Joint Beamforming and Scheduling for a Multi-Antenna Downlink with Imperfect Transmitter Channel Knowledge

Mari Kobayashi* and Giuseppe Caire†

∗Centre Tecnològic de Telecomunicacions de Catalunya
Barcelona, Spain
E-mail: mari.kobayashi@cttc.es

† University of Southern California,
Los Angeles, CA 90089
E-mail: caire@usc.edu

January 11, 2007

Abstract

We consider the downlink of a wireless system where the base-station has $M \geq 1$ antennas and K user terminals have one antenna each. We study the weighted rate sum maximization in the case of non-perfect Channel State Information at the Transmitter (CSIT). Some relevant downlink optimization problems, such as the stabilization of the transmission queues under random packet arrivals and the proportional fair scheduling for infinite backlogged systems, can be solved as special cases of the proposed problem. We restrict the transmitter strategy to be based on Gaussian coding and beamforming. Even under this simplifying condition, the problem at hand is non-convex and it does not appear to lend itself to a simple algorithmic solution. Therefore, we introduce some approximations that yield a definition of signal-to-interference plus noise ratio (SINR) commonly used in the classical array-processing/beamforming literature. For the simpler (but still non-convex) approximated problem, we propose a powerful heuristic solution based on greedy user selection and a gradient iteration that converges to a local maximum of the objective function. This method yields very competitive results with relatively low computational complexity. Extensive simulations show that, in the case of perfect CSIT, the proposed heuristic scheme performs very closely to the optimal (dirty-paper coding) strategy while, in the case of non-perfect CSIT, it significantly outperforms previously proposed suboptimal approaches, such as random beamforming and approximated zero-forcing with greedy user selection.

Keywords : MIMO Gaussian Broadcast Channel, Weighted Sum Rate, Non-perfect CSIT, Linear Beamforming, Scheduling

1 Introduction

In modern data-oriented wireless communications traffic is highly asymmetric. As in today wired internet, the downlink is recognized to be the system bottleneck since users are likely to use the network mostly to download large files and access multimedia streaming services. On the other hand, differently from traditional mobile telephony that requires low bit-rates but very strict delay constraints, such data-oriented communications are generally delay-tolerant. These considerations motivated a number of recent proposals for high-rate data-oriented downlink schemes.

Coarsely speaking, we can identify two "application-motivated" settings: 1) systems with random arrivals and transmission queues [1–5]; 2) systems with infinite backlog [6–10]. In the following we provide an informal discussion of data-oriented downlink schemes in order to motivate the work presented in this paper.

We assume a wireless downlink channel with K single-antenna receivers (users) and one transmitter (base station) equipped with $M \geq 1$ antennas. The complex baseband model is defined by

$$\mathbf{y}_k(t) = \mathbf{h}_k^H(t)\mathbf{X}(t) + \mathbf{n}_k(t) \tag{1}$$

where $k = 1, \ldots, K$ denotes the user index, time is slotted with a slot duration of T channel uses and the time index t ticks at the slot time, $\mathbf{y}_k(t) \in \mathbb{C}^{1\times T}$ is the signal received by user k in slot t, $\mathbf{n}_k(t) \in \mathbb{C}^{1\times T}$ is the corresponding Additive White Gaussian Noise (AWGN), with i.i.d. components $\sim \mathcal{CN}(0, 1)$, $\mathbf{X}(t) \in \mathbb{C}^{M\times T}$ is the transmitted signal space-time array, subject to the input power constraint

$$\frac{1}{T}\mathrm{tr}(\mathbf{X}(t)^H\mathbf{X}(t)) \leq P \tag{2}$$

and $\mathbf{H}(t) = [\mathbf{h}_1(t), \ldots, \mathbf{h}_K(t)] \in \mathbb{C}^{M\times K}$ denotes the matrix of fading channel coefficients. We adopt an ergodic block-fading model where $\mathbf{H}(t)$ is constant on each slot and changes from slot to slot according to some stationary and ergodic process. Furthermore, we assume that T is large enough such that very powerful capacity-approaching codes can be assumed.

At each slot t, some downlink resource allocation policy serves the users with rates $\{R_k(t) \geq 0 : k = 1, \ldots, K\}$ (bit/channel use). We say that a user is *scheduled* (or "served") at slot t if $R_k(t) > 0$. We also assume that the users are aware of the rate allocation choice (e.g., via some control channel that is not taken into account here). For the time being, assume that these rates are allocated according to some Channel State Information available at the Transmitter (CSIT), and that when rate $R_k(t)$ is allocated to user k, then this user successfully receives a packet of $TR_k(t)$ information bits.

In the first setting above (see Fig. 1), the main problem consists of achieving the *stability* of the transmission queues: the main goal of the downlink resource allocation policy is to stabilize the transmit queues such that they all have a finite average buffer size [11]. Notice that, by Little's theorem, finite average size implies finite average delay, that is clearly a very desirable property in a system with random arrivals.

For the second setting above (see Fig. 1), the notion of stability is irrelevant since all data are already present at the transmitter (infinite backlog). However, for self-evident reasons, the channel must be fairly shared among the users. Hence, the main goal of the downlink resource allocation policy is to maximize the average throughput subject to some *fairness* criterion.

In very general terms, a downlink resource allocation policy is defined by a *signaling scheme* $\mathcal{S}$ and a *rate scheduling algorithm* $\mathcal{A}$. The signaling scheme defines the type of coding that can be used (this includes, for example, orthogonal signaling, superposition coding, spatial beamforming, other forms of spatial precoding, etc..). The rate scheduling algorithm chooses at each slot t a rate K-tuple $\mathbf{R}(t) = \{R_k(t)\}$ as a function of CSIT and some user priority parameters (weights) that will be specified later. Therefore, it determines the users that are scheduled in that slot. Let $\mathcal{R}_{\mathcal{S}}(t)$ denote the region of rates that can be *reliably* transmitted over slot t (notice: this depends on $\mathcal{S}$ and on the channel conditions at slot t). The scheduled rate vector $\mathbf{R}(t)$ must belong to $\mathcal{R}_{\mathcal{S}}(t)$, otherwise, the condition of vanishing error probability would not be satisfied for some users. A scheduling algorithm $\mathcal{A}$

that satisfies $\mathbf{R}(t) \in \mathcal{R}_{\mathcal{S}}(t)$ for all t, is said to be *feasible* for the signaling scheme $\mathcal{S}$. The set of all feasible scheduling algorithms shall be denoted by $\mathcal{F}_{\mathcal{S}}$.

We define the region $\Omega_{\mathcal{S}}$ of achievable *ergodic* rates (also referred to as "long-term average throughput region") for a given signaling scheme $\mathcal{S}$ as

$$\Omega_{\mathcal{S}} = \bigcup_{\mathcal{A} \in \mathcal{F}_{\mathcal{S}}} \left\{ \overline{\mathbf{R}} \in \mathbb{R}_+^K : \overline{R}_k \leq \liminf_{\tau \to \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} R_k(t), \ \ \forall\, k \right\} \tag{3}$$

Notice that $\Omega_{\mathcal{S}}$ is convex, since for any two scheduling algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ in $\mathcal{F}_{\mathcal{S}}$ and for all $\beta \in [0,1]$ the time-sharing scheduling $\mathcal{A}$ that operates according to $\mathcal{A}_1$ with probability β and according to $\mathcal{A}_2$ with probability $1 - \beta$ is also feasible.

If queue stability is the goal, it can be shown that for a certain class of problems where the arrival processes and the channel states are jointly stationary and ergodic (see for example [4, 11, 12]) the set of arrival rates (expressed in bit per channel use) for which the system can be stabilized coincides with $\Omega_{\mathcal{S}}$. Furthermore, letting $\{Q_k(t)\}$ denote the sizes of the users' queue lengths (expressed in bits) and $\{\theta_k\}$ denote a set of positive weights, the rate scheduling algorithm $\mathcal{A}$ that for each t solves the maximization problem

$$\max \ \sum_{k=1}^{K} \theta_k Q_k(t) R_k(t), \ \ \text{subject to} \ \ \mathbf{R}(t) \in \mathcal{R}_{\mathcal{S}}(t) \tag{4}$$

achieves stability for any arrival rate K-tuple $\boldsymbol{\lambda} \in \Omega_{\mathcal{S}}$, even without the explicit knowledge of $\boldsymbol{\lambda}$ (i.e., it achieves stability in an "adaptive" way). The queue lengths evolve according to the stochastic difference equation

$$Q_k(t+1) = \max\{Q_k(t) - TR_k(t), 0\} + A_k(t) \tag{5}$$

where $A_k(t)$ denotes the arrival process (number of bits arrived at the k-th queue at the beginning of slot t).

