The 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'07)

A TWO-STAGE APPROACH TO FEEDBACK DESIGN IN MULTI-USER MIMO CHANNELS WITH LIMITED CHANNEL STATE INFORMATION

Randa Zakhour
Institut Eurecom
06560 Sophia Antipolis, France

David Gesbert
Institut Eurecom
06560 Sophia Antipolis, France

ABSTRACT

We consider the downlink of a multiuser MIMO channel, corresponding to a single cell with an N_t-antenna base station and K single-antenna mobile terminals (MTs). It is known that when full channel state information (CSI) is available at the transmitter (full CSIT) the capacity of the system scales as $N_t \log(\frac{P}{N_t} \log K)$, under a total power constraint P [1]. While, when the transmitter has no CSI, scaling reduces to that of a TDMA system. This paper examines the more realistic case of having an intermediate state of CSI. The key idea is based on a split of the allotted feedback between two stages: A first stage devoted to scheduling followed by a second stage for precoder design for the selected users. Based on an approximation of the achievable sum rate, we introduce a method for determining the splitting of the feedback rate so as to maximize performance and provide intuitions. We illustrate the gains of the 2-stage approach via Monte Carlo simulations.

I. INTRODUCTION

Integration of multiple antennas at the transmitter and receiver leads to enhanced capacity through spatial diversity and multiplexing gains. In multiuser configurations, a further capacity increase due to multiuser diversity (MUD) is attainable through judicious scheduling [2].

While in the single-user case CSIT contributes little to achieving the multiplexing gain it is crucial for multiuser MIMO (MU-MIMO). However, CSIT is usually gained at the expense of feedback overhead. A lot of recent research has focused on systems with partial CSIT (limited feedback) so as to circumvent this problem (see [3] and references therein).

Most approaches to MU-MIMO transmission under partial CSI have centered on linear precoding, which is much simpler than nonlinear processing (required for implementing the optimal dirty-paper coding (DPC) scheme, for example), this simplicity coming at the cost of tolerable performance loss [4]. These approaches fall under two categories: orthogonal random beamforming (ORBF) and zero-forcing beamforming (ZFBF).

ORBF, introduced in [1], maintains the same throughput scaling with the number of users as the DPC approach: it consists of generating a number of random orthonormal beams and transmitting on each of them to the user with the corresponding highest signal to interference and noise ratios (SINR); thus, each user need only feed back its SINR and the index of the optimal beam instead of its entire channel information. However, this approach is only efficient when the number of users is large. Several publications have been devoted to improving RBF for cases where the number of users is finite, as well as to analyzing the effect of assigning a finite number of bits to quantizing the SINR information.

When ZFBF is used, the fed-back CSI is used to design a zero-forcing (ZF) precoding channel matrix, which eliminates inter-user interference when perfect CSI is available. In the latter case this scheme also exhibits the same scaling as the optimal strategy. However, when the feedback rate is fixed, the scaling is only maintained if the feedback rate is linearly scaled with SNR in dB [5, 6, 7].

One can also categorize limited feedback approaches according to whether all [6, 7, 8] or only a subset [9, 10, 11] of the users feed back their CSI. In the latter case, for a given maximum feedback rate, criteria are established so that the number of users that feedback their quantized channel is kept limited.

In all the approaches considered so far, the feedback resource exploited to provide CSIT was used at once for both purposes of user selection (or scheduling) *and* precoding matrix design to serve the selected users. The key idea behind this paper is two-fold: (i) It is the scheduling stage of MU-MIMO which consumes most of the feedback as information for a potentially large number K of users must be gathered, while the final precoding concerns at most N_t users, with typically $N_t << K$. (ii) While the final precoder design relies on accurate channel information to allow for a fine spatial separation of the selected users (so as to maintain spatial multiplexing gain), the scheduling algorithm might get away with a poorer representation of the channel. From these two observations, we propose that the allotted feedback bits be split among two tasks (scheduling followed by precoder design) so as to maximize the attainable sum rate under a fixed total feedback constraint. An analysis is presented in this paper that leads to an algorithm for choosing the feedback split factor. The value of the proposed ideas is further confirmed by simulations.

