MIMO downlink joint processing and scheduling: a survey of classical and recent results

Giuseppe Caire
University of Southern California
Los Angeles, CA, USA
email: caire@usc.edu

***Abstract*— Using M antennas at the base station can boost the downlink throughput by a factor M (multiplexing gain), even though the receivers have a single antenna and do not cooperate. In this semi-tutorial paper, we discuss some low-complexity alternatives to achieve a downlink throughput very close to the optimal sum capacity. We review and compare these alternatives and we show that, in the simple setting of independently fading channels, conventional *linear beamforming* achieves the best tradeoff between performance and complexity. This calls for a reconsideration of the emphasis that some recently proposed non-linear downlink precoding techniques have been given both in the research literature and in the technology development. We address also the problem of joint downlink procesing and scheduling in a packet data system (cross-layer design) and compare recently proposed scheme that require very simple CSIT feedback.**

I. Optimal downlink precoding

A downlink channel with one base station equipped with M transmit antennas and K users with one antenna each is the simplest form of *Multi-Input Multi-Output Gaussian Broadcast Channel* (MIMO-GBC), whose sum capacity and more recently the whole capacity region have been fully characterized in a series of recent papers (see [1] and references therein). A time sample of this channel is characterized by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}, \tag{1}$$

where $\mathbf{x}$ is the signal vector transmitted in parallel by the M transmit antennas, $\mathbf{y} = (y_1, \ldots, y_K)^{\mathsf{T}}$ is the vector of signals individually received by the K users and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ is an i.i.d. proper Gaussian noise vector. The matrix $\mathbf{H} = [\mathbf{h}_1^{\mathsf{H}}, \ldots, \mathbf{h}_K^{\mathsf{H}}]^{\mathsf{H}}$ contains the channel coefficients from the M antennas to the K users, where the *row vector* $\mathbf{h}_k \in CC^{1\times M}$ is the channel of user k. The input is subject to the total average transmit power constraint $\mathrm{tr}(\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{H}}]) \leq P$

For each given channel matrix $\mathbf{H}$, the largest achievable rate region (capacity region) $\mathcal{C}(\mathbf{H}, P)$ is is achieved by "Gaussian codes" and Dirty-Paper Coding (DPC). The maximum rate sum (throughput), solution of $\max_{\mathbf{R}\in\mathcal{C}(\mathbf{H},P)} \sum_k R_k$, is given by

$$R_{\mathrm{sum}}(\mathbf{H}, P) = \max_{\mathbf{q}} \ \log\det\left(\mathbf{I} + \mathbf{H}^{\mathsf{H}}\mathrm{diag}(\mathbf{q})\mathbf{H}\right) \tag{2}$$

where maximization is over $\mathbf{q} \in \mathbb{R}_+^K$ such that $\sum_i q_i \leq P$. The convex maximization in (2) can be solved by the simple iterative algorithm of [2]. The solution $\mathbf{q}^\star$ of (2) is referred to as *dual uplink power allocation*. Once $\mathbf{q}^\star$ is found, the corresponding optimal transmission scheme is obtained as follows: 1) fix an arbitrary successive user ordering (e.g., $K, K-1, \ldots, 1$); 2) compute the transmit "beamforming" vectors $\mathbf{f}_k = \beta_k \mathbf{\Gamma}_k^{-1}\mathbf{h}_k^{\mathsf{H}}$, where $\mathbf{\Gamma}_k = \mathbf{I} + \sum_{i=1}^{k-1} \mathbf{h}_i^{\mathsf{H}}\mathbf{h}_i q_i^\star$ and β_k is such that $|\mathbf{f}_k|^2 = 1$; 3) compute the cross-talk matrix $\mathbf{\Phi}$ such that $[\mathbf{\Phi}]_{i,j} = |\mathbf{h}_i\mathbf{f}_j|^2$; 4) compute the user SINRs

$$\gamma_k = \frac{[\mathbf{\Phi}]_{k,k} q_k^\star}{1 + \sum_{i<k} q_i^\star [\mathbf{\Phi}]_{i,k}} = q_k^\star \mathbf{h}_k \mathbf{\Gamma}_k^{-1} \mathbf{h}_k^{\mathsf{H}} \tag{3}$$