If maximum throughput subject to fairness is the goal, solutions differ according to the definition of fairness [9]. A popular criterion is *proportional fairness*, implemented in the 1xEV-DO system, the data-oriented downlink scheme of CDMA2000 [6,7,13]. A Proportional

Fair Scheduling (PFS) algorithm maximizes $\sum_{k=1}^{K} \log \overline{R}_k$ subject to $\overline{\mathbf{R}} \in \Omega_{\mathcal{S}}$ [7]. Equivalently, denoting by $\overline{\mathbf{R}}^{\star}$ and $\overline{\mathbf{R}}$ the ergodic rates achieved by PFS and by any arbitrary $\mathcal{A} \in \mathcal{F}_{\mathcal{S}}$, the inequality

$$\sum_{k=1}^{K} \frac{\overline{R}_k}{\overline{R}_k^{\star}} \leq K \tag{6}$$

holds [14]. The rate scheduling algorithm that allocates at each t the rate solution of the maximization problem

$$\max \sum_{k=1}^{K} \frac{R_k(t)}{\overline{R}_k^{\star}}, \text{ subject to } \mathbf{R}(t) \in \mathcal{R}_{\mathcal{S}}(t) \tag{7}$$

is a PFS [14]. In practice, since $\overline{\mathbf{R}}^{\star}$ is generally unknown, a causal version of the above algorithm is obtained by replacing $\overline{\mathbf{R}}^{\star}$ with the empirical time-averaged rates $\{\widetilde{R}_k(t)\}$ defined by the stochastic difference equation

$$\widetilde{R}_k(t+1) = (1 - 1/t_c)\widetilde{R}_k(t) + (1/t_c)R_k(t). \tag{8}$$

The parameter t_c governs the time interval over which fairness is imposed (see discussion in [6, 7]).

From the above discussion it follows that, for both the stability and the proportional fairness problems, the desired rate scheduling algorithm is the result of a weighted rate maximization problem of the form

$$\max \sum_{k=1}^{K} w_k(t) R_k(t), \text{ subject to } \mathbf{R}(t) \in \mathcal{R}_{\mathcal{S}}(t) \tag{9}$$

for some non-negative time-varying weights $\{w_k(t)\}$ that are recursively computed. Most previous work has been focused on the ideal case of perfect CSIT. In the case of a single antenna at the transmitter, the underlying information theoretic model is the classical degraded Gaussian broadcast channel [15], the optimal signaling scheme $\mathcal{S}$ is known to be Gaussian superposition coding and $\mathcal{R}_{\mathcal{S}}(t)$ coincides with the capacity region of the Gaussian broadcast channel with channel coefficients $h_1(t), \ldots, h_K(t)$. Downlink schemes are driven by the theoretical results on the ergodic capacity region of the fading broadcast channel [16]. When the transmitter is equipped with $M > 1$ antennas, the underlying information theoretic model is the MIMO Gaussian broadcast channel (see for example [17–20]). In this case the optimal signaling scheme $\mathcal{S}$ is given by a combination of beamforming (i.e., linear precoding), Gaussian coding and an interference pre-cancellation strategy at the transmitter known as "Dirty-Paper Coding" (DPC). The maximization of the weighted sum rate under the optimal signaling scheme was addressed in [21] and recently in [22].

When CSIT is non-perfect, the problem becomes much more complicated and the exact solution of (9) for a general CSIT model is yet unknown. There are at least two main theoretical hurdles that make the sought solution hard: 1) for a general CSIT model the optimal signaling scheme $\mathcal{S}$ is not known; 2) even if we fix the signaling scheme $\mathcal{S}$, the corresponding region $\mathcal{R}_{\mathcal{S}}(t)$ might be non-convex. Therefore, the maximization in (9) yields a non-convex problem, for which we have no simple algorithmic solution.

The above difficulties call for some *good* and computationally simple heuristic solution. Several such solutions have been proposed. For example, the schemes based on random beamforming and SNR feedback, in [7] and [8] may be regarded as heuristics to solve (9). Other popular heuristics are based on some form of approximated zero-forcing beamforming computed with respect to the information provided by the CSIT (e.g., using quantized channel feedback [23, 24], or some more general CSIT model [25]). These heuristics are essentially based on the following approach: first, fix the signaling scheme $\mathcal{S}$ in some sensible way and then, solve (9) in order to determine the corresponding rate scheduling algorithm $\mathcal{A}$. This paper is no exception. We propose a new algorithm that can be applied when the channel second-order statistics (covariance) conditioned on the CSIT is known. Our method, albeit not optimal, has some nice features. In particular, it achieves results close to optimal DPC when CSIT is perfect, and significantly outperforms previously proposed methods when CSIT is non-perfect.

2 Simplifications and Approximations

We denote by $\boldsymbol{\alpha}(t)$ the CSIT available to the transmitter at time t. We assume that the channel state process $\{\mathbf{H}(t)\}$ and the CSIT process $\{\boldsymbol{\alpha}(t)\}$ are jointly stationary and ergodic. Furthermore, we restrict the transmitter to use *instantaneous* scheduling policies that allocate the rates $\mathbf{R}(t)$ based on $\boldsymbol{\alpha}(t)$, i.e., by disregarding the observation of the past CSIT values $\boldsymbol{\alpha}(1), \ldots, \boldsymbol{\alpha}(t-1)$. This restriction is motivated by the current proposals for practical CSIT schemes. These can be roughly grouped into the following families: 1) quantized channel feedback [23–25], where $\boldsymbol{\alpha}(t)$ is a quantized version of the channel matrix $\mathbf{H}(t-d)$ and d is the feedback loop delay (in slots)[1], 2) analog feedback [26–29], where $\boldsymbol{\alpha}(t)$ is a noisy version of $\mathbf{H}(t)$.[2]

In this work we fix the signaling scheme $\mathcal{S}$ to be *Gaussian superposition coding*. Furthermore, we assume that the receivers do not perform multiuser decoding (e.g., interference cancellation), that is, each user k treats the interference from signals destined to other users $j \neq k$ as noise. Given the current value of CSIT $\boldsymbol{\alpha}$ (where hereafter we drop the slot index t due to stationarity), the transmitted space-time signal is given by

$$\mathbf{X}(\boldsymbol{\alpha}) = \sum_{k=1}^{K} \mathbf{X}_k(\boldsymbol{\alpha}) \tag{10}$$

where the codewords $\mathbf{X}_k(\boldsymbol{\alpha}) \in \mathbb{C}^{M \times T}$ are conditionally mutually independent given $\boldsymbol{\alpha}$ and, for each k, the columns of $\mathbf{X}_k(\boldsymbol{\alpha})$ are i.i.d., $\sim \mathcal{CN}(\mathbf{0}, \mathbf{S}_k(\boldsymbol{\alpha}))$. For later use, we introduce the matrix-valued function $\mathbf{S}(\boldsymbol{\alpha}) = \{\mathbf{S}_k(\boldsymbol{\alpha}) : k = 1, \ldots, K\}$ that maps the CSIT $\boldsymbol{\alpha}$ into the transmit signal covariance matrices.

For given $\boldsymbol{\alpha}, \mathbf{H}$ and covariance matrices $\mathbf{S}(\boldsymbol{\alpha})$, the rate scheduling algorithm $\mathcal{A}$ must

[1] As a matter of fact, most works on quantized channel feedback consider only the case $d = 0$ or, equivalently, assume implicitly a very slowly-varying fading channel. In practice a feedback delay of at least one slot is necessary, and often d is even larger.

[2] This case includes time-division duplex, where the uplink channel is used to estimate the downlink channel coefficients by exploiting the reciprocity of the radio channel [26, 30].

choose the user rates $\mathbf{R}$ based on $\boldsymbol{\alpha}$. Notice that $\mathbf{H}$ is known to the transmitter only *statistically*, through its conditional first-order probability distribution $p(\mathbf{H}|\boldsymbol{\alpha})$, assumed to be known. The signal to interference plus noise ratio (SINR) seen at receiver k is given by

$$\mathsf{SINR}_k(\mathbf{H}, \mathbf{S}(\boldsymbol{\alpha})) = \frac{\mathbf{h}_k^H \mathbf{S}_k(\boldsymbol{\alpha})\mathbf{h}_k}{1 + \sum_{j \neq k} \mathbf{h}_k^H \mathbf{S}_j(\boldsymbol{\alpha})\mathbf{h}_k} \tag{11}$$

Following the same argument as in [31] (see also [28]), it can be shown that the non-perfect CSIT case can be reduced to an equivalent "virtual" perfect CSIT case where $\boldsymbol{\alpha}$ plays the role of the new channel state and the instantaneous user rates are replaced by the average "outage" rates

$$R_k = \max_{r \geq 0} \; r \left(1 - \Pr\left(\log(1 + \mathsf{SINR}_k(\mathbf{H}, \mathbf{S}(\boldsymbol{\alpha}))) \leq r | \boldsymbol{\alpha}\right)\right) \tag{12}$$

This approach is taken, for example, in [31] for a class of simple signaling strategies and in [28] for the Gaussian superposition coding strategy considered here. However, in the latter case the outage rate maximization in (12) is complicated and requires heavy numerical computation.

