGETEL, Motorola, France Tlcom, Bouygues Telecom, Hitachi Europe Ltd. and Texas Instruments. The work reported herein was also partially supported by the French RNRT project ERMITAGES.

2. CAPACITY DECOMPOSITION

We shall consider here the usual block fading model, except that we shall refer to a block as burst; consider then transmission over a MIMO flat fading channel $y = Hx + v$, for a particular burst of T symbol periods. The accumulated received signal over the burst is then:

$$Y = (I_T \otimes H)\ X + V \tag{1}$$

where Y and V are $N_r T \times 1$ and X is $N_t T \times 1$. H is a $N_r \times N_t$ channel matrix. N_t (resp. N_r) is the number of transmit (resp. receive) antennas. We assume the use of a pilot training sequence of length $N_t T_p$, the length of the transmitted data is then $N_t T_d$ so that $T_p + T_d = T$. We can decompose the burst signal into training and data parts $X = (X_p^T, X_d^T)^T, Y = (Y_p^T, Y_d^T)^T$ and $V = (V_p^T, V_d^T)^T$. As stated in [6], the mutual information between the input and the output of this system verifies:

$$\begin{aligned} I(Y_p, Y_d; X_d|X_p) &= I(Y_d; X_d|X_p, Y_p) + I(Y_p; X_p|X_d) \\ &= I(Y_d; X_d|X_p, Y_p) \end{aligned} \tag{2}$$

$I(Y_p, X_p; X_d) = 0$ due to the independence between X_d and (Y_p, X_p). Consider now a partition of X_d in Q blocks X_i, $i = 1, \ldots, Q$, $X_d = (X_1^T, \ldots, X_Q^T)^T$, of different lengths $T_i, i = 1, \ldots, Q$, $\sum_i T_i = T_d$. X_i and V_i are assumed independent from block to block (block-wise coding across bursts). Let's define for,
$j \geq i$, $X_i^j = (X_i^T, X_{i+1}^T, \ldots, X_j^T)^T$
and $Y_i^j = (Y_i^T, X_{i+1}^T, \ldots, Y_j^T)^T$. Then,

$$\begin{aligned} &I(Y_d; X_d|X_p, Y_p) = I(Y_1^Q; X_1, X_2^Q|X_p, Y_p) \\ &= I(Y_1^Q; X_1|X_p, Y_p) + I(Y_1^Q; X_2^Q|X_p, Y_p, X_1) \\ &= I(Y_1^Q; X_1|X_p, Y_p) + I(Y_2^Q; X_2^Q|X_p, Y_p, X_1, Y_1) \\ &+ \underbrace{I(Y_1; X_2^Q|X_p, Y_p, X_1)}_{=0} \end{aligned} \quad (3)$$

where $I(Y_1; X_2^Q|X_p, Y_p, X_1) = h(Y_1|X_p, Y_p, X_1) - h(Y_1|X_p, Y_p, X_1, X_2^Q)$, and considering the fact that X_2^Q is independent of Y_1 conditioned on (X_p, Y_p, X_1), this leads to $h(Y_1|X_p, Y_p, X_1, X_2^Q) = h(Y_1|X_p, Y_p, X_1)$ and finally $I(Y_1; X_2^Q|X_p, Y_p, X_1) = 0$. Iterating the equation for $i = 2, \ldots, Q$, this leads to the following:

$$\begin{aligned} I(Y_d; X_d|X_p, Y_p) &= \textstyle\sum_{i=1}^{Q} I(Y_i^Q; X_i|X_p, Y_p, X_1^{i-1}, Y_1^{i-1}) \\ &= \underbrace{\sum_{i=1}^{Q} I(Y_i; X_i|X_p, Y_p, X_1^{i-1}, Y_1^{i-1})}_{I1} \\ &+ \underbrace{\sum_{i=1}^{Q-1} I(Y_{i+1}^Q; X_i|X_p, Y_p, X_1^{i-1}, Y_1^{i})}_{I2} \end{aligned}$$