5) solve for the downlink power allocation

$$\mathbf{p}^\star = \left[\mathbf{I} - \mathrm{diag}(\mathbf{a})\mathcal{U}(\mathbf{\Phi})\right]^{-1}\mathbf{a} \tag{4}$$

where $\mathcal{U}(\cdot)$ denotes upper triangular part, and $\mathbf{a}$ is the vector with components $a_k = \frac{\gamma_k}{(1+\gamma_k)[\mathbf{\Phi}]_{k,k}}$; 6) obtain the downlink transmit matrix as $\mathbf{G} = \mathbf{F}\mathrm{diag}(\mathbf{p}^\star)^{1/2}$. DPC for the precoding matrix $\mathbf{G}$ goes as follows: users are encoded in the reverse ordering (in our case, $1, 2, \ldots, K$). The signal x_k of user k has variance 1 and is obtained by DPC treating the signals already produced for users $1, \ldots, k-1$ as *non-causally known interference*. This allows to completely cancel the effect of these users on the receiver of user k, without paying in terms of transmit power. The SINR realized at user k receiver is given by

$$\frac{[\mathbf{\Phi}]_{k,k} p_k^\star}{1 + \sum_{i>k} [\mathbf{\Phi}]_{k,i} p_i^\star} \tag{5}$$

and it is equal to γ_k given in (3). The individual rate of user k is given by $R_k = \log(1+\gamma_k)$ (nat/symbol). Although the individual user rates and powers depend on the chosen ordering, it can be shown that $\sum_k R_k = C(\mathbf{H}, P)$ and that $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{H}}] = \mathrm{tr}(\mathbf{G}^{\mathsf{H}}\mathbf{G}) = \sum_k p_k^\star = P$ for all orderings.

In packet data communications, information bits arrive randomly at the transmitter with given arrival rates $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ and are locally stored into K queues, each associated to one user. The base station operates in SDMA/TDMA mode: at each slot t, a user subset $\mathcal{K}(t)$ is selected. Data bits from the queues of users $k \in \mathcal{K}(t)$ are jointly encoded and transmitted over the M antennas. The resource-allocation policy formed by queue selection (scheduling) and the space-time signaling strategy is referred to as a SDMA/TDMA policy. In this case, the goal of a SDMA/TDMA policy is to *stabilize* the queues of all users in the system, i.e., to achieve finite average delay for all users. Under mild conditions, the maximum achievable stability region coincides

with the *ergodic* capacity region $\mathcal{C}(P)$ of the underlying MIMO GBC [3], [4]. Furthermore, letting $S_k(t)$ denote the queue buffer length of user k in slot t, and $R_k(t)$ denote the instantaneous rate supported by user k in slot t, the *adaptive* policy that maximizes the weighted sum of the instantaneous rates $\sum_{k=1}^{K} S_k(t)R_k(t)$ stabilizes any K-tuple of arrival rates $\boldsymbol{\lambda} \in \mathcal{C}(P)$. Then, efficient maximization of the weighted rate sum for a arbitrary channel matrix $\mathbf{H}$ and non-negative weights $w_1, \ldots, w_K$ is of paramount relevance to implement the adaptive queue-stability SDMA/TDMA policy.

The following algorithm (see [5]) computes the dual uplink powers for the weighted sum-rate maximization problem. It can be shown that the solution is always given by the dual uplink decoding order $\pi_K, \pi_{K-1}, \ldots, \pi_1$ where π is the permutation that sorts the weights in non-increasing order

$$w_{\pi_1} \geq w_{\pi_2} \geq \cdots \geq w_{\pi_K}$$

W.l.o.g. and in order to simplify notation we can consider $\pi_k = k$. The objective function is given by

$$f(\mathbf{q}) = \sum_{k=1}^{K} \Delta_k \log\det\left(\mathbf{I} + \sum_{j=1}^{k} \mathbf{h}_j \mathbf{h}_j^H q_j\right) \tag{6}$$

where we let $\Delta_k \stackrel{\triangle}{=} w_k - w_{k+1} \geq 0$ and we define $w_{K+1} = 0$. Then, we have

Algorithm 1: Initialize $\mathbf{q}^{(0)} = \mathbf{0}$, For $n = 1, 2, \ldots$ do:

1) Water-filling step: let $\boldsymbol{\gamma}^{(n)}$ be the solution of

$$\boldsymbol{\gamma}^{(n)} = \arg \max_{\boldsymbol{\gamma} \geq \mathbf{0},\, \sum_k \gamma_k \leq P} \sum_{j=1}^{K} \Delta_j \sum_{k=1}^{j} \log(1 + \gamma_k \alpha_{k,j}^{(n)}) \tag{7}$$

where, for $k = 1, \ldots, K$ and $j \geq k$, we let

$$\alpha_{k,j}^{(n)} = \mathbf{h}_k^{\mathsf{H}} \left(\mathbf{I} + \sum_{i=1, i\neq k}^{j} q_i^{(n-1)} \mathbf{h}_i \mathbf{h}_i^{\mathsf{H}}\right)^{-1} \mathbf{h}_k \tag{8}$$

2) Update step: let $\mathbf{q}^{(n)} = \frac{1}{K}(\boldsymbol{\gamma}^{(n)} + (K-1)\mathbf{q}^{(n-1)})$.