On the other hand, practical systems such as 1xEV-DO [6,13] make use of fast incremental redundancy coding, in order to cope with the residual uncertainty of $\mathbf{H}$ given $\boldsymbol{\alpha}$: when a user is scheduled to transmit in a slot with given nominal rate (allocated based on $\boldsymbol{\alpha}$), the effective coding rate is adapted such that, eventually, it is slightly less than $\log(1 + \mathsf{SINR}_k(\mathbf{H}, \mathbf{S}(\boldsymbol{\alpha})))$, even though the latter is not known. This can be done by sending parity symbols until the receiver sends an "ACK" back to the transmitter, via an uplink control channel.

Somehow optimistically, in this work we "postulate" the existence of an ideal rate adaptation scheme of this kind, and assume that the rates $\log(1+\mathsf{SINR}_k(\mathbf{H}, \mathbf{S}(\boldsymbol{\alpha})))$ are achievable. This corresponds to considering the new objective function

$$\sum_{k=1}^{K} w_k \mathbb{E}\left[\log\left(1 + \mathsf{SINR}_k(\mathbf{H}, \mathbf{S}(\boldsymbol{\alpha}))\right) | \boldsymbol{\alpha}\right] \tag{13}$$

for the sake of weighted rate sum maximization. This assumption provides an upperbound to the outage rate case (12), since it is easy to see that for any $\boldsymbol{\alpha}$

$$\max_{r\geq 0}\, r\left(1-\Pr\left(\log(1+\mathsf{SINR}_k(\mathbf{H},\mathbf{S}(\boldsymbol{\alpha})))\leq r|\boldsymbol{\alpha}\right)\right)\leq \mathbb{E}\left[\log(1+\mathsf{SINR}_k(\mathbf{H},\mathbf{S}(\boldsymbol{\alpha})))|\boldsymbol{\alpha}\right] \qquad (14)$$

Under the above simplifications, finding the optimal scheduling algorithm $\mathcal{A}$ is reduced to the maximization of (13) with respect to $\mathbf{S}(\boldsymbol{\alpha})$, subject to the transmitted power constraint (2), that yields the sum trace constraint

$$\sum_{k=1}^{K}\mathrm{tr}(\mathbf{S}_k(\boldsymbol{\alpha}))\leq P,\;\; \mathbf{S}_k(\boldsymbol{\alpha})\geq \mathbf{0}\;\;\forall\, k,\;\;\forall\,\boldsymbol{\alpha} \qquad (15)$$

Unfortunately, a simple solution to this problem seems to be unavailable. In order to obtain a *tractable* problem, we shall introduce further restrictions and approximations. In particular, we replace the average user rates in (13) with the approximated rates given by

$$\begin{aligned}\mathbb{E}\left[\log\left(1+\frac{\mathbf{h}_k^H\mathbf{S}_k(\boldsymbol{\alpha})\mathbf{h}_k}{1+\sum_{j\neq k}\mathbf{h}_k^H\mathbf{S}_j(\boldsymbol{\alpha})\mathbf{h}_k}\right)\middle|\,\boldsymbol{\alpha}\right] &\approx \log\left(1+\frac{\mathbb{E}[\mathbf{h}_k^H\mathbf{S}_k(\boldsymbol{\alpha})\mathbf{h}_k|\boldsymbol{\alpha}]}{1+\sum_{j\neq k}\mathbb{E}[\mathbf{h}_k^H\mathbf{S}_j(\boldsymbol{\alpha})\mathbf{h}_k|\boldsymbol{\alpha}]}\right)\\ &= \log\left(1+\frac{\mathrm{tr}(\boldsymbol{\Sigma}_k(\boldsymbol{\alpha})\mathbf{S}_k(\boldsymbol{\alpha}))}{1+\sum_{j\neq k}\mathrm{tr}(\boldsymbol{\Sigma}_k(\boldsymbol{\alpha})\mathbf{S}_j(\boldsymbol{\alpha}))}\right) \quad (16)\end{aligned}$$

where we define the channel conditional covariance matrix $\boldsymbol{\Sigma}_k(\boldsymbol{\alpha})=\mathbb{E}[\mathbf{h}_k\mathbf{h}_k^H|\boldsymbol{\alpha}]$. Since we applied Jensen's inequality to both the numerator and the denominator of the SINR, (16) yields just an approximation of the conditional expected rates. Furthermore, we shall restrict the signaling scheme to the class of "beamforming" Gaussian superposition coding, that is, the covariance matrices $\mathbf{S}_k(\boldsymbol{\alpha})$ are constrained to have rank at most 1. In particular, the transmitted signal is given by

$$\mathbf{X}(\boldsymbol{\alpha})=\mathbf{G}(\boldsymbol{\alpha})\mathbf{U} \qquad (17)$$

where $\mathbf{G}(\boldsymbol{\alpha})=[\mathbf{g}_1(\boldsymbol{\alpha}),\ldots,\mathbf{g}_K(\boldsymbol{\alpha})]\in\mathbb{C}^{M\times K}$ is the beamforming matrix and $\mathbf{U}\in\mathbb{C}^{K\times T}$ is a Gaussian i.i.d. matrix with elements $\sim\mathcal{CN}(0,1)$ that contains on each k-th row the codeword intended to user k. This corresponds to considering input covariances in the form $\mathbf{S}_k(\boldsymbol{\alpha})=\mathbf{g}_k(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})^H$. The transmit power constraint is given by

$$\mathrm{tr}\left(\mathbf{G}(\boldsymbol{\alpha})^H\mathbf{G}(\boldsymbol{\alpha})\right)\leq P,\quad\forall\,\boldsymbol{\alpha} \qquad (18)$$

Plugging the beamforming covariances into (16), we obtain the average rate approximation $\log(1 + \widetilde{\mathsf{SINR}}_k(\boldsymbol{\alpha}))$ where

$$\widetilde{\mathsf{SINR}}_k(\boldsymbol{\alpha}) = \frac{\mathbf{g}_k(\boldsymbol{\alpha})^H \boldsymbol{\Sigma}_k(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{1 + \sum_{j\neq k} \mathbf{g}_j(\boldsymbol{\alpha})^H \boldsymbol{\Sigma}_k(\boldsymbol{\alpha})\mathbf{g}_j(\boldsymbol{\alpha})} \tag{19}$$

It is interesting to notice that the definition of SINR given in (19) is commonly used in the beamforming/signal processing literature (see for example [32–34]). These works address the problem of minimum transmit power subject to quality-of-service constraints. This can be formulated as

$$\min P, \;\; \text{subject to } \widetilde{\mathsf{SINR}}_k(\boldsymbol{\alpha}) \geq \gamma_k, \;\; \forall\, k, \;\; \text{and to } \operatorname{tr}\left(\mathbf{G}(\boldsymbol{\alpha})^H \mathbf{G}(\boldsymbol{\alpha})\right) \leq P \tag{20}$$

where $\gamma_k \geq 0$ are the users' target SINR requirements (quality-of-service). Although never said explicitly, it should be noticed that the SINR constraint in (20) does not correspond to any rigorous notion of information-theoretic achievable rates: the connection between achievable rates and the "signal-processing" SINR defined in (19) is provided by the approximations made in this section.

Also, it should be noticed that (20) can be reduced to a convex problem and it can be solved using standard tools such as uplink-downlink duality and semi-definite programing [32–34]. On the contrary, the *approximated* weighted rate sum maximization

$$\max_{\mathbf{G}(\boldsymbol{\alpha})} \sum_{k=1}^{K} w_k \log\left(1 + \widetilde{\mathsf{SINR}}_k(\boldsymbol{\alpha})\right), \;\; \text{subject to } \operatorname{tr}\left(\mathbf{G}(\boldsymbol{\alpha})^H \mathbf{G}(\boldsymbol{\alpha})\right) \leq P \tag{21}$$

is a non-convex problem. In the next section we provide a heuristic method to solve (21). Then, in Section 4 we shall apply the proposed heuristic method to the original problems (queue stability and PFS) and compare the resulting performance to other known heuristic solutions.

As a final remark, we notice that for the special case of perfect CSIT ($\boldsymbol{\alpha} = \mathbf{H}$) we have $\boldsymbol{\Sigma}_k = \mathbf{h}_k\mathbf{h}_k^H$. Therefore the approximation in (16) holds with equality and the rates $\log(1 + \widetilde{\mathsf{SINR}}_k(\mathbf{H}))$ are actually achievable by downlink beamforming. The maximization of the sum rate (equal weights) is still a non-convex problem, and a heuristic algorithm was proposed in [35].

3 Heuristic rate scheduling algorithm

We can eliminate the constraint in (21) by letting the transmit signal be

$$\mathbf{X}(\boldsymbol{\alpha}) = \sqrt{\frac{P}{\mathrm{tr}(\mathbf{G}(\boldsymbol{\alpha})^H \mathbf{G}(\boldsymbol{\alpha}))}} \mathbf{G}(\boldsymbol{\alpha}) \mathbf{U} \tag{22}$$

where $\mathbf{G}(\boldsymbol{\alpha})$ is now unconstrained. For the sake of notation simplicity, in this section we drop the explicit dependency of $\boldsymbol{\Sigma}_k$ and $\mathbf{G}$ on $\boldsymbol{\alpha}$. The resulting *unconstrained* problem corresponding to (21) consists of maximizing the objective function

$$f(\mathbf{G}) = \sum_{k=1}^{K} w_k \log \left(1 + \frac{\mathbf{g}_k^H \boldsymbol{\Sigma}_k \mathbf{g}_k}{\mathrm{tr}(\mathbf{G}^H \mathbf{G})/P + \sum_{j \neq k} \mathbf{g}_j^H \boldsymbol{\Sigma}_k \mathbf{g}_j} \right) \tag{23}$$

As already noticed, $f(\mathbf{G})$ is not concave in $\mathbf{G}$. Our heuristics consist of finding an iterative method that provably converges to a local maximum of $f(\mathbf{G})$, and using it jointly with a greedy user selection in order to avoid being trapped in "bad" local maxima (this concept will be defined more precisely in the following).