This clearly shows the way of processing to achieve capacity: for every block we use the already detected blocks as a (data-aided (DA)) training sequence (in addition to the actual training) and use the not yet detected blocks as blind information. Also, the mutual information can be seen as the sum of two parts: $I(Y_d; X_d|X_p, Y_p) = I1 + I2$ where:
$I1$: is the rate that we can achieve by using only the already treated blocks for side information (DA training).
$I2$: is the additional amount of rate that can be achieved by exploiting the blind information contained in the not yet detected blocks.
The average mutual information is defined as $I_{avg} = \frac{1}{T} I(Y_d; X_d|X_p, Y_p)$. We want to show below that this quantity goes to the coherent MI $I(y; x|H)$ as T and Q grow for finite T_p, where $I(y; x|H) = \lim_{T\to\infty} \frac{1}{T} I(Y; X|H) = \lim_{i\to\infty} \frac{1}{T_i} I(Y_i; X_i|H)$. An upper bound is given by:

$$\begin{aligned} I_{avg} &= \tfrac{1}{T}(h(X_d) - h(X_d|X_p, Y_p, Y_d)) \\ &\leq \tfrac{1}{T}(h(X_d) - h(X_d|X_p, Y_p, Y_d, H)) \\ &= \tfrac{1}{T}(h(X_d) - h(X_d|Y_d, H)) \\ &= \tfrac{1}{T}(I(Y; X|H) - I(Y_p; X_p|H)) \\ &\Rightarrow \lim_{T\to\infty} I_{avg} \leq I(y; x|H) \end{aligned} \quad (4)$$

Also

$$\begin{aligned} &I(Y_i^Q; X_i|X_p, Y_p, X_1^{i-1}, Y_1^{i-1}) \\ &= h(X_i) - h(X_i|X_p, Y_p, X_1^{i-1}, Y_1^{i-1}, Y_{i+1}^Q, Y_i) \\ &\geq h(X_i) - h(X_i|Y_i, \widehat{H}(X_p, Y_p, X_1^{i-1}, Y_1^{i-1}, Y_{i+1}^Q, Y_i)) \\ &= h(X_i) - h(X_i|Y_i, \widehat{H}(X_p, X_1^{i-1}, \underbrace{Y_p, Y_1^{i-1}, Y_{i+1}^Q}_{=\overline{Y}_i})) \\ &= I(Y_i; X_i|\widehat{H}^{(i)}) \end{aligned} \quad (5)$$

where $\widehat{H}^{(i)} = \widehat{H}(X_p, X_1^{i-1}, \overline{Y}_i)$ is the optimal estimate of the channel (statistic of reduced dimension), that verifies $\lim_{i\to\infty} \widehat{H}^{(i)} = H$ $a.s.$ Given that, $\lim_{i\to\infty} \frac{1}{N_i} I(Y_i^Q; X_i|X_p, Y_p, X_1^{i-1}, Y_1^{i-1}) \geq I(y; x|H)$. This allows us to conclude that $\lim_{T\to\infty} I_{avg} \geq I(y; x|H)$. Combining this result with (4) we conclude that: $\lim_{T\to\infty} \frac{1}{T} I(Y_d; X_d|X_p, Y_p) = I(y; x|H)$. This means that when T grows, adopting the detection method per block, and the associated channel estimation allows to reach the capacity of the system assuming perfect channel knowledge (see [3] for related results at high SNR).
Remark1: When T grows, the use of detected blocks only to estimate the channel allows to achieve asymptotically the MI I_{avg} of the system. But for finite T, it is necessary to also use the blind information to reach it.
Remark2: For a fixed T, and when all the entries of X (training and data) are iid, it's easy to see that the average MI I_{avg} of the system is maximized when the number of the training symbols T_p is as small as possible (i.e. allows semiblind identifiability of the channel).
Remark3: In the case where H is no longer constant, and varies from block to block $H = H^{(i)}$, the (coherent) capacity of the system assuming perfect channel knowledge is no longer achievable for large T and the average MI is bounded by:

$$\frac{1}{T}\sum_{i=1}^{Q} I(Y_i; X_i|\widehat{H}_i) \leq I_{avg} \leq \frac{1}{T}\sum_{i=1}^{Q} I(Y_i; X_i|H_i) \quad (6)$$

for any channel estimate $\widehat{H}_i = \widehat{H}_i(X_p, X_1^{i-1}, \overline{Y}_i)$.

3. CHANNEL ESTIMATION

3.1. Bayesian case (random channel with prior)

The capacity C of the system is the maximum MI I_{avg} over all input distributions, under a given power constraint. We have

$$\begin{aligned} &\tfrac{1}{T}\textstyle\sum_{i=1}^{Q} I(Y_i; X_i|\widehat{H}^{(i)}) \leq C \leq \\ &\max_{p(X_d), tr(R_{X_d}) \leq T_d N_t \sigma_x^2} \frac{1}{T}\sum_{i=1}^{Q} I(Y_i; X_i|H) \,. \end{aligned} \quad (7)$$