II. SYSTEM MODEL

We consider a multiuser MIMO channel, where a transmitter equipped with N_t antennas communicates with $K \geq N_t$ single-antenna receivers. The latter are assumed to have perfect channel knowledge. The received signal at user k, denoted $y_k \in \mathbb{C}$ can be written as:

$$y_k = \mathbf{h}_k \mathbf{x} + n_k \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^{N_t \times 1}$ is the transmitted signal vector, $\mathbf{h}_k \in \mathbb{C}^{1 \times N_t}$ and and $n_k \in \mathbb{C}$ represent the channel vector and the noise at the kth user, respectively. We assume perfect channel knowledge at the receiver (CSIR), and that both the channel vector entries and the noise are independent identically distributed

1-4244-1144-0/07/$25.00

(i.i.d.) zero mean unit variance complex Gaussian random variables (r.v.'s), $CN(0,1)$. $\mathbf{x}$ is subjected to a total power constraint P such that $\mathbb{E}[\mathbf{x}^{\mathrm{H}}\mathbf{x}] = P$. Furthermore, we assume a block-fading channel and focus on the ergodic sum rate as system performance measure.

III. MU-MIMO PRECODING WITH FULL CSIT (ZERO-FORCING)

Since capacity-achieving DPC is quite complex, we restrict ourselves to a suboptimal but simple linear precoding scheme, namely ZF beamforming, which is known to achieve optimal performance for large K [4]. Under full CSIT, the transmitted signal $\mathbf{x}$ is given by:

$$\mathbf{x} = \sqrt{P}\mathbf{W}_{ZF}\mathbf{s} \tag{2}$$

where $\mathbf{W}_{ZF} \triangleq \mathbf{H}_A^{\dagger}/\sqrt{\mathrm{tr}((\mathbf{H}_A\mathbf{H}_A^{\mathrm{H}})^{-1})}$ is the designed ZF precoding matrix, $\mathbf{H}_A \in \mathbb{C}^{N_t \times N_t} \triangleq [\mathbf{h}_{A_1}^{\mathrm{T}} \dots \mathbf{h}_{A_{N_t}}^{\mathrm{T}}]^{\mathrm{T}}$ being the aggregate channel of the group A of N_t selected users and $\mathbf{s} = [s_1, \dots, s_{N_t}] \in \mathbb{C}^{N_t \times 1}$ is the vector of independent transmit symbols such that $\mathbb{E}[\mathbf{s}^{\mathrm{H}}\mathbf{s}] = \mathrm{I}_{N_t}$.

The ZF process effectively transforms the MIMO channel into N_t parallel subchannels with equal power gain. Consequently, for each $k \in A$, the received signal (1) becomes:

$$y_k = \sqrt{\frac{P}{\mathrm{tr}((\mathbf{H}_A\mathbf{H}_A^{\mathrm{H}})^{-1})}}s_k + n_k \tag{3}$$

The achievable sum rate for a given A is thus equal to:

$$SR_{ZF\text{-}CSIT} = N_t \log_2\left(1 + \frac{P}{\mathrm{tr}((\mathbf{H}_A\mathbf{H}_A^{\mathrm{H}})^{-1})}\right) \tag{4}$$

And the scheduling rule that maximizes (4) is:

$$A_{opt} = \arg \max_{A\subset\{1,\dots,K\},|A|=N_t} \frac{1}{\mathrm{tr}((\mathbf{H}_A\mathbf{H}_A^{\mathrm{H}})^{-1})} \tag{5}$$

IV. MU-MIMO WITH 2-STAGE FEEDBACK

The two steps of scheduling (finding the optimal set A) and precoding matrix design are mapped into two feedback stages. In the first stage, each of the K receivers feeds back a "coarse" quantized version of its channel vector, $\hat{\mathbf{h}}_{1,k}$, $k = 1, \dots, K$; a group of users, denoted by $\hat{A}$, is selected as in (5), but with the real channels replaced by their quantized versions. Thus,

$$\hat{A} = \arg \max_{A\subset\{1,\dots,K\},|A|=N_t} \frac{1}{\mathrm{tr}((\hat{\mathbf{H}}_{1,A}\hat{\mathbf{H}}_{1,A}^{\mathrm{H}})^{-1})} \tag{6}$$

In the second stage, users in $\hat{A}$ send back refinements of their channels (e.g. quantized versions of $\mathbf{h}_k - \hat{\mathbf{h}}_{1,k}$). The new channel estimates $\hat{\mathbf{h}}_{2,k}$ are used to design the ZF precoding matrix $\hat{\mathbf{W}}_{ZF}$:

$$\hat{\mathbf{W}}_{ZF} = \frac{\hat{\mathbf{H}}_{2,\hat{A}}^{\dagger}}{\sqrt{\mathrm{tr}((\hat{\mathbf{H}}_{2,\hat{A}}\hat{\mathbf{H}}_{2,\hat{A}}^{\mathrm{H}})^{-1})}} \tag{7}$$