$\triangle$

The "water-filling step" in Algorithm 1 is a convex optimization that admits a unique solution of the water-filling type that can be found by one-dimensional line search. Once $\lim_{n\to\infty} \mathbf{q}^{(n)} = \mathbf{q}^{\star}$ has been determined, the downlink scheme is obtained via the steps outlined before, for the DPC successive encoding order $1, 2, \ldots, K$.

II. Linear precoding

Linear precoding (or "beamforming") represents a low-complexity simple alternative to DPC. In a linear precoding scheme, the transmit signal $\mathbf{x}$ is given by

$$\mathbf{x} = \mathbf{G}\mathbf{u} \tag{9}$$

where $\mathbf{G} \in \mathbb{C}^{M\times K}$ and the vector $\mathbf{u} = (u_1, \ldots, u_K)^{\mathsf{T}}$ contains the user code symbols. The symbols u_k are independently generated by channel encoders for users $k = 1, \ldots, K$. Assuming w.l.o.g. that $\mathbb{E}[\mathbf{u}\mathbf{u}^{\mathsf{H}}] = \mathbf{I}$, the power constraint imposes that $\mathrm{tr}(\mathbf{G}^{\mathsf{H}}\mathbf{G}) \leq P$. User k SINR under linear precoding is given by

$$\mathsf{SINR}_k = \frac{[\boldsymbol{\Phi}]_{k,k} p_k}{1 + \sum_{i\neq k} [\boldsymbol{\Phi}]_{k,i} p_i} \tag{10}$$

where $\boldsymbol{\Phi}$ was defined before, with $\mathbf{f}_i = \mathbf{g}_i/\sqrt{p_i}$, $\mathbf{g}_i$ denoting the i-th column of $\mathbf{G}$, and $p_i = |\mathbf{g}_i|^2$. Under Gaussian codes and minimum distance decoding at the receivers, the user rates $R_k = \log(1 + \mathsf{SINR}_k)$ are achievable.

The weighted rate sum maximization under linear beamforming is a non-convex problem that has escaped so far to a simple solution. In [6], the constrained maximization with respect to $\mathbf{G}$ is replaced by an unconstrained maximization with respect to $\mathbf{B}$, where the transmit signal is given by $\mathbf{x} = \sqrt{\frac{P}{\mathrm{tr}(\mathbf{B}^{\mathsf{H}}\mathbf{B})}}\,\mathbf{B}\mathbf{u}$. The algorithm proposed in [6] consists finds (iteratively) a stationary point of the rate sum, obtained by setting the gradient of $\sum_k R_k$ w.r.t. $\mathbf{B}$ to zero. We shall refer to this algorithm as "SVH" from the initials of the authors of [6]. The SVH algorithm is plagued by possible convergence to local maxima. Improvements (at the expenses of complexity) are obtained by combining the basic SVH algorithm with some global maximization strategy, such as *differential evolution* (DE) (see [7] and references therein).

Another option consists of combining the basic SVH algorithm with the following general purpose *greedy user selection* strategy. For a fixed transmission scheme, denote by $R(\mathcal{K})$ the achievable throughput obtained by restricting the system to users in the subset $\mathcal{K} \subseteq \{1, \ldots, K\}$.

Algorithm 2: Let $\mathcal{K}_0 = \emptyset$ and $R(\emptyset) = 0$. For $k = 1, \ldots, K$, do:

1) Find
$$i_k = \arg \max_{i \notin \mathcal{K}_{k-1}} R(\mathcal{K}_{k-1} \cup \{i\})$$

2) If $R(\mathcal{K}_{k-1} \cup \{i_k\}) < R(\mathcal{K}_{k-1})$, let $\mathcal{K} = \mathcal{K}_{k-1}$ and exit. Else, if $k = K$ let $\mathcal{K} = \{1, \ldots, K\}$ and exit. Else, let $\mathcal{K}_k = \mathcal{K}_{k-1} \cup \{i_k\}$, and go to 1.