3.1 Convergence to local maxima

The gradient of $f(\mathbf{G})$ with respect to $\mathbf{g}_k$ is given by

$$\frac{\partial f}{\partial \mathbf{g}_k} = \frac{w_k}{d_k} \boldsymbol{\Sigma}_k \mathbf{g}_k - \frac{\mathrm{tr}(\mathbf{D})}{P} \mathbf{g}_k - \sum_j [\mathbf{D}]_{j,j} \boldsymbol{\Sigma}_j \mathbf{g}_k \tag{24}$$

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

$$\mathbf{D} = \mathrm{diag} \left(\frac{w_k \mathbf{g}_k^H \boldsymbol{\Sigma}_k \mathbf{g}_k}{d_k (d_k + \mathbf{g}_k^H \boldsymbol{\Sigma}_k \mathbf{g}_k)} \right) \tag{25}$$

and where

$$d_k = \mathrm{tr}(\mathbf{G}^H \mathbf{G})/P + \sum_{j \neq k} \mathbf{g}_j^H \boldsymbol{\Sigma}_k \mathbf{g}_j \tag{26}$$

is the denominator of $\widetilde{\mathsf{SINR}}_k$.

Stationary points of $f(\mathbf{G})$ satisfy the zero-gradient condition $\nabla f(\mathbf{G}) = \mathbf{0}$, where

$$\nabla f(\mathbf{G}) = \left[\left(\frac{\partial f}{\partial \mathbf{g}_1} \right)^H, \left(\frac{\partial f}{\partial \mathbf{g}_2} \right)^H, \ldots, \left(\frac{\partial f}{\partial \mathbf{g}_K} \right)^H \right]^H$$

denotes the gradient of $f(\mathbf{G})$ (written as a column vector). If $\mathbf{G}$ is a stationary point, then it must satisfy the equation

$$\left[\frac{\mathrm{tr}(\mathbf{D})}{P} \mathbf{I} + \sum_j [\mathbf{D}]_{j,j} \mathbf{\Sigma}_j \right] \mathbf{g}_k = \frac{w_k}{d_k} \mathbf{\Sigma}_k \mathbf{g}_k \tag{27}$$

for all $k = 1, \ldots, K$. While solving explicitly for (27) is not possible, the following iterative algorithm is proposed in order to converge to a local maximum, solution of (27).

1. Let $\mathbf{G}^{(0)}$ be an initial choice for the beamforming matrix.

2. At iteration $n = 1, 2, \ldots$, for all $k = 1, \ldots, K$ compute

$$\mathbf{a}_k^{(n)} = \frac{w_k}{d_k^{(n-1)}} \left[\frac{\mathrm{tr}(\mathbf{D}^{(n-1)})}{P} \mathbf{I} + \sum_j [\mathbf{D}]_{j,j}^{(n-1)} \mathbf{\Sigma}_j \right]^{-1} \mathbf{\Sigma}_k \mathbf{g}_k^{(n-1)} \tag{28}$$

   and let

$$\mathbf{g}_k^{(n)} = \mu \mathbf{a}_k^{(n)} + (1 - \mu) \mathbf{g}_k^{(n-1)} \tag{29}$$

   where μ is the result of the line search

$$\mu = \arg \max_{\nu \in [0,1]} f \left(\nu [\mathbf{a}_1^{(n)}, \ldots, \mathbf{a}_K^{(n)}] + (1 - \nu) \mathbf{G}^{(n-1)} \right) \tag{30}$$

3. Update $\mathbf{D}^{(n)}$ and $d_k^{(n)}$ using (25) and (26) computed in $\mathbf{G} = \mathbf{G}^{(n)}$, respectively.

We have the following result:

Proposition 1. For all $\mathbf{G}^{(0)}$, the sequence of objective function values $f(\mathbf{G}^{(0)}), f(\mathbf{G}^{(1)}), f(\mathbf{G}^{(2)}), \ldots$ is non-decreasing.

Proof. See Appendix B. □

Since $f(\mathbf{G})$ is bounded for every finite P, then the sequence of objective function values generated by the algorithm converges. Unfortunately, since $f(\mathbf{G})$ may have several local maxima, the final point of the iterative algorithm depends in general on the initial condition. Several approaches can be followed here. For example, we may repeat the above recursion for several randomly chosen initial points and select the best found local maximum. Another approach consists of initializing $\mathbf{G}^{(0)}$ according to the dominant spatial direction of the channels. Suppose that $\boldsymbol{\Sigma}_k$ has dominant eigenvector $\mathbf{v}_k$. Then, a sensible choice for the initial point consists of $\mathbf{G}^{(0)} = [\mathbf{v}_1, \ldots, \mathbf{v}_K]^H$, that is, the spatial filter "matched" to the dominant channel directions. This choice is sensible if the CSIT yields directly the directions $\mathbf{v}_k$ without need for an eigenvalue decomposition, i.e., without any computational effort. For the analog feedback model illustrated in Appendix A this is indeed the case. In fact, we have used this initialization method in the simulations of Section 4.

Furthermore, we have also observed that the line search (30) can be avoided, and the value $\mu = 1$ can be safely used. Actually, we conjecture that the algorithm for $\mu = 1$ always converges to a local maximum.

3.2 Greedy user selection

A system is said to be interference limited if the weighed rate sum is bounded in the limit of large SNR. In the system at hands, $\widetilde{\mathsf{SINR}}_k$ is bounded for $P \to \infty$ if $\mathbf{g}_j^H \boldsymbol{\Sigma}_k \mathbf{g}_j > 0$ for some $j \neq k$. In particular, if the column space of $\boldsymbol{\Sigma}_k$ coincides with $\mathbb{C}^M$ then there is no "zero-forcing" direction such that we can have simultaneously $\widetilde{\mathsf{SINR}}_j \to \infty$ and $\widetilde{\mathsf{SINR}}_k \to \infty$ as $P \to \infty$. If all channel covariance matrices have rank M, any rate scheduling algorithm that serves more than one user per slot yields an interference limited system. For channel matrices of rank less than M the situation is more complicated. In particular, for perfect CSIT where all $\boldsymbol{\Sigma}_k = \mathbf{h}_k \mathbf{h}_k^H$ have rank 1, it is possible to find subsets of $1 \leq m \leq M$ *linearly* *independent* users that can be scheduled at the same time while the system is not interference limited.

This observation is the key to understand the need of greedy user selection in conjunction with the above iterative algorithm: for sufficiently large P the objective function $f(\mathbf{G})$ is maximized by scheduling only a subset of users. Unfortunately, the algorithm might converge to an interference limited local maximum, i.e., it might not be able to select a subset of linearly independent users. In this case, the gap between the absolute maximum and such local maxima can be made *arbitrarily large* for sufficiently large P. Hence, without some mechanism that prevents "bad" local maxima, the performance loss of the proposed iterative algorithm with respect to the optimum can be unbounded (in contrast with the claims made in [35] for a similar algorithm that suffers from the very same problem).

We conclude that in order to operate at high SNR with linear beamforming it is essential to select a good subset $\mathcal{K} \subseteq \{1, \dots, K\}$ of users [36–39]. Brute-force optimal user selection requires searching for all $\binom{K}{m}$ subsets of m out of K users, for all $m = 1, 2, \dots$. The problem of *efficient* optimal user selection is still open and only heuristic algorithms have been proposed. We shall use the following simple greedy scheme. Let $f_{\mathcal{K}}$ denote the objective function value obtained by forcing all users $j \notin \mathcal{K}$ to have $\mathbf{g}_j = \mathbf{0}$ (hence zero rate) and then running the iterative algorithm of Section 3.1 restricted to the user subset $\mathcal{K}$. Then, the greedy user selection algorithm works as follows: initialize $\mathcal{K}_0 = \emptyset$ and $f_\emptyset = 0$. For $m = 1, 2, \dots$, do:

1. Find
$$k_m = \arg \max_{k \notin \mathcal{K}_{m-1}} f_{\mathcal{K}_{m-1} \cup \{k\}}$$
2. If $f_{\mathcal{K}_{m-1} \cup \{k_m\}} < f_{\mathcal{K}_{m-1}}$, let $\mathcal{K} = \mathcal{K}_{m-1}$ and stop.
3. Otherwise, let $\mathcal{K}_m = \mathcal{K}_{m-1} \cup \{k_m\}$, and go to 1.