Due to quantization error, interference will not be entirely eliminated by the ZF-precoder. Thus the received signal vector will be given by:

$$\mathbf{y} = \sqrt{P}\mathbf{H}_{\hat{A}}\frac{\hat{\mathbf{H}}_{2,\hat{A}}^{\dagger}}{\sqrt{\mathrm{tr}((\hat{\mathbf{H}}_{\mathbf{2},\hat{\mathbf{A}}}\hat{\mathbf{H}}_{\mathbf{2},\hat{\mathbf{A}}}^{\mathrm{H}})^{-1})}}\mathbf{s} + \mathbf{n} \tag{8}$$

Rewriting $\mathbf{H}_{\hat{A}}$ as the sum of the quantized channel and an error term, $\mathbf{H}_{\hat{A}} = \hat{\mathbf{H}}_{2,\hat{A}} + \mathbf{E}_{2,\hat{A}}$, (8) becomes:

$$\mathbf{y} = \frac{\sqrt{P}}{\sqrt{\mathrm{tr}((\hat{\mathbf{H}}_{\mathbf{2},\hat{\mathbf{A}}}\hat{\mathbf{H}}_{\mathbf{2},\hat{\mathbf{A}}}^{\mathrm{H}})^{-1})}}\mathbf{s} + \frac{\sqrt{P}\mathbf{E}_{2,\hat{A}}\hat{\mathbf{H}}_{2,\hat{A}}^{\dagger}}{\sqrt{\mathrm{tr}((\hat{\mathbf{H}}_{2,\hat{A}}\hat{\mathbf{H}}_{2,\hat{A}}^{\mathrm{H}})^{-1})}}\mathbf{s} + \mathbf{n} \tag{9}$$

Since the second term in (9) corresponds to the deviation from the scaled identity matrix obtained when ZF is perfect (cf. (3)), we use the following performance metric to approximate the achieved sum-rate:

$$SR_{ZF\text{-}Q2} = \sum_{i=1}^{N_t} \log_2(1 + SINR_{\hat{A}_i}) \tag{10}$$

where

$$SINR_{\hat{A}_i} = \frac{P}{\mathrm{tr}((\hat{\mathbf{H}}_{2,\hat{A}}\hat{\mathbf{H}}_{2,\hat{A}}^{\mathrm{H}})^{-1}) + P\|(\mathbf{E}_{2,\hat{A}}\hat{\mathbf{H}}_{2,\hat{A}}^{\dagger})_i\|^2} \tag{11}$$

where $(.)_i$ denotes the i^{th} row of a given matrix.

V. ANALYSIS

A. Feedback Rate Splitting Formalization

Let B_{total} denote the total number of bits available for feedback (the total feedback "rate" across all users), and $\alpha \in [0,1]$ the splitting factor between the two stages of our scheme. Thus $B_1 = \alpha B_{\mathrm{total}}$ and $B_2 = (1-\alpha)B_{\mathrm{total}}$ bits will be dedicated to the scheduling and precoding matrix design stages, respectively.

B. Quantization Model

For user k, entries in the channel vector $\mathbf{h}_k$ are i.i.d. $CN(0,1)$ r.v.'s. To model their quantization we adopt an ideal model from rate-distortion theory [12], corresponding to the successive refinement of a Gaussian with mean squared-error as distortion measure[1]. The resulting achievable distortions per vector entry at each stage in terms of α and B_{total}, are:

$$\sigma_{e_1}^2 = 2^{-\alpha B_{total}/(K\times N_t)} \tag{12}$$

$$\sigma_{e_2}^2 = 2^{-\frac{B_{total}}{N_t}(\frac{\alpha}{K}+\frac{1-\alpha}{N_t})}, \tag{13}$$

Furthermore, entries in the first and second stage quantized channel vectors, $\hat{\mathbf{h}}_{1,k}$ and $\hat{\mathbf{h}}_{2,k}$ respectively, are i.i.d. and re-

[1]This may not be the optimal distortion measure [13] but this model serves our purposes quite well.

lated to each other and to the true CSIT by the following distributions (where $j = 1, \ldots, N_t$):

$$\hat{h}_{1,k,j} \sim CN(0, 1-\sigma_{e_1}^2) \quad (14)$$

$$\hat{h}_{2,k,j}|\hat{h}_{1,k,j} \sim CN(\hat{h}_{1,k,j}, \sigma_{e_1}^2 - \sigma_{e_2}^2) \quad (15)$$

$$h_{k,j}|\hat{h}_{2,k,j} \sim CN(\hat{h}_{2,k,j}, \sigma_{e_2}^2). \quad (16)$$

C. Extreme cases

Before tackling the optimum splitting factor and the resulting performance estimation, we analyze the extreme cases of α being either 0 or 1. Their comparison serves as justification for the adopted rate splitting approach.
Both cases correspond to having a single quantization stage, but differ in the following manner:

- $\alpha = 0$: N_t randomly selected users feed back their channels to enable the design of the precoding matrix.
- $\alpha = 1$: all users feed back their channels; scheduling and precoding matrix design are done based on the quantized channel. Quantization model aside, this is equivalent to the strategy in [6] and most existing work.