$\triangle$

The greedy user selection of Algorithm 2 can be combined with the SVH algortihm, or with zero-forcing (ZF) linear beamforming with water-filling (WF) power allocation (briefly, ZFWF). The resulting "greedy-ZFWF" has been proposed in [8]. It is wortwhile noticing here that the greedy-ZFWF yields, by definition, equal or better rate sum than the greedy algorithm presented in [9], as well as of any greedy user selection based uniquely on $\mathbf{H}$. In fact, any algorithm that adds one user at each step and does not re-process the set of already selected users (greedy), under the same beamforming optimization, cannot outperform Algorithm 2. Furthermore, both greedy-SVH and greedy-ZFWF can be easily modified in order to take into account arbitrary weighted rate sum maximization [10].

III. Non-linear precoding

Seeking a tradeoff between the performance of DPC and and the low complexity of linear precoding, several non-linear

precoding schemes have been proposed. In general, a non-linear precoder produces the transmitted signal $\mathbf{x}$ in the form

$$\mathbf{x} = \frac{1}{\sqrt{\rho}}\mathbf{G}(\mathbf{u}+\mathbf{d}-\boldsymbol{\lambda}) \tag{11}$$

where $\mathbf{G}$ is some transmit matrix such that $\mathrm{tr}(\mathbf{G}^{\mathsf{H}}\mathbf{G}) \leq P$, $\mathbf{d}$ is a random *dithering* signal assumed known at the transmitters' side and $\boldsymbol{\lambda}$ is a data-dependent lattice point, i.e., $\boldsymbol{\lambda}$ is a function of $\mathbf{u}$ and $\mathbf{d}$. The normalization coefficient ρ is defined by

$$\rho = \frac{\mathrm{tr}(\mathbf{G}^{\mathsf{H}}\mathbf{G}\mathbb{E}[(\mathbf{u}+\mathbf{d}-\boldsymbol{\lambda})(\mathbf{u}+\mathbf{d}-\boldsymbol{\lambda})^{\mathsf{H}}])}{P} \tag{12}$$

where expectation is with respect to the data vector $\mathbf{u}$ and the dither vector $\mathbf{d}$.

In order to allow independent decoding at the receivers, $\boldsymbol{\lambda}$ is restricted to take values in a K-fold Cartesian-product lattice Λ^K, where Λ denotes a one-dimensional (complex) lattice (typically, $\Lambda = \tau\mathbb{Z}[j]$ for some scaling factor τ).

Each user k "sees" the single-input single-output channel given by

$$y_k = \frac{1}{\sqrt{\rho}}(\mathbf{h}_k\mathbf{g}_k)(u_k + d_k - \lambda_k) + \eta_k \tag{13}$$

where η_k is noise plus residual interference. The receiver subtracts the known dither component d_k and, after appropriate scaling, applies a modulo-Λ operation[1] in order to remove the data-dependent lattice point λ_k. In general, the resulting modulo-Λ channel at receiver k is given by

$$y'_k = u_k + \zeta_k \mod \Lambda \tag{14}$$

For an appropriately designed dithering signal $\mathbf{d}$, ζ_k and u_k are independent or *almost* independent. Under the independence assumption, the achievable rate of user k can be lower-bounded by

$$R_k \geq \left[\log \mathrm{Vol}(\Lambda) - \log\left(\pi e \sigma_\zeta^2\right)\right]_+ \tag{15}$$

where $\sigma_\zeta^2 = \mathbb{E}[|\zeta_k|^2]$ and where we used the fact that the differential entropy of the output of the modulo-Λ channel is maximized by the uniform distribution $\mathrm{Uniform}(\mathcal{V})$, where $\mathcal{V}$ denotes the Voronoi cell of Λ [11], of volume $\mathrm{Vol}(\Lambda)$. A simple choice for the dither signal is to let d_k with real and imaginary parts i.i.d., uniformly distributed in the interval $[0, \kappa]$, for some $\kappa \gg 1$.

We examine two well-known non-linear precoding schemes. The first is a modified version of the classical THP [11]. In this case, the component of $\boldsymbol{\lambda}$ are computed in sequence, for a given encoding order (w.l.o.g., assume $1, 2, \ldots, K$). Using the DPC transmit matrix and power allocation, it can be shown that the modified THP achieves the sum rate lower bound [11]

$$R_{\mathrm{sum}}^{\mathrm{thp}}(\mathbf{H}, P) \geq \sum_{k=1}^{K} \left[\log(1+\gamma_k) - \log \pi e G(\Lambda)\right]_+ \tag{16}$$

where $G(\Lambda)$ is the normalized second moment of Λ (equal to $1/6$ for the Gaussian integer grid). Assuming that all the terms for which $\gamma_k > 0$ in (16) get positive rate, it follows that the THP incurs a penalty of $\log \pi e G(\Lambda)$ nat/symbol per active user with respect to DPC (this is roughly 0.5 bit/symbol per active users in the case of the Gaussian integer grid). This loss might be significant. Hence, we propose to combine the modified THP with the greedy user selection of Algorithm 2 in order to limit the number of active users to at most M users. The resulting scheme shall be referred to as "greedy-THP".