Notice that this greedy selection can be applied to any heuristic weighted rate sum maximization approach that converges to some local maximum of the objective function. For example, in [36,37] the greedy search is used with zero-forcing beamforming and waterfilling power allocation (in the case of perfect CSIT), and in the simulations of Section 4 we shall consider the same approach with an approximated zero-forcing beamforming based on the non-perfect CSIT.

As said before, when the channel covariance matrices have rank M and P is sufficiently large the objective function is maximized by letting only one user transmit (TDMA). In this case, the greedy user selection is optimal since it chooses the user with the largest weighted rate, given by

$$\begin{aligned} k &= \arg \max_{j=1,\ldots,K} w_j \log \left(1 + P \max_{\mathbf{g}_j : ||\mathbf{g}_j||=1} \mathbf{g}_j^H \boldsymbol{\Sigma}_j \mathbf{g}_j \right) \\ &= \arg \max_{j=1,\ldots,K} w_j \log \left(1 + P \lambda_{\max}(\boldsymbol{\Sigma}_j) \right) \end{aligned} \tag{31}$$

where $\lambda_{\max}(\boldsymbol{\Sigma}_j)$ denotes the maximum eigenvalue of $\boldsymbol{\Sigma}_j$. For the case of no CSIT, we have that $\boldsymbol{\Sigma}_k = \mathbb{E}[\mathbf{h}_k \mathbf{h}_k^H]$. Hence, the values $\lambda_{\max}(\boldsymbol{\Sigma}_k)$ are constants that depend only on the a priori "spatial distribution" of the users. If, in addition, the channels are identically and isotropically distributed (e.g., as in a rich scattering Rayleigh fading situation), then $\boldsymbol{\Sigma}_k = \mathbf{I}$ and (31) corresponds to scheduling the user with the largest weight. In the case of PFS, this yields round-robin scheduling, and in the case of random arrivals this yields the so-called *Serve-the-Largest-Connected-Queue* policy [11].

4 Numerical Results

4.1 Simulation setting

The channel and CSIT models are defined in Appendix A. We matched the Gauss-Markov correlation coefficient in (32) to Jakes' autocorrelation model, such that $r = J_0(2\pi v f_c T/c)$ where v is the mobile speed (m/sec) and f_c is the carrier frequency in Hz, c is the light speed (m/sec), T denotes the slot length (sec) and $J_0(\cdot)$ is the Bessel function of the first kind and order 0. Inspired by the HDR/1xEV-DO system, we considered a slot duration of 1.67 msec and a feedback delay of $d = 2$ slots. For the analog feedback scheme of Appendix A, we let the uplink SNR be equal to the downlink SNR, i.e. $P = 1/N_0^{\rm up}$.

Under the assumption of perfect CSIT (ideal noiseless and delay-free feedback) we considered: 1) the optimal DPC signaling strategy and rate scheduling implemented using the algorithm of [22], 2) the linear zero-forcing beamforming scheme with waterfilling power allocation, enhanced by greedy user selection as proposed by [36, 37] (referred to as"greedy-ZFWF"), 3) the algorithm proposed in this paper, that for the special case of perfect CSIT is very similar to that presented in [35] enhanced by the greedy user selection (referred to as "greedy-SVH" from the name of the authors of [35]), 4) a version of opportunistic beamforming scheme (RB) proposed in [7, 8] as described in [31]. For the RB scheme, we consider that the number of beams is fixed a priori and, if not specified otherwise, it is equal to the number of antennas $B = M$ as proposed in [8].

Under non-perfect CSIT, we compare the proposed algorithm (referred to as "greedy-KC" from the name of the authors of this paper) with a "greedy-ZFWF" that computes the precoding matrix as if $\boldsymbol{\alpha}$ was the true channel (mismatched zero-forcing), as considered for example in [23–25] and the RB scheme selecting users based on the SINR predicted by the same CSIT model of the other schemes. For the RB scheme, we make the optimistic assumption that the sequence of random beamforming matrices is known a priori to all the users [31]. Furthermore, since the RB does not require the transmitter to track the channel, we assume that the feedback link is *noiseless* so that the only source of nuisance is due to the feedback delay. We feel that this represents a fair comparison with other more complicated signaling schemes and reflects the fact that RB requires in general less feedback: each user sends back M SINR values (one for each random beam) based on the channel prediction $\boldsymbol{\alpha}$ (that can be computed at the transmitter) [31].

4.2 Sum rate vs. number of users under PFS

Here we consider the PFS scheduling. We chose $t_c = 5K$, that imposes fairness over a rather short time interval (e.g., this value would be appropriate for a multimedia streaming application, where users cannot tolerate long time intervals without receiving source bits). Fig.2 shows the average sum rate vs. the number of users K for the case of perfect CSIT. Fig.3 shows analogous curves for the case of non-perfect CSIT case with mobile speed $v = 10$ km/h, that yields the prediction MMSE $\sigma_e^2 = 0.05$ under an SNR of 20 dB. The performance of greedy-ZFWF is omitted since it coincides with the performance achieved by greedy-SVH. We observe that for this choice of t_c comparable to the number of users, the RB scheme with a single beam [7, 31] converges to a round-robin by serving one user after another and thus cannot exploit multiuser diversity. On the contrary, the systems based on multiple beams are able to achieve very large spectral efficiency since they can schedule multiple users simultaneously while maintaining the same degree of fairness. Remarkably, the proposed system performs close to the optimal DPC-based scheme in the case of perfect CSIT (especially if M is not too large), and yields dramatic improvements over RB for the case of non-perfect CSIT.

4.3 Average delay vs. sum arrival rate

For a system with queues and random arrivals we consider mutually independent Poisson arrival processes $A_k(t)$ with exponentially distributed packet length (in bit).

Fig. 4 shows the averaged delay (computed by Little's theorem) vs. the sum arrival rate for a system with $K = 20$ users and $M = 2, 4$ antennas under perfect CSIT. The average SNR is 10dB and we consider i.i.d. varying channel ($r = 0$ in the Gauss-Markov model). We remark that both greedy-ZFWF and greedy-SVH achieve near DPC performance especially for $M = 2$ although the performance gap increases for $M = 4$. With the RB scheme, the queue buffers overflow at much smaller arrival rate than other schemes. Notice also that for a 20-user system, the RB scheme with $M = 4$ antennas is harmful (much worse than $M = 2$) because it requires a dramatically larger number of users compared to the number of beams to exploit multiuser diversity.

4.4 Average delay vs. mobile speed

We evaluate the average delay performance of greedy-ZFWF, greedy-KC, and RB as a function of the mobile speed v by letting the total arrival rate fixed to $\lambda_{\text{sum}} = 6.0$ bit/channel use. In Fig. 5 we consider a 20-user system with $M = 4$ antennas ($M = 1$ refers to a standard single-antenna TDMA system). The SNR equal to 20dB. Under the given parameters, the mobile speeds $v = 0,\ 10,\ 15,\ 22,\ 33.5,\ 44.5,\ 58$ km/h yield the channel prediction MMSE $\sigma_e^2 = 0.00,\ 0.05,\ 0.10{,}0.20,\ 0.40,\ 0.60,\ 0.80$ respectively. A closed form expression for the d-step prediction MMSE that depends on the mobility is provided in Appendix A. As said before, for RB the feedback link is considered noise-free which gives the prediction MMSE slightly smaller than the analog feedback for a given mobile speed. These results show that greedy-KC and greedy-ZFWF achieve substantially equivalent performance and much better stability with respect to RB for low speed. However, when the Doppler bandwidth increases, greedy-KC is much more robust to channel prediction error than the greedy-ZFWF, that becomes rapidly mismatched. Interestingly, despite the fact that RB needs only SINR feedback, it is much more fragile to mobility (Doppler) than the other two schemes that require a more accurate CSIT.

5 Conclusions

We studied the weighted sum rate maximization problem in the MIMO-BC with linear beamforming strategies under non-perfect CSIT. We showed that both the adaptive policy that stabilizes all arrival rates in the system stability region and the PFS joint scheduling and beamforming policy for systems with infinite backlog can be compactly stated as a weighted rate sum maximization for suitable time-varying weights. The problem with Gaussian superposition coding and non-perfect CSIT appears to be very difficult to be treated in an exact manner. Therefore, we formulated a simplified problem where we adopted the definition of SINR commonly used in the signal-processing/beamforming literature. Although simpler, the resulting problem is still non-convex. Hence, we provided a simple greedy beamforming algorithm that yields very good results with relatively low complexity. This algorithm requires the knowledge of the channel covariance matrices given the CSIT at each slot time t. This is particularly suited with time-division duplex (TDD) or frequency-division duplex (FDD) with analog feedback, that we have discussed to some detail extent in Appendix A. However, our approach might also be applied to quantized feedback or other form of CSIT feedback as long as the channel vector conditional covariances can be characterized.

Numerical results confirm that under perfect CSIT our simple linear beamforming scheme achieves near DPC performance both in terms of sum rate and average delay. On the other hand, in the presence of feedback delay and noise, it is found that some previously proposed heuristic schemes (notably, those based on random beamforming or on mismatched zero-forcing beamforming) rapidly degrade, while the new scheme appears to be much more robust to non-perfect CSIT and therefore it is a good candidate for systems that have to handle some mobility, such as IEEE 802.16e (mobile WiMax).

