RATE BALANCING IN THE BROADCAST CHANNEL WITH INDEPENDENT ENCODING

Pedro Tejera, Christoph K. Schmid, Wolfgang Utschick

Associate Institute for Signal Processing, Munich University of Technology
Arcisstraße 21, 80333 Munich
Email: {pedro.tejera,wolfgang.utschick}@tum.de, ckschmid@web.de
Phone: +49-89-289-28512, Fax: +49-89-289-28504

ABSTRACT

In the work at hand the rate balancing problem in the broadcast channel is stated and discussed under the restriction of independent encoding of transmitted information streams. For a short term power constraint, the solution to this problem is characterized and the number of time slots allocated by an optimum scheduler is found to be limited by the number of users. Furthermore, a practical algorithm is proposed that is able to achieve optimality at the cost of complexity. A suboptimum but efficient implementation of this algorithm is discussed and compared with more elemental scheduling and beamforming approaches. A major observation is the nearly optimality of zero-forcing (ZF) beamforming vectors for all kinds of rate balancing constraints and SNR values.

1. INTRODUCTION

The present work considers the rate balancing problem in a point to multipoint communication system. In this setting, a transmitter with M antennas communicates with K receivers, each equipped with a single antenna.

Using successive encoding at the transmitter, a rate balancing algorithm has been presented in [1] that is essentially the same as the algorithm proposed in [2] for power minimization with given rate constraints. Based on some of the results in [2] a rate balancing algorithm has been presented in [3] that applies to any number of receive antennas and comprises optimization of encoding order and link scheduling.

Due to practical reasons, independent encoding of information sent to the users in the network may be preferable. If Gaussian inputs are assumed and no link scheduling is considered, i.e., all users are bound to be served in the same time slot and using the same frequency band, the rate balancing problem is intimately related to the SINR balancing problem, which has been solved in [4]. Due to the fact that the achievable rate region without link scheduling (time sharing) is generally non-convex, serving all users simultaneously can be extremely suboptimum, especially (but not only) if the number of users in the network is larger than the number of transmit antennas. In this paper we incorporate time sharing into the formulation of the rate balancing problem. If the same power constraint applies in each of the scheduled time slots, we characterized the optimum solution and show that at most K time slots are needed to achieve the optimum solution. In order to solve this problem an algorithm is presented that trades off complexity for performance and is able to achieve the optimum if no complexity constraints are considered. The algorithm basically consists of a preselection of rate vectors from within the achievable rate region without time sharing and the subsequent solution of a linear optimization problem that determines the optimum time allocation for the preselected rate vectors. If no constraints are imposed regarding the number of preselected vectors the algorithm has the potential to achieve optimality. However, in order to keep complexity to reasonable levels an efficient preselection method is proposed whose asymptotical complexity depends polynomially on K.

Still, significant complexity regarding computation of optimum beamforming vectors as well as scheduling of users motivates consideration of simpler beamforming and grouping approaches. ZF beamforming vectors are considered as an alternative to the optimum minimum mean square error (MMSE) beamforming vectors. On the other hand, as an alternative to the polynomially complex preselection method mentioned above, a simpler successive grouping approach similar to that proposed in [5] is discussed. ZF forcing beamformers turn out to be nearly optimum for all scenarios investigated, independently of SNR values and rate balancing constraints. As we illustrate with the help of an example, this is due to the fact that, even though the ZF achievable rate region without time sharing and the MMSE achievable rate region without time sharing significantly differ, the convex hulls of both are nearly equal. Simpler scheduling approaches, as the one presented here and those in [5] and [6], turn out to perform good at low to moderate SNR values but are clearly suboptimum at high SNR values. This is intimately related to the restriction assumed by these approaches of serving each user in no more than one time slot.

The rest of the paper is structured as follows. Section 2 introduces the system model and states the rate balancing problem. In Section 3 the optimum solution to this problem is characterized and discussed. In Section 4 an optimum algorithm is presented and a suboptimum but efficient implementation of this algorithm is proposed. Section 5 discusses suboptimum beamforming and scheduling approaches that are usually found in literature. Finally, numerical results are shown in Section 6 and conclusions are drawn in Section 7.