This is valid for all choices of the partitioning of X_d into $X_i, i = 1, \ldots, Q$, and in particular for $Q = T_d$ and $T_i \equiv 1$. For an AWGN V with power σ_v^2 and in the absence of side information on the channel at the transmitter (see [5]), the max in the upper bound of the capacity (coherent capacity) is attained for a centered white Gaussian input with covariance $R_{X_d} = \sigma_x^2 I_{N_t T_d}$.

$$\frac{1}{T}\sum_{i=1}^{T_d} I(Y_i;X_i|\widehat{H}^{(i)}) \le C \le \frac{T-T_p}{T} E\ln\det(I+\frac{\sigma_x^2}{\sigma_v^2}H\,H^H) \quad (8)$$

where $\widehat{H}^{(i)} = \widehat{H}(X_p, X_1^{i-1}, \overline{Y}_i)$.
The received signal is $Y_i = H\,X_i + V_i = \widehat{H}^{(i)}\,X_i + \widetilde{H}^{(i)}\,X_i + V_i = \widehat{H}^{(i)}\,X_i + V_i + Z_i$, where we assume $\widehat{H}^{(i)}$ to satisfy the Pythagorean Theorem (PT) (i.e. $\widehat{H}$ is decorrelated with $\widetilde{H}$). Now $I(Y_i,X_i|\widehat{H}^{(i)}) = h(Y_i|\widehat{H}^{(i)}) - h(Y_i|X_i,\widehat{H}^{(i)}) = h(Y_i|\widehat{H}^{(i)}) - h(V_i+Z_i|X_i,\widehat{H}^{(i)})$. Under the above conditions and for uncorrelated and centered Gaussian V_i and $X_i = X_i^G$ (variable with same 1st and 2nd order moments, but Gaussian), it was shown in [6] that a lower bound for $I(Y_i,X_i|\widehat{H}^{(i)})$ is given by considering Z_i as an independent and white Gaussian noise, with covariance $\sigma_z^2\,I$, where $\sigma_z^2 = \frac{1}{N_r}tr\ \mathrm{E}(Z_iZ_i^H) = \sigma_x^2\frac{tr\ \mathrm{E}(\widetilde{H}^{(i)}\widetilde{H}^{(i)\,H})}{N_r} = N_t\sigma_x^2\sigma^2_{\widetilde{H}^{(i)}}$ and $\sigma^2_{\widetilde{H}^{(i)}} = \frac{tr\ \mathrm{E}(\widetilde{H}^{(i)}\widetilde{H}^{(i)\,H})}{N_r\,N_t}$. The new lower bound is now:

$$\begin{aligned} C &\ge \tfrac{1}{T}\textstyle\sum_{i=1}^{T_d} I(Y_i;X_i|\widehat{H}^{(i)}) \\ &\ge \tfrac{T-T_p}{T}\textstyle\sum_{i=1}^{T_d}\mathrm{E}\ln\det(I+\frac{\sigma_x^2}{\sigma_v^2+N_t\sigma_x^2\sigma^2_{\widetilde{H}^{(i)}}}\widehat{H}^{(i)}\widehat{H}^{(i)\,H}) \end{aligned} \quad (9)$$

Let $\sigma^2_{\widehat{H}^{(i)}} = \frac{tr\ \mathrm{E}(\widehat{H}^{(i)}\widehat{H}^{(i)\,H})}{N_r\,N_t}$ and $\overline{H}^{(i)} = \frac{\widehat{H}^{(i)}}{\sigma_{\widehat{H}^{(i)}}}$, then due to the fact that the channel estimator satisfies the Pythagorean Theorem, we have that $\sigma^2_{\widehat{H}^{(i)}} + \sigma^2_{\widetilde{H}^{(i)}} = \sigma_H^2$. Now:

$$\begin{aligned} C &\ge \tfrac{T-T_p}{T}\textstyle\sum_{i=1}^{T_d}\mathrm{E}\ln\det(I+\frac{\sigma_x^2(\sigma_H^2-\sigma^2_{\widetilde{H}^{(i)}})}{\sigma_v^2+N_t\sigma_x^2\sigma^2_{\widetilde{H}^{(i)}}}\overline{H}^{(i)}\overline{H}^{(i)\,H}) \\ &= C_{LB} \end{aligned} \quad (10)$$