APPENDIX

A Channel model and Analog Feedback

In our simulation we assume that the channel vectors $\mathbf{h}_k(t)$ are Gaussian with mean zero, unit variance per component and spatially white, that is $\sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. Furthermore, the channels are mutually statistically independent for different users.

We consider a stationary ergodic Gauss-Markov block-fading process, given by

$$\mathbf{h}_k(t) = r\mathbf{h}_k(t-1) + \boldsymbol{\nu}_k(t) \tag{32}$$

where $\boldsymbol{\nu}_k(t) \sim \mathcal{CN}(\mathbf{0}, (1-r^2)\mathbf{I})$ is a Gaussian i.i.d. process.

We assume that each receiver can estimate *exactly* its channel vector. Several methods for providing CSIT are proposed and implemented. In particular, when uplink and downlink use different frequency bands (FDD), the CSIT must be explicitly fed back from the users to the base station [23–29]. We hasten to say that the beamforming and rate scheduling algorithm proposed in this paper can be applied as long as the conditional covariances of the channel vectors, $\boldsymbol{\Sigma}_k(\boldsymbol{\alpha}(t)) = \mathbb{E}[\mathbf{h}_k(t)\mathbf{h}_k(t)^H|\boldsymbol{\alpha}(t)]$ are known. In this paper, however, we consider analog feedback [26, 27] for FDD systems, where the receivers send back to the transmitter their (ideal) channel measurements as unquantized modulation symbols, over the uplink channel. The analog feedback model applies also to the case of TDD, assuming reciprocity [26].

In general, the feedback link is affected by a delay of d slots. We model the noisy feedback link as

$$\mathbf{z}_k(t) = \mathbf{h}_k(t-d) + \boldsymbol{\nu}'_k(t)$$

where $\boldsymbol{\nu}'_k(t) \sim \mathcal{CN}(\mathbf{0}, N_0^{\mathrm{up}}\mathbf{I})$ is an i.i.d. process and $\{\boldsymbol{\nu}_k(t)\}$ and $\{\boldsymbol{\nu}'_k(t)\}$ are independent. The uplink SNR is equal to $1/N_0^{\mathrm{up}}$. This may depend on several factors such as the number of uplink pilot symbols used (for TDD) or the repetition/power allocation of the analog feedback symbols for FDD (see [26] for a very thorough discussion and comparison between TDD and FDD).

We define the CSIT $\boldsymbol{\alpha}_k(t)$ of user k in slot t as the MMSE *prediction* of $\mathbf{h}_k(t)$ given the observation $\{\mathbf{z}_k(t') : t' = 1, \ldots, t\}$, namely,

$$\boldsymbol{\alpha}_k(t) = \mathbb{E}[\mathbf{h}_k(t)|\mathbf{z}_k(t-d), \ldots, \mathbf{z}_k(0)] = r^d\mathbb{E}[\mathbf{h}_k(t-d)|\mathbf{z}_k(t-d), \ldots, \mathbf{z}_k(0)] \qquad (33)$$

where the last equality follows from the Gauss-Markov model (32).

This can be computed recursively using a Kalman filter/predictor approach. As a result, $\mathbf{h}_k(t) = \boldsymbol{\alpha}_k(t) + \mathbf{e}_k(t)$, $\mathbf{h}_k(t)$ and $\boldsymbol{\alpha}_k(t)$ are jointly Gaussian and the estimation error $\mathbf{e}_k(t)$ is statistically independent of $\boldsymbol{\alpha}_k(t)$, with i.i.d. components $\sim \mathcal{CN}(0, \sigma_e^2)$ where σ_e^2 is the noisy

d-step MMSE prediction error variance. For the Gauss-Markov process considered in (32), we have a closed-form expression for σ_e^2, given by

$$\sigma_e^2 = r^{2d}\epsilon_0 + (1-r^2)\sum_{l=0}^{d-1} r^{2l} \tag{34}$$

where ϵ_0 denotes the *estimation* MMSE for the delay-free observation, given by

$$\epsilon_0 = \left(\frac{1-r^2}{2r^2}\right)\left(-\left(1+\frac{1}{N_0^{\text{up}}}\right)+\sqrt{1+\left(\frac{1}{N_0^{\text{up}}}\right)^2+2\left(\frac{1}{N_0^{\text{up}}}\right)\frac{1+r^2}{1-r^2}}\right)$$

It follows that

$$\boldsymbol{\Sigma}_k(\boldsymbol{\alpha}(t)) = \boldsymbol{\alpha}_k(t)\boldsymbol{\alpha}_k^H(t) + \sigma_e^2\mathbf{I} \tag{35}$$

In the case of FDD, this form of analog feedback requires feeding back the components of $\boldsymbol{\alpha}_k(t)$ (M complex channel uses per user).

B Proof of Proposition 1

With some abuse of notation, in this section we denote by $\mathbf{G} = [\mathbf{g}_1^H, \ldots, \mathbf{g}_K^H]^H$ the MK-dimensional vector obtained by stacking the beamforming vectors. From (24), we can write

$$\frac{\partial f}{\partial \mathbf{g}_k} = \left[\frac{w_k}{d_k}\boldsymbol{\Sigma}_k - \mathbf{M}\right]\mathbf{g}_k$$

where

$$\mathbf{M} = \frac{\text{tr}(\mathbf{D})}{P}\mathbf{I} + \sum_j [\mathbf{D}]_{j,j}\boldsymbol{\Sigma}_j$$

We let

$$\mathbf{a}_k = \frac{w_k}{d_k}\mathbf{M}^{-1}\boldsymbol{\Sigma}_k\mathbf{g}_k$$

and $\mathbf{A} = [\mathbf{a}_1^H, \ldots, \mathbf{a}_K^H]^H$.

First, we show that the difference vector $\mathbf{A}^{(n)} - \mathbf{G}^{(n-1)}$, where $\mathbf{A}^{(n)}$ is obtained from $\mathbf{G}^{(n-1)}$ according to (28), has a non-negative projection onto the gradient $\nabla f(\mathbf{G}^{(n-1)})$, for all $\mathbf{G}^{(n-1)} \in \mathbb{C}^{MK}$. For any arbitrary $\mathbf{G}$, let

$$\begin{aligned}
\nabla f(\mathbf{G})^H(\mathbf{A}-\mathbf{G}) &= \sum_{k=1}^{K} \mathbf{g}_k^H \left[\frac{w_k}{d_k}\boldsymbol{\Sigma}_k - \mathbf{M}\right]\left[\frac{w_k}{d_k}\mathbf{M}^{-1}\boldsymbol{\Sigma}_k - \mathbf{I}\right]\mathbf{g}_k \\
&= \sum_{k=1}^{K} \mathbf{g}_k^H \left[\left(\frac{w_k}{d_k}\right)^2 \boldsymbol{\Sigma}_k\mathbf{M}^{-1}\boldsymbol{\Sigma}_k - 2\frac{w_k}{d_k}\boldsymbol{\Sigma}_k + \mathbf{M}\right]\mathbf{g}_k \qquad (36)
\end{aligned}$$

Since $\mathbf{M}$ is Hermitian symmetric positive definite for all $\mathbf{G}$ and finite P, then there exists a matrix $\mathbf{B}$ also symmetric positive definite such that $\mathbf{M} = \mathbf{B}\mathbf{B}^H$. Using this, we can write

$$\begin{aligned}
\left(\frac{w_k}{d_k}\right)^2 \boldsymbol{\Sigma}_k\mathbf{M}^{-1}\boldsymbol{\Sigma}_k - 2\frac{w_k}{d_k}\boldsymbol{\Sigma}_k + \mathbf{M} &= \left(\frac{w_k}{d_k}\right)^2 \boldsymbol{\Sigma}_k\mathbf{B}^{-H}\mathbf{B}^{-1}\boldsymbol{\Sigma}_k - 2\frac{w_k}{d_k}\boldsymbol{\Sigma}_k + \mathbf{B}\mathbf{B}^H \\
&= \left[\frac{w_k}{d_k}\mathbf{B}^{-1}\boldsymbol{\Sigma}_k - \mathbf{B}\right]^H \left[\frac{w_k}{d_k}\mathbf{B}^{-1}\boldsymbol{\Sigma}_k - \mathbf{B}\right] \qquad (37)
\end{aligned}$$

Hence, the quadratic form in (36) is Hermitian symmetric positive semidefinite and $\nabla f(\mathbf{G})^H(\mathbf{A}-\mathbf{G}) \geq 0$ for all $\mathbf{G}$, as we wanted to show.

It follows that by starting from any $\mathbf{G}$ and moving in the direction of $\mathbf{A}-\mathbf{G}$ we cut level contours of $f(\cdot)$ of non-decreasing level. The function of the real variable ν defined by

$$F(\nu) = f(\nu\mathbf{A} + (1-\nu)\mathbf{G})$$

is non-decreasing in $\nu = 0$. Hence, letting μ be the solution of $\max_{\nu>0} F(\nu)$ and replacing $\mathbf{G}$ with $\mu\mathbf{A} + (1-\mu)\mathbf{G}$ (as in (29) and (30)) yields a value of the objective function that is at least as large as $f(\mathbf{G})$. Furthermore, it is clear that the search can be restricted to $\nu \in [0,1]$ and we still obtain a non-decreasing objective function value.