To compare performance as a function of feedback rate in both schemes, we estimate the ergodic sum rate by averaging (10) over the quantized channel of the selected users and the corresponding quantization error statistics (we drop the 2 from $SR_{ZF\text{-}Q2}$ since we only have a single stage). Taking the expectation over the quantization error statistics first, we are able to bound $\overline{SR}_{ZF\text{-}Q} \triangleq \mathbb{E}_{\hat{\mathbf{H}}_{\hat{A}}, \mathbf{E}_{\hat{A}}} SR_{ZF\text{-}Q}$ as shown in (17), where $\mathrm{eE}_1(x) \triangleq e^x \mathrm{E}_1(x)$, $\mathrm{E}_1(.)$ being the exponential integral, defined as $E_1(x) = \int_x^\infty \frac{e^{-t}}{t} dt$. A sketch of the derivation of these bounds is provided in Appendix I.
But $\overline{SR}_{ZF\text{-}Q}$ and its bounds will differ for the two α's since:

- $\sigma^2_{e,\alpha=0} \leq \sigma^2_{e,\alpha=1}$ (since $K \geq N_t$)
- The statistics of the quantized channels of users in $\hat{A}$ are different, because of differences in both quantization error and scheduling schemes.

1. $\alpha = 0$: since users are selected randomly, the entries of $\hat{\mathbf{H}}_{\hat{A}}$ are all i.i.d. $CN(0, 1-\sigma_e^2)$. Consequently (using results in [14]), bounds on $\overline{SR}_{ZF\text{-}Q}$ (cf. (17)) are given by (19), where

$$c_0 = \frac{1-\sigma_e^2}{1+P\sigma_e^2} \quad (18)$$

and γ_{EM} is the Euler-Mascheroni constant.

Lemma 1. *For fixed σ_e^2, the given scheme has a multiplexing gain of 0.*

Proof. For fixed σ_e^2, $Pc_0 \to \frac{1-\sigma_e^2}{\sigma_e^2}$ as $P \to \infty$: both upper and lower bound will tend to constants confirming the lemma. □

Theorem 1. *A necessary and sufficient condition for the given scheme to maintain the multiplexing gain of N_t under quantization error is to have $\sigma_e^2 = O(1/P)$.*

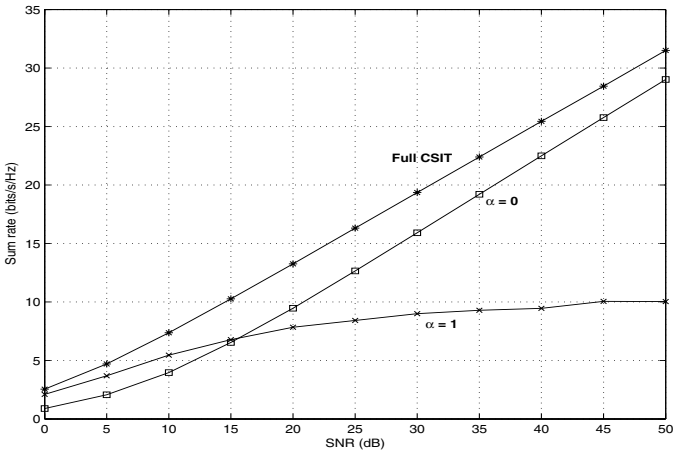

Figure 1: Sum capacities for $M = 2, K = 20, B_{total} = 80$ bits.

Proof. This can be shown by calculating the limit defining the multiplexing gain. Omitted due to space limitations. Note that a similar condition is obtained in [5]. □

Remark 1. *Comparing upper and lower bounds in* (19)*, c_0 can be identified as a scaling factor which quantifies power loss with respect to perfect channel knowledge.*

2. $\alpha = 1$: We lower bound $\lambda_{\min}$'s cumulative distribution function (cdf) by that of the maximum of $N_1 \triangleq \binom{K}{N_t}$ i.i.d. exponentially distributed r.v.s of mean $(1-\sigma_e^2)/N_t$ (effectively ignoring the dependencies between the N_1 possible sets of scheduled users). Similarly, we upper bound it by that of choosing the maximum out of $N_2 \triangleq \left\lfloor \frac{K}{N_t} \right\rfloor$ with the same distribution (only considering disjoint groups). Thus

$$\left(1 - e^{-\frac{N_t x}{1-\sigma_e^2}}\right)^{N_1} \leq F_{\lambda_{\min}}(x) \leq \left(1 - e^{-\frac{N_t x}{1-\sigma_e^2}}\right)^{N_2} \quad (20)$$

Using these cdfs and applying Jensen's inequality to the upper bound (concavity of the log) (cf. (17)), $\overline{SR}_{ZF\text{-}Q}$ is bounded as shown in (21), where H_{N_1} denotes the N_1-th harmonic number.