The expectation is over the distribution of $\overline{H}^{(i)}$, which remains close to that of $H^{(i)}$. Then the given capacity lower bound C_{LB} depends primarily on the Mean Square Error (MSE) of the channel estimator $\widehat{H}^{(i)}$. Since C_{LB} is a decreasing function of the MSE, the optimum estimator is the Minimum Mean Square Error(MMSE) estimator:

$$\begin{aligned} \widehat{H}^{(i)}_{MMSE} &= \widehat{H}^{(i)}_{MMSE}(X_p, X_1^{i-1}, \overline{Y}_i) \\ &= \mathrm{E}(H|X_p, X_1^{i-1}, \overline{Y}_i) \end{aligned} \quad (11)$$

which is an unbiased estimator of H. The performance of any unbiased estimator is bounded by the Cramer-Rao lower bound:

$$R_{\widetilde{\mathbf{h}}^{(i)}\widetilde{\mathbf{h}}^{(i)}} = \mathrm{E}\,\widetilde{\mathbf{h}}^{(i)}\widetilde{\mathbf{h}}^{(i)\,T} \ge J^{-(i)} \quad (12)$$

where $\mathbf{h} = [Re(vect(H))^T\ \ Im(vect(H))^T]^T$ and $J^{(i)}$ is the Bayesian Fischer Information Matrix (FIM) for the a posteriori distribution of H, and is in this case:

$$\begin{aligned} J^{(i)} &= -\mathrm{E}\,\tfrac{\partial}{\partial\mathbf{h}}\left(\tfrac{\partial\ln p(H|X_p,X_1^{i-1},\overline{Y}_i)}{\partial\mathbf{h}}\right)^T \\ &= \underbrace{-\mathrm{E}\,\frac{\partial}{\partial\mathbf{h}}\left(\frac{\partial\ln p(X_p,Y_p,X_1^{i-1},Y_1^{i-1}|H)}{\partial\mathbf{h}}\right)^T}_{J^{(i)}_{DA\ training}} \\ &\quad \underbrace{-\mathrm{E}\,\frac{\partial}{\partial\mathbf{h}}\left(\frac{\partial\ln p(Y_{i+1}^{T_d}|H)}{\partial\mathbf{h}}\right)^T}_{J^{(i)}_{blind}} \underbrace{-\mathrm{E}\,\frac{\partial}{\partial\mathbf{h}}\left(\frac{\partial\ln p(H)}{\partial\mathbf{h}}\right)^T}_{J^{(i)}_{prior}} \end{aligned}$$

We have $\sigma^2_{\widetilde{H}^{(i)}} \ge \frac{tr\,J^{-(i)}}{N_t\,N_r}$. This is an absolute lower bound on the channel estimation MSE. The MMSE estimator achieves this bound asymptotically ($T_d \to \infty$). The above result is also valid for the case when the channel varies from block to block, since channel re-estimation for every block is assumed.

3.2. Asymptotic Behavior

We focus here on the asymptotic behavior of the capacity loss for small channel estimation error. We suppose that the channel is constant over every block, the mutual information for the particular block (i) at the receiver assuming channel estimation is then $I(Y_i;X_i|\widehat{H}^{(i)}) = \mathrm{E}(I_{H^{(i)}}(Y_i;X_i|\widehat{H}^{(i)}))$, where $I_{H^{(i)}}(Y_i;X_i|\widehat{H}^{(i)})$ is the capacity for a particular realization of the channel. It is assumed here that the channel $H^{(i)}$ may vary from block to block (while allowing an asymptotic regime). The MI assuming perfect channel knowledge and for a particular realization is $I_{H^{(i)}}(Y_i;X_i|H^{(i)})$. Let's temporarily drop the block index (i).
In the following, we derive a weighting matrix that appears in the asymptotic MI decrease and optimal semiblind channel estimate. The first order derivative of $I_H(Y;X|\widehat{H})$ with respect to $\widetilde{H} = H - \widehat{H}$ evaluated at $\widetilde{H} = 0$ is zero:

$$\begin{aligned} &\tfrac{\partial}{\partial\widetilde{\mathbf{h}}}I_H(Y;X|\widehat{H})|_{\widetilde{H}=0} = \tfrac{\partial}{\partial\widetilde{\mathbf{h}}}\left(h(X) - h_H(X|Y,\widehat{H})\right)|_{\widetilde{H}=0} \\ &= \int\int\tfrac{p_H(Y,X|H)}{p_H(X|Y,H)}\tfrac{\partial}{\partial\widetilde{\mathbf{h}}}p_H(X|Y,\widehat{H})|_{\widetilde{H}=0}dXdY \\ &= \int\int p_H(Y|H)\tfrac{\partial}{\partial\widetilde{\mathbf{h}}}p_H(X|Y,\widehat{H})|_{\widetilde{H}=0}dXdY \\ &= \int p_H(Y|H)\tfrac{\partial}{\partial\widetilde{\mathbf{h}}}(\underbrace{(\int p_H(X|Y,\widehat{H})dX)}_{=1}|_{\widetilde{H}=0}dY = 0\,. \end{aligned} \quad (13)$$