Finally, the proof follows by repeating the same reasoning at every point $\mathbf{G}^{(n)}$ of the trajectory generated by the iterative algorithm and by the initial condition $\mathbf{G}^{(0)}$.

References

[1] E. M. Yeh and A. S. Cohen, "Information Theory, Queueing, and Resource Allocation in Multi-user Fading Communications," *in Proc. of the 2004 CISS, Princeton, NJ*, March 2004.

[2] H.Boche and M.Wiczanowski, "Stability Region of Arrival Rates and Optimal Scheduling for MIMO-MAC - A Cross-Layer Approach," *Proceeding of IZS,Zurich*, February 2004.

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

[4] T. Ren, R. J. La, and L. Tassiulas, "Optimal Transmission Scheduling with Base Station Antenna Array in Cellular Networks," *Proc. of IEEE Computer Communications Conference (Infocom)*, 2004.

[5] S. Shakkottai and A.Stolyar, "Scheduling for Multiple flows Sharing a Time-Varying Channel: The Exponential Rule," *American Mathematical Society Translations*, vol. 207, September 2002.

[6] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, vol. 38, pp. 70–77, July 2000.

[7] 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.

[8] M.Sharif and B.Hassibi, "On the Capacity of MIMO Broadcast Channel with Partial Side Information," *IEEE Trans. on Inform. Theory*, vol. 51, no. 2, pp. 506 – 522, February 2005.

[9] A.Sang, X.Wang, M.Madihian, and R.D.Gitlin, "Downlink scheduling schemes in cellular packet data systems of multiple-input multiple-output antennas," *Proceeding of IEEE Globecom 2004*, pp. 421–427, December 2004.

[10] S. Borst and P. Whiting, "The Use of Diversity Antennas in High-Speed Wireless Systems: Capacity Gains, Fairness Issues," *Bell Laboratories Technical Memorandum, 2001*, 2001.

[11] M. J. Neely, E. Modiano, and C. E. Rohrs, "Power Allocation and Routing in Multibeam Satellites With Time-Varying Channels," *IEEE/ACM Transaction on Networking*, vol. 11, pp. 138–152, February 2003.

[12] E. M. Yeh and A. S. Cohen, "Throughput optimal power and rate control for multiaccess and broadcast communications," in *IEEE Int. Symp. Inform. Theory, (ISIT 2004)*, Chicago, Illinois, June 27– July 2 2004, p. 112.

[13] "CDMA2000 High Rate Packet DataAir Interface Specification," *TIA/EIA/3GPP2 Standard IS-856/3GPP2 C.S.0024, v3.0*, December 2001.

[14] V. K. N. Lau, "Proportional Fair Space-Time Scheduling for Wireless Communications," *IEEE Trans. on Commun.*, vol. 53, no. 8, August 2005.

[15] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[16] L. Li and A. J. Goldsmith, "Capacity and Optimal resource Allocation for Fading Broadcast Channels-Part I:Ergodic Capacity," *IEEE Trans. on Inform. Theory*, vol. 47, pp. 1083–1102, March 2001.

[17] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multi-antenna gaussian broadcast channel," *IEEE Trans. on IT.*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[18] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the gaussian mimo broadcast channel," in *proc. of ISIT 2004*, Chicago, IL, June 27 – July 2, 2004, p. 174.

[19] S. Vishwanath, N. Jindal, and A. J. Goldsmith, “Duality, achievable rates, and sum-rate capacity of gaussian mimo broadcast channels,” *IEEE Trans. Inform. Theory*, vol. 49, pp. 2658–2668, Oct. 2003.

[20] P. Viswanath and D.N.C. Tse, “Sum capacity of the vector gaussian broadcast channel and uplink-downlink duality,” *IEEE Trans. Inform. Theory*, vol. 49, pp. 1912–1921, Aug. 2003.

[21] H. Viswanathan, S. Venkatesan, and H. Huang, “Downlink capacity evaluation of cellular networks with known interference cancellation,” *IEEE J. Selec. Areas in Commun*, vol. 21, no. 5, pp. 802–811, June 2003.

[22] M. Kobayashi and G. Caire, “An Iterative Water-Filling Algorithm for Maximum Weighted Sum-Rate of Gaussian MIMO-BC,” *IEEE J. Select. Areas Commun.*, vol. 24, August 2006.

[23] N.Jindal, “MIMO Broadcast Channels with Finite Rate Feedback,” *IEEE Global Telecommunications Conference (Globecom), Nov. 2005.*

[24] N. Jindal, “MIMO Broadcast Channels with Finite Rate Feedback,” *IEEE Trans. on Inform. Theory*, vol. 52, November 2006.

[25] P. Ding, D. J. Love, and M. D. Zoltowski, “On the Sum Rate of Multi-Antenna Broadcast Channels with Channel Estimation Error,” *Asilomar*, October 2005.

[26] 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.

[27] N. Mandayam and S. Verdú, “Analysis of an approximate decorrelating detector,” *Wireless Personal Communications*, vol. 6, no. 1–2, pp. 97–111, Jan. 1998.

[28] M. Kobayashi, “On the use of multiple antennas in a downlink of wireless systems,” *Ph.D dissertation, ENST*, June 2005.

[29] M. Kobayashi and G. Caire, “Joint Beamforming and Scheduling for a MIMO Downlink with Random Arrivals,” *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2006.

[30] M. Guillaud, D. Slock, and R. Knopp, “A Practical Method for Wireless Channel Reciprocity Exploitation Through Relative Calibration,” *ISSPA ’05, Sydney, Australia*, 2005.

[31] M. Kobayashi, G. Caire, and D. Gesbert, “Transmit diversity vs. opportunistic beamforming in data packet mobile downlink transmission,” *IEEE Trans. on Commun.*, January 2007.

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

[33] M. Bengtsson and B. Ottersten, “Optimal Downlink Beamforming using Semidefinite Optimization,” *Proc. of Annual Allerton Conference on Communication, Control, and Computing*, September 1999.

[34] H. Xu, J. Liu, F. Rubio, and A. I. Perez-Neira, “Orthogonal transmit beamforming scheme based on semidefinite optimization,” *European Trans. on Telecomm.*, vol. 17, no. 2, March/April 2006.

[35] M. Stojnic, H. Vikalo, and B. H. Hassibi, “Rate maximization in multi-antenna broadcast channels with linear preprocessing,” *Proceeding of IEEE Globecom ’2004*, 2004.

[36] G. Dimic and N. Sidiropoulos, “Low-complexity downlink beamforming for maximum sum capacity,” in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing*

(ICASSP 2004), Montreal, Quebec, Canada, May 17–21 2004, vol. 4, pp. iv–701 – iv–704.

[37] 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.

[38] T. Yoo and A. Goldsmith, “On the optimality of Multiantenna Broadcast Scheduling using Zero-Forcing Beamforming,” *IEEE J. Select. Areas Commun.*, vol. 24, no. 3, pp. 528–541, 2006.

[39] T. Yoo and A. Goldsmith, “Sum rate optimal multi-antenna downlink beamforming strategy based on clique search,” in *Proc. of IEEE Global Telecommunications Conference (Globecom)*.