The second non-linear precoding scheme is the vector precoding (VP) introduced in [12]. VP chooses $\boldsymbol{\lambda}$ according to the instantaneous transmit energy minimization

$$\boldsymbol{\lambda} = \arg \min_{\boldsymbol{\lambda}' \in \Lambda^K} |\mathbf{G}(\mathbf{u}+\mathbf{d}-\boldsymbol{\lambda})|^2 \tag{17}$$

This amounts to a closest lattice point search to find the point of the lattice $\Lambda' = \mathbf{G}\Lambda^K$ closest to the vector $\mathbf{v} = \mathbf{G}(\mathbf{u}+\mathbf{d})$, that can be efficiently implemented by using universal lattice decoding algorithms known as "sphere-decoders".

In [12], a VP scheme is proposed in conjunction with ZF precoding or with a regularized pseudo-inverse precoding, and uniform power allocation (this shall be referred to as the ZFVP algorithm). We generalize this idea and examine VP when the matrix $\mathbf{G}$ is designed according to different criteria. In particular, for $\mathbf{G}$ designed according to the greedy-ZFWF scheme (referred to as the greedy-ZFWFVP algorithm), we obtain the achievable sum rate lower bound

$$R_{\mathrm{sum}}^{\mathrm{zfwfvp}}(\mathbf{H}, P) \geq \sum_{k \in \mathcal{K}} \left[\log p_k - \log(\rho \pi e G(\Lambda))\right]_+ \tag{18}$$

where p_k is the waterfilling downlink power for active user k and $\mathcal{K}$ is the set of users selected by Algorithm 2. A disadvantage of both the ZFVP and the greedy-ZFWFVP algorithms is that the transmit matrix $\mathbf{G}$ is designed a priori, without taking into account the fact that the effective noise variance with VP is decreased from 1 to $\rho \leq 1$. As an alternative, we have examined an iterative optimization of $\mathbf{G}$ under the VP strategy such that, at every iteration, ρ is computed for given $\mathbf{G}$ using (12) and Monte Carlo average, and then $\mathbf{G}$ is re-optimized for the new equivalent noise variance equal to ρ [7]. Furthermore, when computing $\mathbf{G}$, we directly optimize the modulo-Λ channel rate sum, based on the lower bound (15). This yields a modified waterfilling solution that allocates equal downlink power to all users $k \in \mathcal{K}$. This scheme shall be referred to as the "modified greedy-ZFWFVP".

We observe the following important facts. 1) The THP makes use of "inflation factors" $\alpha_k = \frac{\mathsf{SINR}_k}{1+\mathsf{SINR}_k}$ in order to achieve user rates $R_k = \log(1+\mathsf{SINR}_k) - \log \pi e G(\Lambda)$ [11]. On the contrary, VP must use $\alpha_k = 1$ and therefore it can achieve user rates $R_k = \log \mathsf{SINR}_k - \log(\rho \pi e G(\Lambda))$, where SINR_k denotes the "pre-modulo" Λ SINR of user k. 2) On the other hand, THP achieves always $\rho = 1$, while VP achieves generally $\rho \leq 1$. Hence, it is not a priori clear which one of the two schemes yields the best performance. 3) The VP scheme is significantly more computationally demanding than THP since it needs an $|\mathcal{S}|$-dimensional sphere encoder at the

[1] We define $[a] \mod \Lambda = a - Q_\Lambda(a)$ where $Q_\Lambda(a) = \arg\min_{\lambda \in \Lambda} |a-\lambda|$ is the minimum distance lattice quantizer based on Λ.

transmitter, while THP computes just a sequence of (complex) scalar quantizations.

For the realistic case where K is significantly larger than M (Fig. 1), the greedy linear precoder finds easily a good subset of users with almost mutually orthogonal channels. No non-linear scheme beats the SVH with DE for any range of SNR. The non-linear gain is not sufficient to compensate for the modulo-Λ loss. Furthermore, the greedy-SVH and the greedy-ZFWF for $K \gg M$ yield performance almost identical to the SVH with DE at much lower complexity, in agreement with the result of [9].