The second order derivative w.r.t. $\widetilde{H}$ evaluated at 0 is:

$$\begin{aligned} &\tfrac{\partial}{\partial\widetilde{\mathbf{h}}}\left(\tfrac{\partial I_H(Y;X|\widehat{H})}{\partial\widetilde{\mathbf{h}}}\right)^T|_{\widetilde{H}=0} = -\mathrm{E}\,\tfrac{\partial}{\partial\widetilde{\mathbf{h}}}\left(\tfrac{\partial\ln p_H(Y|\widehat{H})}{\partial\widetilde{\mathbf{h}}}\right)^T|_{\widetilde{H}=0} \\ &+\mathrm{E}\,\tfrac{\partial}{\partial\widetilde{\mathbf{h}}}\left(\tfrac{\partial\ln p_H(Y|X,\widehat{H})}{\partial\widetilde{\mathbf{h}}}\right)^T|_{\widetilde{H}=0} = J_Y(H) - J_{Y|X}(H) \end{aligned} \quad (14)$$

J_Y and $J_{Y|X}$ are FIMs describing the MI decrease due to the channel estimation error and evaluated at $\widetilde{H} = 0$. To compute J_Y and $J_{Y|X}$, we consider a receiver point of view in which, given a certain realization H and a certain $\widehat{H}$, $\widetilde{H}$ is an unknown constant. Then $p_H(Y|X,\widehat{H}) = p(V + (\widehat{H}+\widetilde{H})X|X,\widehat{H}) = p(V+\widetilde{H}\,X|X) = p_{\widetilde{H}}(\widetilde{V}|X)$ where $\widetilde{V} = V + \widetilde{H}\,X$ and the variable $\widetilde{V}|X$ follows the same distribution as V but with an offset (mean) $\widetilde{H}\,X$ (notation assumes $N_i = 1$). Now:

$$J_{Y|X}(H) = -\mathrm{E}\left(\frac{\partial}{\partial\widetilde{\mathbf{h}}}\left(\frac{\partial \ln p_{\widetilde{H}}(\widetilde{V}|X)}{\partial\widetilde{\mathbf{h}}}\right)^T|_{\widetilde{H}=0}\right)$$
$$J_Y(H) = \mathrm{E}(\mathbf{B}\mathbf{B}^T), \quad \mathbf{B} = \frac{\partial \ln p_H(Y|\widehat{H})}{\partial\widetilde{\mathbf{h}}}|_{\widetilde{H}=0}$$

$$\begin{aligned}\mathbf{B} &= \frac{1}{p_H(Y|H)}\frac{\partial\int p_H(Y,X|\widehat{H})dX}{\partial\widetilde{\mathbf{h}}}|_{\widetilde{H}=0}\\ &= \frac{1}{p_H(Y|H)}\frac{\partial\int p_H(Y|X,\widehat{H})p(X)dX}{\partial\widetilde{\mathbf{h}}}|_{\widetilde{H}=0}\\ &= \frac{1}{p_H(Y|H)}\int\frac{\partial p_{\widetilde{H}}(\widetilde{V}|X)}{\partial\widetilde{\mathbf{h}}}|_{\widetilde{H}=0}p(X)dX\end{aligned}$$

For $\widehat{H}$ in a small neighborhood of H, the MI decrease is approximated by a quadratic function:

$$\begin{aligned}&I_H(Y;X|\widehat{H}) - I_H(Y;X|H) = -\widetilde{\mathbf{h}}^T W(H)\widetilde{\mathbf{h}}\\ &= -(\widehat{\mathbf{h}}-\mathbf{h})^T(J_{Y|X}(H)-J_Y(H))(\widehat{\mathbf{h}}-\mathbf{h})\end{aligned} \quad (15)$$

The weighting matrix $W(H) = J_{Y|X}(H) - J_Y(H)$ is non-negative definite since $I_H(Y;X|\widehat{H}) \le I_H(Y;X|H)$.