1.1. Notation

In the following, vectors and matrices are denoted by lower case bold and capital bold letters, respectively. We use $(\bullet)^{\mathrm{T}}$ for matrix transposition and $(\bullet)^{\mathrm{H}}$ for conjugate transposition. $[\boldsymbol{A}]_{i,j}$ represents the entry in the ith row and jth column of matrix $\boldsymbol{A}$ and $[\boldsymbol{A}]_{\bullet,j}$ represents its jth column. $\mathrm{Co}\{\mathcal{R}\}$ denotes the convex hull of set $\mathcal{R}$ and $|\mathcal{R}|$ its cardinality.

2. SYSTEM MODEL AND PROBLEM STATEMENT

At a given time instant during the sth time slot, user k receives

$$y_{s,k} = \boldsymbol{h}_k^{\mathrm{H}} \sum_{j=1}^{K} \boldsymbol{p}_{s,j} x_j + n_k,$$

where $x_j \sim \mathcal{CN}(0,1)$ is the transmit signal sent to user j, $\boldsymbol{h}_k \in \mathbb{C}^{M\times 1}$ is the channel corresponding to user k, $\boldsymbol{p}_{s,j} \in \mathbb{C}^{M\times 1}$ represents the beamforming vector for user j at time slot s, and $n_k \sim \mathcal{CN}(0,1)$ represents a source of additive white Gaussian noise. Perfect channel knowledge is assumed at the transmitter and the receivers. Let $\mathcal{U} = \{1,\ldots,K\}$ denote the set of users in the network and $\mathcal{P}$ the power set of $\mathcal{U}$ minus the empty set. We define a scheduler as a pair $S = (\mathcal{S}, \boldsymbol{t})$, wherein $\mathcal{S} \subseteq \mathcal{P}$ and $\boldsymbol{t} \in \mathbb{R}_+^{|\mathcal{S}|\times 1}$, with $\|\boldsymbol{t}\|_1 = 1$, indicates the relative transmission time allocated to each element of $\mathcal{S} = \{\mathcal{G}_1,\ldots,\mathcal{G}_{|\mathcal{S}|}\}$.

Let $\boldsymbol{\rho} = [\ \rho_1 \ \ \cdots \ \ \rho_K \ ]^{\mathrm{T}}$ be the rate balancing constraint vector indicating the proportions that should be fulfilled by the rates achieved by the users in the network. The rate balancing problem with link scheduling can be stated as follows,

$$\max_{S,\boldsymbol{p}_{s,k}} \gamma \quad \text{s. t.} \quad r_k = \gamma \rho_k$$
$$\sum_{s=1}^{|\mathcal{S}|} t_s \sum_{k=1}^{K} \|\boldsymbol{p}_{s,k}\|_2^2 \leq P_{\mathrm{Tx}}, \qquad (1)$$

where $r_k = \sum_s t_s \log_2(1+\mathrm{SINR}_{s,k})$,

$$\mathrm{SINR}_{s,k} = \frac{\boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{p}_{s,k} \boldsymbol{p}_{s,k}^{\mathrm{H}} \boldsymbol{h}_k}{1 + \boldsymbol{h}_k^{\mathrm{H}} \sum_{j\neq k} \boldsymbol{p}_{s,j} \boldsymbol{p}_{s,j}^{\mathrm{H}} \boldsymbol{h}_k}, \qquad (2)$$

and P_{Tx} is the maximally allowed average transmit power. Note that consistency of the problem statement requires $\boldsymbol{p}_{s,k} = \boldsymbol{0}$ if $k \notin \mathcal{G}_s$, i.e., if user k is not served in time slot s, no power is allocated to that user in that slot. Analysis and solution of this problem are very complex tasks. A more tractable problem is obtained if (1) is replaced by

$$\sum_{k=1}^{K} \|\boldsymbol{p}_{s,k}\|_2^2 \leq P_{\mathrm{Tx}}, \ \forall s, \qquad (3)$$

i.e., in each time slot the power constraint must be fulfilled. As the noise have been assumed to have unit variance, for the rest of the paper we define $\mathrm{SNR} = P_{\mathrm{Tx}}$.