These bounds can be used to reach conclusions about the scaling and power loss with respect to the perfect CSIT case that are similar to those made when $\alpha = 0$.

Comparing both cases, in order to maintain the same approximate power loss with respect to perfect CSIT for a fixed P, $\alpha = 1$ would necessitate K/N_t times higher feedback rate than $\alpha = 0$, the rate for the former being however greater due to MUD. Fig. 1 shows for the same B_{total}, the achievable sum rates in both cases. At low SNR, the accuracy of the quantization for $\alpha = 1$ is sufficient to achieve some of the MUD gains. This no longer holds at higher SNR, as the system becomes interference-limited. Thus the simple binary scheme of switching between $\alpha = 0$ and $\alpha = 1$ depending on M and K would lead to better performance than either separately. Further improvement would be expected from letting α vary within $[0, 1]$.

$$N_t\mathbb{E}_{\lambda_{\min}}\log\left(1+\frac{\lambda_{\min}}{N_t(1/P+\sigma_e^2)}\right) < \overline{SR}_{ZF\text{-}Q} < N_t\mathbb{E}_{\lambda_{\min}}\left[\log\left(1+P[\lambda_{\min}+\sigma_e^2]\right)\right] - N_t\mathrm{eE}_1\left(\frac{1}{P\sigma_e^2}\right) \quad (17)$$

$$N_t\mathrm{eE}_1\left(\frac{N_t^2}{Pc_0}\right) < \overline{SR}_{ZF\text{-}Q} < N_t\left[\log(1+P\sigma_e^2)+\mathrm{eE}_1\left(\frac{N_t}{Pc_0}\right)-\mathrm{eE}_1\left(\frac{1}{P\sigma_e^2}\right)\right] < N_t\left[\mathrm{eE}_1\left(\frac{N_t}{Pc_0}\right)+\gamma_{EM}\right] \quad (19)$$

D. Optimal α

The optimal α, α_{opt} is defined as :

$$\alpha_{opt} = \arg\max_{\alpha} \overline{SR}_{ZF\text{-}Q2} \quad (22)$$

where $\overline{SR}_{ZF\text{-}Q2} \triangleq \mathbb{E}_{\hat{\mathbf{H}}_1,\hat{\mathbf{H}}_{2,\hat{A}},\mathbf{E}_{2,\hat{A}}}\left[SR_{ZF\text{-}Q2}\right]$ (cf. (10)). Similarly to (29), this expectation can be rewritten as:

$$N_t\mathbb{E}_{\hat{\mathbf{\Lambda}}_{2,\hat{A}},\mathbf{E}_{2,\hat{A}}}\log\left(1+\frac{1}{\sum\lambda_{2,\hat{A},j}^{-1}(\frac{1}{P}+|\mathbf{E}_{2,\hat{A},i,j}|^2)}\right), \quad (23)$$

where $\hat{\mathbf{\Lambda}}_{2,\hat{A}}$ is the diagonal matrix of eigenvalues of $\hat{\mathbf{H}}_{2,\hat{A}}\hat{\mathbf{H}}_{2,\hat{A}}^{\mathrm{H}}$, denoted $\lambda_{2,\hat{A},j}, j=1,\ldots,N_t$, the joint distribution of which depends on that of $\hat{\mathbf{H}}_1$ and $\hat{\mathbf{H}}_{2,\hat{A}}$, alternatively on $\hat{\mathbf{H}}_1$ and $\hat{\mathbf{E}}_{12,\hat{A}} \triangleq \hat{\mathbf{H}}_{2,\hat{A}}-\hat{\mathbf{H}}_{1,\hat{A}}$. Letting $\lambda_{2,\min}$ be the minimum eigenvalue of $\hat{\mathbf{H}}_{2,\hat{A}}\hat{\mathbf{H}}_{2,\hat{A}}^{\mathrm{H}}$, we can bound $\overline{SR}_{ZF\text{-}Q2}$ (cf. (17)) as shown in (24).

$$N_t\mathbb{E}_{\lambda_{2,\min}}\log\left(1+\frac{P\lambda_{2,\min}}{N_t(1+P\sigma_{e_2}^2)}\right) < \overline{SR}_{ZF\text{-}Q2} < N_t\left[\gamma_{EM}+\mathbb{E}_{\lambda_{2,\min}}\log\left(1+\frac{P\lambda_{2,\min}}{1+P\sigma_{e_2}^2}\right)\right] \quad (24)$$

Unfortunately we are unable to bound the distribution of $\lambda_{2,\min}$. Instead we approximate the achievable sum rate by its upper bound obtained from applying Jensen's inequality, i.e. by bringing the expectation in (24) inside the logarithm.