Example: For white Gaussian and decorrelated V_i and X_i, $\widetilde{V}_i|X_i$ is Gaussian with mean $\widetilde{H}\,X$ and covariance $\sigma_v^2 I$: $p_{\widetilde{H}}(\widetilde{V}_i|X_i) = (2\pi\sigma_v^2)^{-N_i N_{rx}}\exp -\frac{|\widetilde{V}_i-(I_{N_i}\otimes\widetilde{H})\,X_i|^2}{\sigma_v^2}$. Then $J_{Y_i}(H) = 0$ and $J_{Y_i|X_i}(H) = N_i\frac{\sigma_x^2}{\sigma_v^2}I$. As a result $W_i(H) = W_i = N_i\frac{\sigma_x^2}{\sigma_v^2}I$ is constant. More generally, for any Gaussian noise V_i and zero mean input X_i; $J_{Y_i}(H) = 0$, $J_{Y_i|X_i}(H) = J_{Y_i|X_i}$ is constant and $W_i(H) = W_i = J_{Y_i|X_i}$. We now want to find the best channel estimator, that minimizes the MI decrease. Then we need to minimize the cost function $I(Y;X|H) - I(Y;X|\widehat{H}(S)) = \mathrm{E}\{I_H(Y;X|H) - I_H(Y;X|\widehat{H}(S))\}$ where for every block i, $\widehat{H}$ is based on $S^{(i)} = (X_p, X_1^{i-1}, \overline{Y}_i)$. Asymptotically

$$\begin{aligned}&\min_{\widehat{\mathbf{h}}(S)}(I(Y;X|H) - I(Y;X|\widehat{H}(S)))\\ &\Rightarrow \mathrm{E}[\min_{\widehat{\mathbf{h}}(S)}\ \mathrm{E}\{(\widehat{\mathbf{h}}(S)-\mathbf{h})^T W(H)(\widehat{\mathbf{h}}(S)-\mathbf{h})|S\}]\\ &\Rightarrow \widehat{\mathbf{h}}_{opt}(S) = (\mathrm{E}\{W(H)|S\})^{-1}(\mathrm{E}\{W(H)\mathbf{h}|S\})\end{aligned} \quad (16)$$

So $\widehat{\mathbf{h}}_{opt}$ is a weighted MMSE estimate. For centered input X and Gaussian noise V, $W(H) = W$ and the optimal channel estimator is the MMSE estimator $\widehat{\mathbf{h}}_{opt}(S) = W^{-1}\,\mathrm{E}\{W\mathbf{h}|S\} = \mathrm{E}\{\mathbf{h}|S\}$. The minimum mean MI decrease w.r.t. the perfectly known channel case is:

$$\begin{aligned}&I(Y;X|H) - I(Y;X|\widehat{H}_{opt}(S)) = \mathrm{E}\,[\mathbf{h}^T W(H)\mathbf{h}\\ &-(\mathrm{E}\{W(H)\mathbf{h}|S\})^T(\mathrm{E}\{W(H)|S\})^{-1}(\mathrm{E}\{W(H)\mathbf{h}|S\})].\end{aligned}$$

and all this so far for block (i). The minimum mean MI decrease for the complete burst becomes, using the recursive MI decomposition of section 2:

$$\begin{aligned}&\textstyle\sum_{i=1}^Q[I(Y_i;X_i|H^{(i)}) - I(Y_i;X_i|\widehat{H}_{opt}^{(i)}(S^{(i)}))]\\ &= \textstyle\sum_{i=1}^Q[\mathrm{E}\{\mathbf{h}^T W(H^{(i)})\mathbf{h}\\ &-(\mathrm{E}\{W_i\mathbf{h}|S^{(i)}\})^T(\mathrm{E}\{W_i|S^{(i)}\})^{-1}(\mathrm{E}\{W_i\mathbf{h}|S^{(i)}\})\}].\end{aligned}$$

where $W_i = W_i(H^{(i)}) = J_{Y_i|X_i}(H^{(i)}) - J_{Y_i}(H^{(i)})$.
In the case of a constant channel over the different blocks (i.e $H^{(i)} = H$, $i = 1,\ldots,Q$), $\sum_{i=1}^Q I(Y_i;X_i|H^{(i)})$ gets modified to $I(Y_d;X_d|H)$ and $W_i(H^{(i)})$ to $W_i(H)$.