IV. Schemes with non-perfect CSI

While the general case of non-perfect CSIT is yet to be solved in an information theoretic sense [13], several schemes have been proposed in order to limit the amount of required CSIT feedback while achieving good performance. A popular scheme consists of quantized channel state information (via noiseless feedback) [14]. Another possibility is to use random beamforming (RB) and SINR feedback, opportunistically allocating users to the beams where they achieve their best SINR [15].

Here we consider another option, that easily incorporates the relevant case where the channel changes in time and the feedback has a delay. Hence, CSIT is outdated even in the case of ideal noiseless channel state feedback. Each receiver estimates *exactly* their channel vector $\mathbf{h}_k(t)$ and sends it back without coding and quantization (we referto this scheme as "Analog Feedback" [16]). Due to a delay of d slots and to the noise feedback link, the trasmitter at time t has a sequence of delayed and noisy observations of the channel matrix process $\{\mathbf{H}(t)\}$. Finally, the CSIT $\boldsymbol{\alpha}(t)$ at time t is defined as the MMSE *prediction* of $\mathbf{H}(t)$ given these observations. Assuming that the channel evolves according to a Gauss-Markov process, the MMSE prediction can be computed efficiently using a Kalman filter.

We consider a newly proposed greedy beamforming algorithm for weighted rate-sum maximization [10]. Following the formulation in [17], let $\boldsymbol{\Sigma}_k = \mathbb{E}[\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}|\boldsymbol{\alpha}]$ denote the conditional channel covariance given the CSIT and define an "average SINR" conditionally on CSIT as

$$\mathsf{SINR}_k(\boldsymbol{\alpha}) = \frac{\mathbf{b}_k^{\mathsf{H}}\boldsymbol{\Sigma}_k\mathbf{b}_k}{\mathrm{tr}(\mathbf{B}^H\mathbf{B})/P + \sum_{j\neq k}\mathbf{b}_j^{\mathsf{H}}\boldsymbol{\Sigma}_k\mathbf{b}_j} \quad \forall k \qquad (19)$$

where $\mathbf{b}_k \in \mathbb{C}^{M\times 1}$ (k-th column of the matrix $\mathbf{B}$) is the unnormalized beamforming vector of user k. This definition of SINR does not correspond to any achievable rate. Nevertheless, it is widely used in the signal processing literature and enables us to formulate the weighted sum rate in a tractable way (it is therefore an heuristic approach). The new objective function is given by

$$f(\mathbf{B}) = \sum_{k=1}^{K} w_k \log\left(1 + \frac{\mathbf{b}_k^{\mathsf{H}}\boldsymbol{\Sigma}_k\mathbf{b}_k}{\mathrm{tr}(\mathbf{B}^{\mathsf{H}}\mathbf{B})/P + \sum_{j\neq k}\mathbf{b}_j^{\mathsf{H}}\boldsymbol{\Sigma}_k\mathbf{b}_j}\right) \qquad (20)$$

The gradient of f with respect to $\mathbf{b}_k$ is given by

$$\frac{\partial f}{\partial \mathbf{b}_k} = \frac{w_k}{d_k}\boldsymbol{\Sigma}_k\mathbf{b}_k - \frac{\mathrm{tr}(\mathbf{D})}{P}\mathbf{b}_k - \sum_j [\mathbf{D}]_{j,j}\boldsymbol{\Sigma}_j\mathbf{g}_k \qquad (21)$$

where $[\mathbf{D}]_{j,j}$ denotes the j-th diagonal element of $\mathbf{D}$ given by

$$\mathbf{D} = \mathrm{diag}\left(\frac{w_k\mathbf{b}_k^{\mathsf{H}}\boldsymbol{\Sigma}_k\mathbf{b}_k}{d_k(d_k + \mathbf{b}_k^{\mathsf{H}}\boldsymbol{\Sigma}_k\mathbf{b}_k)}\right) \qquad (22)$$

and d_k denotes the denominator of SINR_k. We proposed [10] the following iterative algorithm in order to maximize the weighted sum-rate in (20).

Algorithm 3: Initialize $\mathbf{B}^{(0)} = \frac{\boldsymbol{\alpha}^{\mathsf{H}}}{\sqrt{\mathrm{tr}(\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathsf{H}})}}$. For $n = 1, 2, \ldots$ do:

$$\mathbf{b}_k^{(n)} = \frac{w_k}{d_k^{(n-1)}}\left(\frac{\mathrm{tr}(\mathbf{D}^{(n-1)})}{P}\mathbf{I} + \sum_j [\mathbf{D}]_{j,j}^{(n-1)}\boldsymbol{\Sigma}_j\right)^{-1}\boldsymbol{\Sigma}_k\mathbf{b}_k^{(n-1)}$$

and update $\mathbf{D}^{(n+1)}$ correspondingly. △

This algorithm can be included into the greedy user selection Algorithm 2.