Further noting that

$$\mathbb{E}_{\hat{\mathbf{H}}_{1,\hat{A}},\hat{\mathbf{E}}_{12,\hat{A}}}\mathrm{tr}\left((\hat{\mathbf{H}}_1+\hat{\mathbf{E}}_{12})(\hat{\mathbf{H}}_1+\hat{\mathbf{E}}_{12})^H\right) = \mathbb{E}_{\hat{\mathbf{H}}_{1,\hat{A}}}\mathrm{tr}\left(\hat{\mathbf{H}}_1\hat{\mathbf{H}}_1^H\right)+N_t^2(\sigma_{e_1}^2-\sigma_{e_2}^2) \quad (25)$$

and recalling that the trace is the sum of the eigenvalues, we are intuitively lead to approximate the desired expectation by:

$$\mathbb{E}_{\hat{\mathbf{H}}_{1,\hat{A}},\hat{\mathbf{E}}_{12,\hat{A}}}\lambda_{2,\min} \approx \mathbb{E}_{\hat{\mathbf{H}}_{1,\hat{A}}}\lambda_{1,\min}+c(\sigma_{e_1}^2-\sigma_{e_2}^2) \quad (26)$$

for some finite c.

Guided by the results of the discussion of the two extreme cases and our knowledge of the perfect CSIT case, approximating the expectation of (26) leads us to define the following power loss factor with respect to the ideal case:

$$PL \triangleq \frac{1-\sigma_{e_1}^2}{1+P\sigma_{e_2}^2}+\frac{\sigma_{e_1}^2-\sigma_{e_2}^2}{\log K(1+P\sigma_{e_2}^2)}, \quad (27)$$

where $\log K$ is the expected MUD gain. The first term in the summation is the loss due to the first stage quantization while the second term is that caused by the second stage.

α is thus selected based on the following heuristic:

$$\alpha_{heur} = \arg\max_{\alpha} PL \quad (28)$$

VI. Simulation Results

Simulations were run to investigate the performance of our heuristic algorithm. [2]

Fig. 2 compares the performance for a 30-user system for different values of α. As can be seen, our scheme effectively provides a smooth transition between the two extreme cases, and comparing to the α_{opt} case (found by exhaustive search) the loss due to the non-optimality of the heuristic is reasonable.

Other simulations were aimed at checking how performance varies with the total feedback rate. Results are omitted due to space constraints, but these confirm conclusions from our analysis: when B_{total} is too small, only the multiplexing gain can be achieved at higher SNR, while for increasing B_{total} an increasing part of the MUD gain is achieved; for $\alpha = 0$ increasing the number of bits beyond that necessary to achieving multiplexing gain is no longer useful, and for $\alpha = 1$ saturation still occurs at higher SNR (although later with increasing B_{total}, as expected).

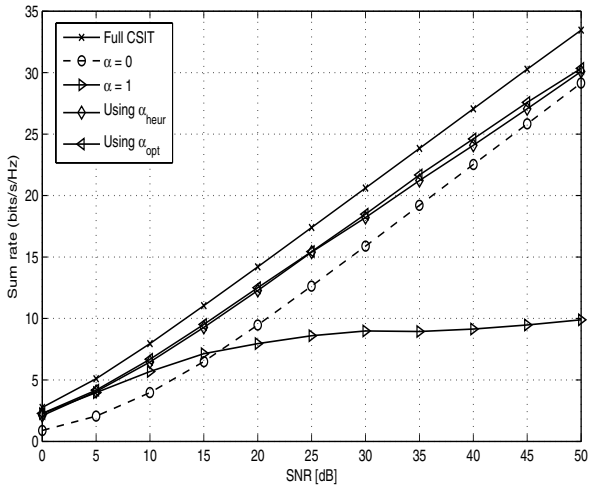

Figure 2: Sum capacities for $M=2$, $K=30$, $B_{total}=120$ bits, and different schemes. α_{heur} provides a smooth transition between extreme α's, with tolerable loss with respect to α_{opt}.