3.3. Deterministic channel

In this case we have no prior information on the channel. The channel is constant during the burst with an unknown value H. The coherent capacity is then $I(Y_d;X_d|Y_p,X_p,H)$. For a given realization Y^o of Y and a (variable) channel $\widehat{H}$, let's define $G(Y^o,X_p,\widehat{H}) = h(X) - \mathrm{E}(-\ln q(X|Y^o,\widehat{H})) = h(X_d) - \mathrm{E}(-\ln q(X|Y_p^o,Y_d^o,\widehat{H}))$ where the expectation here is with respect to X_d (X_p is known), and $q(X|Y_p^o,Y_d^o,\widehat{H}) = p(X|Y_p^o,Y_d^o,H)|_{H=\widehat{H}}$. Let's introduce a partition of Y_d in which the different blocks have the same lengths $N_i = N_1$, $i = 1,\ldots,Q$ and are i.i.d. Then in $\frac{1}{T}G(Y^o,X_p,\widehat{H}) = \frac{1}{T}[\ln q(X_p|Y_p^o,\widehat{H}) + \sum_{i=1}^Q(h(X_i) + \mathrm{E}\ln q(X_i|Y_i^o,\widehat{H}))]$, the averaging over the blocks tends asymptotically to an expectation w.r.t. Y_d^o. We conclude that :

$$\begin{aligned}&\lim_{T\to\infty}\tfrac{1}{T}G(Y^o,X_p,H) = \lim_{T\to\infty}\tfrac{1}{T}\{\ln p(X_p|Y_p^o,H)\\ &-\textstyle\sum_{i=1}^Q(h(X_i) - \mathrm{E}\ln p(X_i|Y_i^o,H)\} = I(y;x|H)\end{aligned} \quad (17)$$

So asymptotically $G(Y^o,X_p,H)$ approaches the mutual information. Let's define $G_\infty(\widehat{H}) = \lim_{T\to\infty}\frac{1}{T}G(Y^o,X_p,\widehat{H})$.

$$\begin{aligned}&G_\infty(H) - G_\infty(\widehat{H})\\ &= \tfrac{1}{N_1}\,\mathrm{E}_{X_1,Y_1|H}[\ln p(X_1|Y_1,H) - \ln q(X_1|Y_1,\widehat{H})]\\ &= \tfrac{1}{N_1}\textstyle\int p(Y_1|H)(\int p(X_1|Y_1,H)\ln\frac{p(X_1|Y_1,H)}{q(X_1,Y_1|\widehat{H})}dX_1)dY_1\\ &= \tfrac{1}{N_1}\,\mathrm{E}_{Y_1|H}[D(p(X_1|Y_1,H)||q(X_1|Y_1,\widehat{H}))] \ge 0\end{aligned}$$

(Kullback-Leibner distance). This means that $G_\infty(\widehat{H})$ is maximized when $q(X_1|Y_1,\widehat{H}) = p(X_1|Y_1,H)$. This fact, combined with the hypothesis that the training part is sufficiently informative to allow, together with the blind information, complete channel identifiability, ensures that asymptotically (for T_d and T_p big enough) $G(Y^o,X_p,\widehat{H})$ is maximized for $\widehat{H} = H$. We can hence use the maximization of $G(Y^o,X_p,\widehat{H})$ as optimization approach to find a consistent channel estimator.
This method is related to the first iteration of an iterative MAP/ML estimation approach for input signal/channel with EM applied to the ML estimation part for the channel. The maximum a posteriori (MAP) estimate of X_d in the MAP/ML approach is:

$$X_d^{MAP} = \arg_{X_d}\max_{X_d,H'}\ln q(X|Y^o,H')\,. \quad (18)$$

Various solution techniques exist for this type of problem. Consider an alternating maximization (between X_d and H) approach in which expectation over X_d is introduced when maximizing over H (EM-like approach). The iterations comprise two steps. The first step for the first iteration gives:

$$\begin{aligned}\widehat{H} &= \arg\max_{H'}\,\mathrm{E}(\ln q(X|Y^o,H'))\\ &= \arg\max_{H'}\,G(Y^o,X_p,H')\,.\end{aligned} \quad (19)$$

This is a semiblind cost function for the channel estimation. The second step consists of the MAP estimation of the input assuming the channel $H' = \widehat{H}$.

Remark1: The recursive decomposition of section 2 remains valid in the deterministic channel case, and we can process by successive detection of the symbols (blocks). To have an acceptable algorithm complexity, one can choose from a variety of channel updating techniques. Along the lines of section 2, this approach allows to maximize the MI for large T.

Remark2: Similarly to the Bayesian case, we can evaluate the asymptotic MI decrease as $\widetilde{\mathbf{h}}^T W(H) \widetilde{\mathbf{h}}$ with $W(H) = J_{Y|X}(H) - J_Y(H)$. In the deterministic case however, the direct minimization of the MI decrease does not lead to a meaningful channel estimator.