As a term of comparison, we consider a RB scheme with SINR feedback, where the feedback link is *noiseless* (the index of the best beam and its strength is sent back quantized and encoded over the feedback link). Thus, the only suboptimality in this case is due to the delay in the feedback. We feel that this is a fair comparison and reflects the fact that RB needs somehow less feedback: for every user, it needs only the index of the random beam over which the user achieves the best SINR, and the corresponding SINR value. However, we shall see that the delay in the feedback, alone, is able to prevent the RB scheme to work properly, as it ends up scheduling users in the wrong beams.

Fig. 2 shows averaged delay vs. the sum arrival rate for DPC (optimal) and linear schemes (greedy-ZFWF and greedy-SVH). We evaluate also the average delay performance of the SDMA/TDMA policy of Algorithm 3 and of RB as a function of the mobile speed v by letting the total arrival rate fixed to 6.0 bit/channel use. We consider a 20-user system with $M = 1, 2, 4$ antennas ($M = 1$ refers to a single-antenna TDMA system). The downlink SNR is set to be 20dB. For the analog feedback scheme, the uplink SNR is also assumed 20dB. Fig. 3 and 4 show the average delay vs. the mobile speed v km/h. The overall superiority of the newly proposed Algorithm 3 is clearly demonstrated.

References

[1] G. Caire, S. S. (Shitz), Y. Steinberg, and H. Weingarten, *On Information Theoretic Aspects of MIMO-Broadcast Channels*. Cambridge Univ. Press, 2005.

[2] N.Jindal, W.Rhee, S.Vishwanath, S.A.Jafar, and A.Goldsmith, "Sum Power Iterative Waterfilling for Multi-Antenna Gaussian Broadcast Channels," *IEEE Trans. on Inform. Theory*, vol. 51, no. 4, April 2005.

[3] E.M.Yeh and A.S.Cohen, "Information Theory, Queueing, and Resource Allocation in Multi-user Fading Communications," *Proceedings of the 2004 Conference on Information Sciences and Systems,Princeton, NJ*, March 2004.

[4] H.Boche and M.Wiczanowski, "Stability Optimal Transmission Policy for the Multiple Antenna Multiple Access Channel in the Geometric View," *To appear in EURASIP Signal Processing Journal, Special Issue on Advances in Signal Processing-assisted Cross-layer Designs*, 2006.

[5] M. Kobayashi and G. Caire, "An Iterative Waterfilling Algorithm for Maximum Weighted Sum-Rate of Gaussian MIMO-MAC and MIMO-BC," *Submitted to "IEEE J. Select. Areas Commun.", Special Issue on Nonlinear Optimization*, September 2005.

[6] M. Stojnic, H. Vikalo, and B. H. Hassibi, "Rate maximization in multi-antenna broadcast channels with linear preprocessing," in *IEEE Global Telecommunications Conference (Globecom 2004)*, Dallas, Texas, USA, November 29 – December 3 2004.

[7] F. Boccardi, F. Tosato, and G. Caire, "Precoding Schemes for the MIMO-GBC," *Int. Zurich Seminar on Commun., Zurich, Switzerland*, February 2006.

[8] G. Dimic and N. Sidiropoulos, "On Downlink Beamforming with Greedy User Selection: Performance Analysis and Simple New Algorithm," IEEE Trans. on Sig. Proc., vol. 53, no. 10, pp. 3857–3868, October 2005.

[9] T. Yoo and A. Goldsmith, "Optimality of zero-forcing beamforming with multiuser diversity," in *Proc. of ICC 2005*, vol. 1, May 2005, pp. 542–546.

[10] M. Kobayashi and G. Caire, "Joint beamforming and scheduling for a MIMO downlink with random arrivals," January 2006, submited to ISIT 2006, Seattle, WA.

[11] R. Zamir, S. Shamai (Shitz), and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1250–1276, June 2002.

[12] C. Peel, B. Hochwald, and A. Swindlehurst, "A vector-perturbation technique for near-capacity multi-antenna multi-user communication-part i: Channel inversion and regularization," *IEEE Transactions on Communications*, vol. 53, no. 1, pp. 195–202, January 2005, b.M. Hochwald, C.B. Peel, and A.L. Swindlehurst, "A Vector-Perturbation Technique for Near-Capacity Multi-Antenna Multi-User Communication-Part II: Perturbation," *IEEE Transactions on Communications*, Vol. 53, No. 3, pp. 537–544, March 2005.