VII. Conclusion

A two-stage resource allocation scheme was proposed for the multiuser MIMO broadcast channel under feedback rate constraint: the available feedback rate is split between the scheduling phase where all receivers feed back a coarse quantization

[2] As scheduling according to (6) is too computationally intensive, we resort to the semiorthogonal user selection (SUS) algorithm of [4] instead.

$$N_t \sum_{k=1}^{N_2} (-1)^{k+1} \binom{N_2}{k} \mathrm{e}\mathrm{E}_1\left(\frac{kN_t^2}{Pc_0}\right) < \overline{SR}_{ZF\text{-}Q} < N_t \left[\gamma_{EM} + \log\left(1 + P\frac{c_0}{N_t} H_{N_1}\right)\right] \quad (21)$$

of their channel state information (CSIT), and the precoding phase where the selected receivers feed back refined versions of their CSI for precoding matrix design. The optimum splitting of the available feedback rate between the two stages was investigated in a Rayleigh fading channel, and a heuristic algorithm was derived and tested.

ACKNOWLEDGEMENTS

The authors would like to acknowledge useful discussions with Marios Kountouris and Nihar Jindal.

APPENDIX I

The expectation $\overline{SR}_{ZF\text{-}Q}$ can be reformulated as shown in equation (29) below,

$$\begin{aligned}
\overline{SR}_{ZF\text{-}Q} &= \mathbb{E}_{\hat{\mathbf{H}}_{\hat{A}}, \mathbf{E}_{\hat{A}}} \sum_{i=1}^{N_t} \log_2(1 + SINR_{\hat{A}_i}) \\
&= N_t \mathbb{E}_{\hat{\mathbf{H}}_{\hat{A}}, \mathbf{E}_{\hat{A}}} \log\left(1 + \frac{1}{\frac{\mathrm{tr}((\hat{\mathbf{H}}_{\hat{A}} \hat{\mathbf{H}}_{\hat{A}}^{\mathrm{H}})^{-1})}{P} + \|(\mathbf{E}_{\hat{A}} \hat{\mathbf{H}}_{\hat{A}}^{\dagger})_i\|^2}\right) \\
&\overset{(a)}{=} N_t \mathbb{E}_{\hat{\mathbf{\Lambda}}_{\hat{A}}, \hat{\mathbf{U}}, \mathbf{E}_{\hat{A}}} \log\left(1 + \frac{1}{\frac{\mathrm{tr}(\hat{\Lambda}_{\hat{A}}^{-1})}{P} + \mathbf{E}_{\hat{A},i} \hat{\mathbf{U}} \hat{\Lambda}_{\hat{A}}^{-1} \hat{\mathbf{U}}^{\mathrm{H}} \mathbf{E}_{\hat{A},i}^{\mathrm{H}}}\right) \\
&\overset{(b)}{=} N_t \mathbb{E}_{\hat{\mathbf{\Lambda}}_{\hat{A}}, \mathbf{E}_{\hat{A}}} \log\left(1 + \frac{1}{\sum_{j=1}^{N_t} \lambda_{\hat{A},j}^{-1} (\frac{1}{P} + |\mathbf{E}_{\hat{A},i,j}|^2)}\right), \quad (29)
\end{aligned}$$

where in (a) $\hat{\mathbf{H}}_{\hat{A}}$ is replaced by its eigenvalue decomposition ($\hat{\mathbf{\Lambda}}_{\hat{A}}$ is the diagonal matrix containing the eigenvalues $\{\lambda_{\hat{A},j}\}$ of $\hat{\mathbf{H}}_{\hat{A}} \hat{\mathbf{H}}_{\hat{A}}^{\mathrm{H}}$; $\hat{\mathbf{U}}$ is unitary), and (b) is obtained by noting that $\mathbf{E}_{\hat{A},i} \hat{\mathbf{U}}$ has the same statistics as $\mathbf{E}_{\hat{A},i}$[3].

$D \triangleq \sum_{j=1}^{N_t} \lambda_{\hat{A},j}^{-1} (1/P + |\mathbf{E}_{\hat{A},i,j}|^2)$ in (29) may be bounded as [15]:

$$\frac{1/P + |\mathbf{E}_{\hat{A},i,j_{\min}}|^2}{\lambda_{\min}} < D < \frac{\sum_{j=1}^{N_t} (1/P + |\mathbf{E}_{\hat{A},i,j}|^2)}{\lambda_{\min}}, \quad (30)$$

where $\lambda_{\min}$ is the smallest eigenvalue in the summation and $j_{\min}$ the index of the corresponding entry in the $\mathbf{E}_{\hat{A},i}$ vector. $\overline{SR}_{ZF\text{-}Q}$ is thus bounded as in equation (31). The sum of N_t squared norms of i.i.d. $CN(0, \sigma_e^2)$ r.v.'s in the lower bound's denominator, and $|\mathbf{E}_{\hat{A},i,j}|^2$ in the upper bound are replaced by a Gamma(N_t, σ_e^2) distributed r.v., γ_{N_t} and a Gamma$(1, \sigma_e^2)$ distributed r.v., γ_1, respectively. Once these changes of variable made, applying Jensen's inequality to the lower bound, and averaging over γ_1 in the upper bound, then upper bounding the result, yields (17).