4. CORRELATED MIMO CHANNEL MODEL

In order to improve channel estimation and reduce capacity loss, it is advantageous to exploit correlations in the channel, if present. So consider the frequency-flat MIMO channel: H $(N_{rx} \times N_{tx})$, $\boldsymbol{h} = \text{vect}(H)$. The correlated channel model we suggest is:

$$\boldsymbol{h} = S\, \boldsymbol{g} \qquad (+ g_0\, \overline{\boldsymbol{h}} \text{ for direct path, } |g_0| = 1)$$

where the elements of $\boldsymbol{g}$ are taken to be i.i.d. Gaussian for a stochastic model. The correlations are captured by S.

Special case 1: Bell-Labs/Saturn model:

$H = R_r^{1/2}\, G\, R_t^{H/2} \Rightarrow S = R_t^{H/2} \otimes R_r^{1/2}$, $\boldsymbol{g} = \text{vect}(G)$

Special case 2: Cioffi-Raleigh model (multipath model):

$$H = \sum_i g_i\, \boldsymbol{a}_i\, \boldsymbol{b}_i^H \quad \Rightarrow \quad S = [\boldsymbol{b}_1^* \otimes \boldsymbol{a}_1 \;\; \boldsymbol{b}_2^* \otimes \boldsymbol{a}_2 \;\; \cdots] .$$

The model $\boldsymbol{h} = S\, \boldsymbol{g}$ is straightforwardly extendible to the non-zero delay-spread case.

5. OBSERVATIONS

Capacity approaching channel estimation should exploit prior info + data/decision aided info + (Gaussian) blind info. The symbol-wise decomposition of the block fading channel capacity involves for each symbol position in a burst a channel estimate that is based on: prior channel distribution info, training and detected inputs up to that symbol position, (Gaussian) blind info in remaining channel outputs. Hence, symbol-wise Gaussian semiblind (Bayesian) channel estimation is required (blind parts: detected data + Gaussian undetected data). To have asymptotically (in SNR and burst length) negligible capacity loss, enough training symbols are required to have (deterministic) identifiability of the parameters that cannot be identified blindly (from the Gaussian undetected symbols), hence blind (Gaussian) info reduces training data requirements. Exploiting channel correlations, the (effective) number of degrees of freedom in the channel and hence the training requirements get reduced. The channel with i.i.d. entries, while optimal from a capacity point of view, is the worst case from the channel estimation point of view. The recursive mutual info decomposition may suggest a practical approach for channel estimation. However, simpler practical approaches would pass through the bursts iteratively, with semiblind (blind info = Gaussian undetected symbols) channel estimation in the first pass, and semiblind (blind info = detected data) channel estimation in the next iterations. Prior channel info (and S in the channel model) gets estimated (sufficiently well) by considering the data in multiple bursts jointly (assuming these parameters are invariant across a (large) set of bursts). Whereas we have considered block fading so far in this paper, we conjecture that these results extend to the continuous transmission (CT) case: in steady-state, channel estimation should be based on the semi-infinite detected past symbols, and semi-infinite future blind channel information. A Gauss-Markov model for the channel variations with a certain Doppler bandwidth will prevent perfect channel extraction from this infinite data though. Finally, the proposed channel model is useful for the introduction of partial channel knowledge at the transmitter. Indeed, if the transmitter can know the channel correlations summarized in S in $\boldsymbol{h} = S\, \boldsymbol{g}$ and only lacks knowledge of the fast fading parameters $\boldsymbol{g}$, the channel capacity may be close to that of the known channel case.

6. REFERENCES

[1] E. Biglieri, J. Proakis and S. Shamai "Fading channels: information-theoretic and communications aspects". *IEEE Trans. Info. Theory*, vol. 44, no. 6, pp. 2619-2692, Oct 1998.

[2] T.L. Marzetta and B.M. Hochwald "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading ". *IEEE Trans. Info. Theory*, vol. 45, no. 1, pp. 139-157, Jan 1999.

[3] L. Zheng and D. Tse "Communication on the Grassmann Manifold: A Geometric Approach to the Noncoherent Multiple Antenna Channel". To appear in *IEEE Trans. Info. Theory*.

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

[5] I.E. Telatar. "Capacity of Multi-antenna Gaussian Channels". *Eur. Trans. Telecom.*, vol. 10, no. 6, pp. 585-595, Nov/Dec 1999.

[6] B. Hassibi and B. M. Hochwald. "How Much Training is Needed in Multiple-Antenna Wireless Links?". Submitted to *IEEE Trans. Info. Theory*.

[7] A. Medles, D.T.M. Slock and E. De Carvalho "Linear prediction based semi-blind estimation of MIMO FIR channels". In Proc. *Third Workshop on Signal Processing Advances in Wireless Communications (SPAWC'01)*, pp. 58-61, Taipei, Taiwan, March 2001.