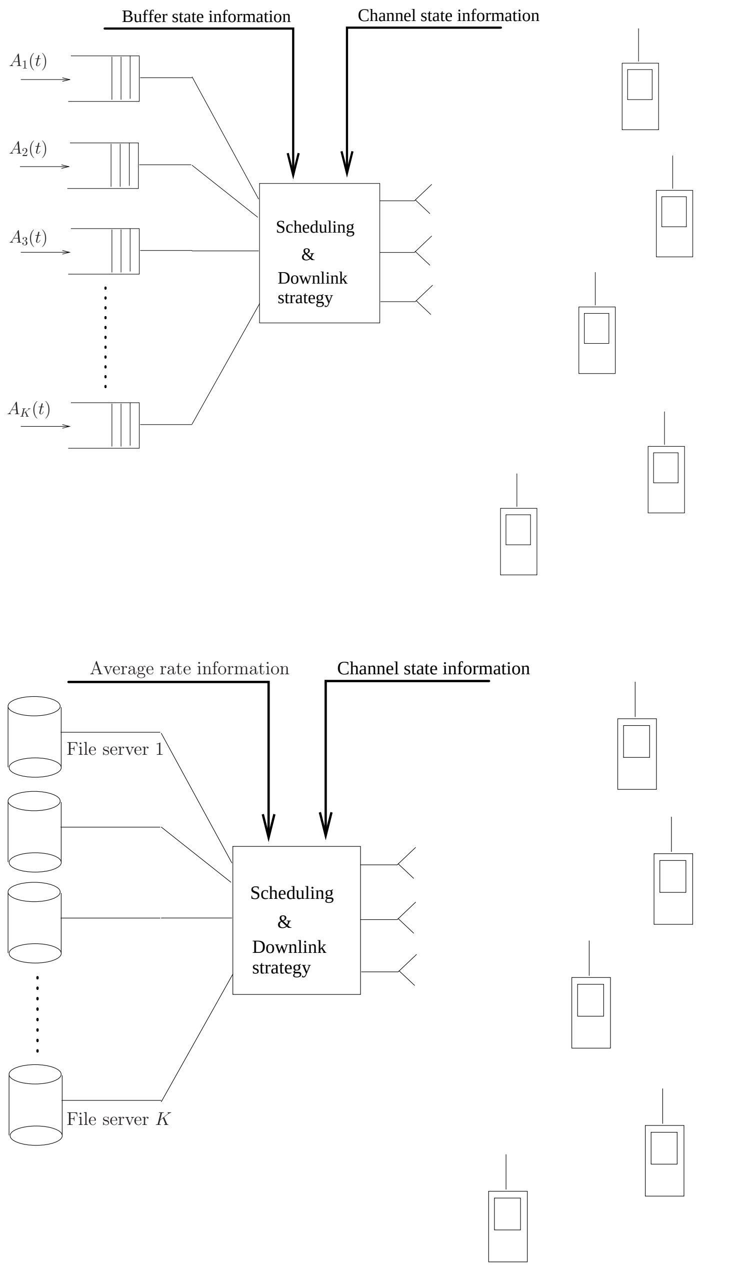
Buffer state information
Channel state information
$A_1(t)$
$A_2(t)$
$A_3(t)$
$A_K(t)$
Scheduling & Downlink strategy
Average rate information
Channel state information
File server 1
File server K
Scheduling & Downlink strategy

Figure 1: MIMO downlink with random arrivals (above) and with infinite backlog (below).

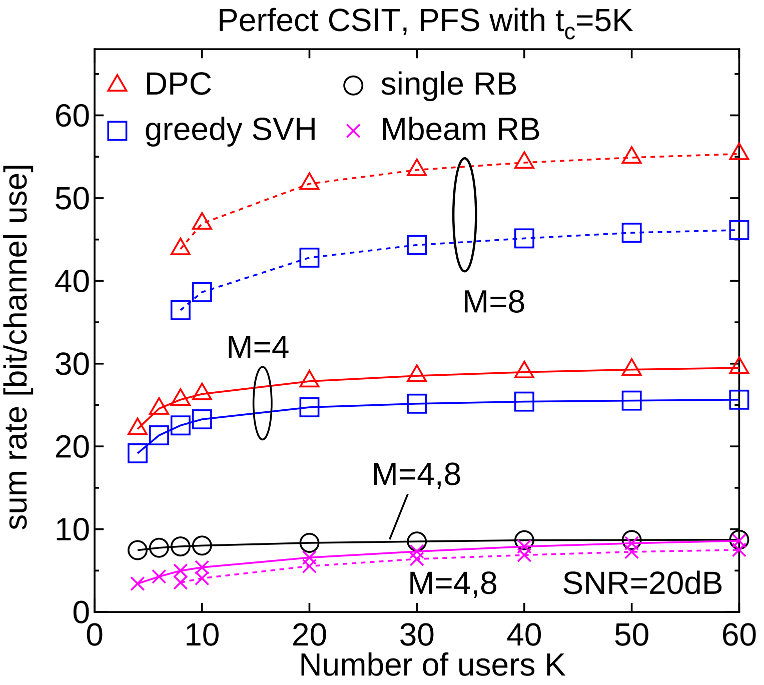

Figure 2: Sum rate vs. number of users with PFS and perfect CSIT.

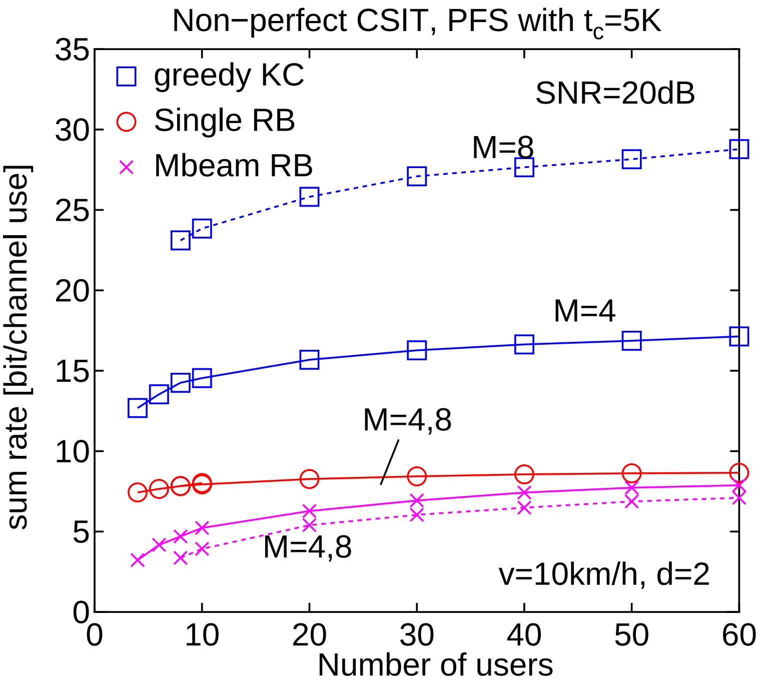

Figure 3: Sum rate vs. number of users with PFS and non-perfect CSIT.

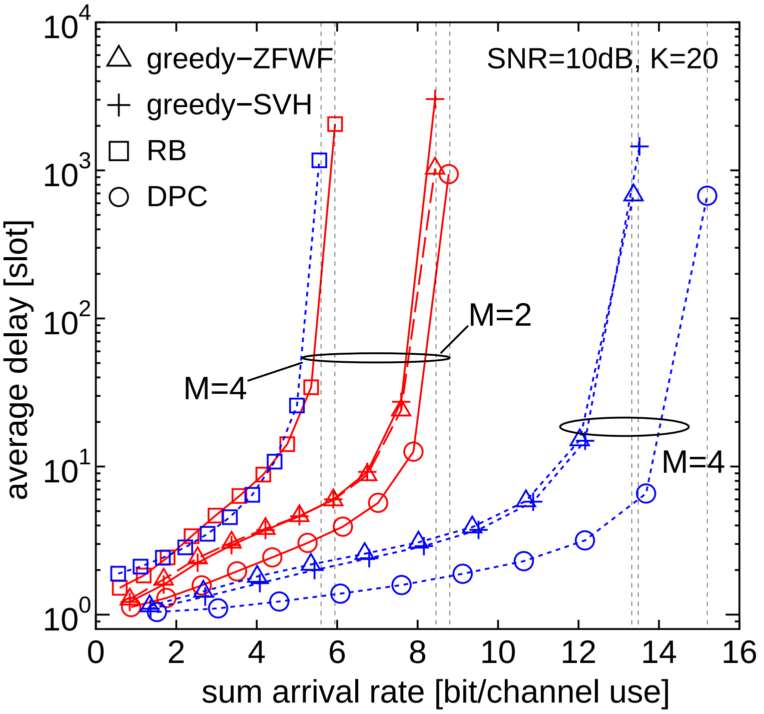

Figure 4: Average delay vs. sum arrival rate with perfect CSIT.

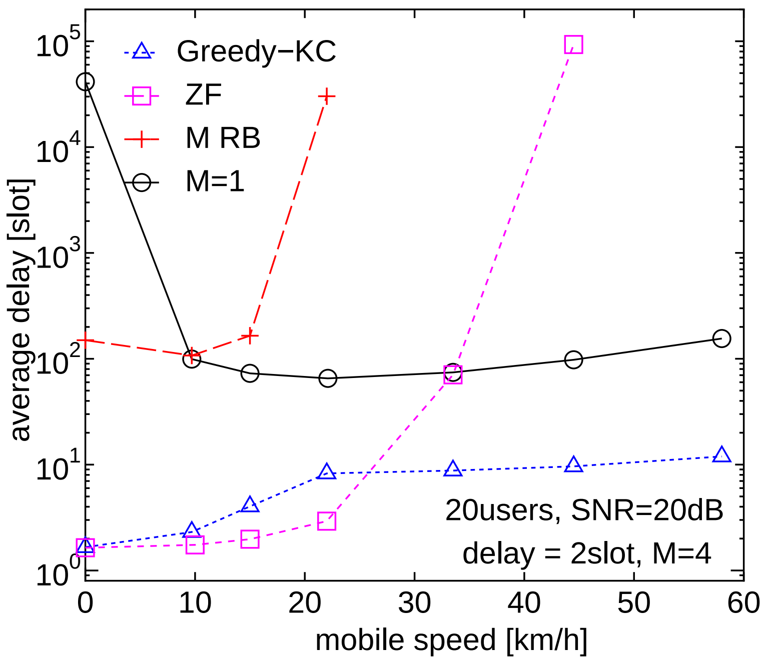

Figure 5: Average delay vs. mobile speed for greedy KC, greedy ZFWF and RB, $M = 4$ antennas.