[13] A. Lapidoth, S. Shamai, and M. Wigger, "On the capacity of fading MIMO broadcast channels with imperfect transmitter side-information," in *Proc. of 43rd Annual Allerton Conf.*, Monticello, IL, September 2005.

[14] N.Jindal, "MIMO Broadcast Channels with Finite Rate Feedback," *IEEE Global Telecommunications Conference (Globecom)*, November 2005.

[15] P.Viswanath, D.N.C.Tse, and R.Laroia, "Opportunistic Beamforming Using Dumb Antennas," *IEEE Trans. on Inform. Theory*, vol. 48, no. 6, June 2002.

[16] T.L.Marzetta and B.M.Hochwald, "Fast Transfer of Channel State Information in Wireless Systems," *Submitted to "IEEE Transactions on Signal Processing"*, June 2004.

[17] M.Schubert and H.Boche, "Solution of Multiuser Downlink Beamforming Problem With Individual SINR Constraints," *IEEE Trans. on Vehic. Tech.*, vol. 53, January 2004.

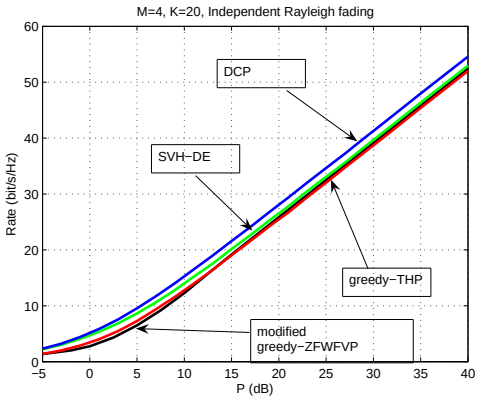

Fig. 1. Average throughput for a $M = 4, K = 20$ system.

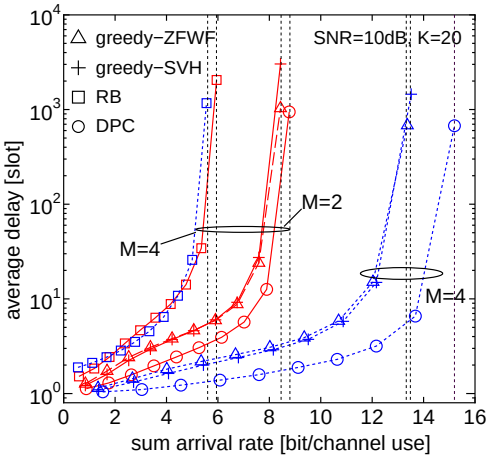

Fig. 2. Average delay vs. sum arrival rate (perfect CSIT).

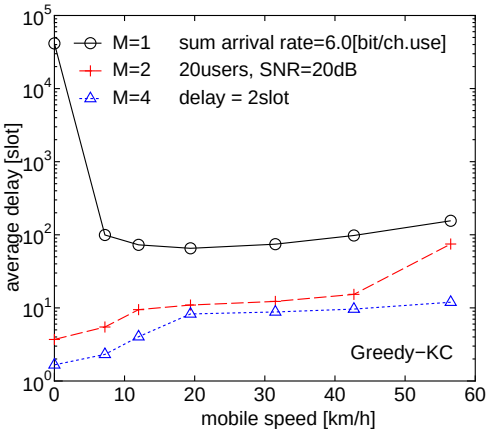

Fig. 3. Average delay vs. mobile speed for Algortihm 3 with greedy user selection.

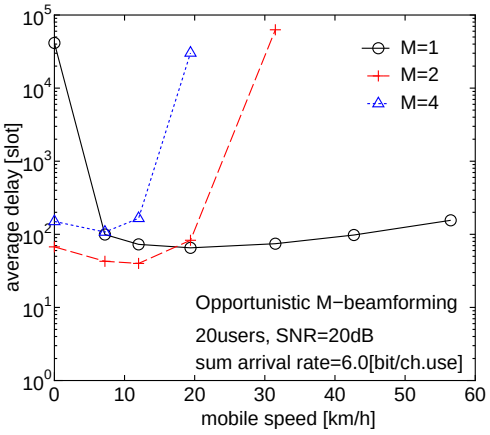

Fig. 4. Average delay vs. mobile speed for random beamforming.