[3] $\mathbf{E}_{\hat{A},i}$ corresponds to the error vector associated with the quantization of user i's channel, user i being an arbitrary user in $\hat{A}$

$$\begin{aligned}
&\mathbb{E}_{\lambda_{\min}, \mathbf{E}_{\hat{A}}} \log\left(1 + \frac{\lambda_{\min}}{\sum_{j=1}^{N_t} (\frac{1}{P} + |\mathbf{E}_{\hat{A},i,j}|^2)}\right) < \frac{\overline{SR}_{ZF\text{-}Q}}{N_t} \\
&< \mathbb{E}_{\lambda_{\min}, \mathbf{E}_{\hat{A}}} \log\left(1 + \frac{\lambda_{\min}}{\frac{1}{P} + |\mathbf{E}_{\hat{A},i,j_{\min}}|^2}\right) \quad (31)
\end{aligned}$$

REFERENCES

[1] M. Sharif, B. Hassibi, "On the Capacity of MIMO Broadcast Channels with Partial Side Information", *IEEE Transactions on Information Theory*, Vol. 51, No. 2, 506-522, Feb. 2005.

[2] R. Knopp and P.A. Humblet, "Information capacity and power control in single-cell multiuser communications", in *Proc. of Int. Conf. on Communications*, Vol. 1, June 1995.

[3] D. Gesbert, M. Kountouris, R. Heath, C-B. Chae and T. Salzer, "From Single User to Multiuser Communications: Shifting the MIMO paradigm", *Signal Processing Magazine*, Accepted. 2007.

[4] T. Yoo and A. Goldsmith, "On the optimality of Multiantenna Broadcast Scheduling Using Zero-Forcing Beamforming", *IEEE Journal on Selected Areas in Communications*, Vol. 24, No. 3, March 2006.

[5] N. Jindal, "MIMO Broadcast Channels with Finite Rate Feedback", *IEEE Transactions on Information Theory*, Vol. 52, No. 11, pp. 5045-5059, Nov. 2006.

[6] T. Yoo, N. Jindal and A. Goldsmith, "Finite-Rate Feedback MIMO Broadcast Channels with a Large Number of Users", *2006 IEEE International Symposium on Information Theory*, July 2006.

[7] T. Yoo, N. Jindal and A. Goldsmith, "Multi-antenna Broadcast Channels with Limited Feedback and User Selection", to appear in *IEEE Journal Selected Areas in Communications*, 2007.

[8] K. Huang, J. G. Andrews and R. W. Heath Jr., "Orthogonal Beamforming for SDMA Downlink with Limited Feedback", in *Proceedings of 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2007.

[9] D. Gesbert and M. Slim Alouini, "How much feedback is multi-user diversity really worth? ", in *Proceedings of IEEE Intern. Conf. On Communications (ICC)*, 2004.

[10] K. Huang, R. W. Heath Jr. and J. G. Andrews, "Multi-user Aware Limited Feedback for MIMO Systems", submitted to *IEEE Transactions on Signal Processing*, January 2007.

[11] C. Swannack, G. W. Wornell and E. Uysal-Biyikoglu, "MIMO broadcast scheduling with quantized channel state information", *2006 IEEE International Symposium on Information Theory*, July 2006.

[12] W. H. R. Equitz, and T. Cover, "Successive Refinement of Information", *IEEE Transactions on Information Theory*, Vol. 37, No. 2, pp. 269-275, March 1991.

[13] D.J. Love, R.W. Heath Jr., W. Santipach, and M.L. Honig, What is the value of limted feedback for MIMO channels?, *IEEE Communications Magazine*, vol. 42, Issue 10, Oct. 2004.

[14] A. Edelman, "Eigenvalues and condition numbers of random matrices", *Ph.D. dissertation, Math. Dept., MIT*, 1989

[15] C. Swannack, E. Uysal-Biyikoglu and G. W. Wornell, "MIMO broadcast scheduling with limited channel state information", in *Proceedings of 43rd Allerton Conference on Communications, Control and Computing*, Sept. 2005.