3. OPTIMUM SOLUTION

The solution of the rate balancing problem with constraint (3) can be characterized as the intersection of the straight line $\boldsymbol{r} = \gamma\boldsymbol{\rho}$ with the boundary of the convex hull of the achievable rate region without time sharing. The achievable region without time sharing is defined as

$$\mathcal{R} = \{\boldsymbol{r} \ : \ \exists \boldsymbol{\Gamma},\ r_k \leq \log_2(1+\mathrm{SINR}_k)\}, \qquad (4)$$

where $\boldsymbol{\Gamma} = [\ \mathrm{SINR}_1 \ \ \cdots \ \ \mathrm{SINR}_K \ ]^{\mathrm{T}}$ is supposed to be a feasible vector of SINR values [4]. This characterization is a direct consequence of the following result.

Theorem 1 *Let S and $\{\boldsymbol{p}_{s,k}\}$ be the optimum scheduler and beamformers for the rate balancing problem with constraint (3). In each time slot i, the corresponding beamformers $\{\boldsymbol{p}_{i,k}\}$ achieve a rate vector $\boldsymbol{r}_i$ on the boundary of the convex hull of $\mathcal{R}$.*

Proof: Let $\boldsymbol{r}_i$ be the rate vector achieved by the optimum solution in time slot i. Assume that this vector lies in the interior of $\mathrm{Co}\{\mathcal{R}\}$. This means that there exists a vector $\boldsymbol{r}'_i = \alpha\boldsymbol{r}_i$ with $\alpha > 1$ that belongs to $\mathrm{Co}\{\mathcal{R}\}$. Since $\boldsymbol{r}'_i \in \mathrm{Co}\{\mathcal{R}\}$, a number of points $\boldsymbol{r}_{ij} \in \mathcal{R}$ can be found such that $\boldsymbol{r}'_i = \sum_j \mu_j \boldsymbol{r}_{ij}$ with $\sum_j \mu_j = 1$. Let $j \in \{1,\ldots,J\}$, t_i be the fraction of time allocated to slot i and $\boldsymbol{r} = \sum_{s=1}^{|\mathcal{S}|} t_s \boldsymbol{r}_s$ the rate vector achieved by the optimum solution. The amplitude of this vector can be increased as follows. First, divide time slot i into two time slots. A time slot a of length $t_i^a = t_i/\alpha$ and another time slot b of length $t_i^b = (\alpha-1)t_i/\alpha$. In time slot a, $\boldsymbol{r}'_i$ can be achieved by dividing this slot into J time slots and assigning to each subslot j a fraction μ_j of the time t_i^a. Doing that, $\boldsymbol{r}$ is achieved without using time slot b. Indeed, we can write

$$\boldsymbol{r} = \sum_{\substack{s=1\\ s\neq i}}^{|\mathcal{S}|} t_s \boldsymbol{r}_s + t_i^a \sum_{j=1}^{J} \mu_j \boldsymbol{r}_{ij}.$$

Applying the same scheduling policy within slot b as during the time $(1-t_i^b)$ a rate vector $\boldsymbol{r}' = \boldsymbol{r} + t_i^b \boldsymbol{r}$ results that has larger amplitude than vector $\boldsymbol{r}$ and its same direction. This contradicts the optimality of the initial solution. *q.e.d.*

Interestingly, for the rate balancing problem with constraint (3) an optimum scheduler can always be found that allocates the users in at most $|\mathcal{S}| \leq K$ time slots. This is stated by the following theorem.

Theorem 2 *If $S = (\mathcal{S}, \boldsymbol{t})$ is an optimum scheduler with $|\mathcal{S}| > K$ achieving the longest possible rate vector $\boldsymbol{r} = \gamma^* \boldsymbol{\rho}$, there exists a scheduler $S' = (\mathcal{S}', \boldsymbol{t}')$ such that $\mathcal{S}' \subset \mathcal{S}$ and $|\mathcal{S}'| \leq K$ that also achieves $\boldsymbol{r}$.*

Proof: Let $S = (\mathcal{S}, \boldsymbol{t})$ be an optimum scheduler with $|\mathcal{S}| > K$ and $t_s > 0\, \forall s$. In addition, let $\mathcal{Q} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_{|\mathcal{S}|}\}$ be a set formed by the rate vectors achieved in the different time slots. It is well known that the convex hull of a finite set of points in $\mathbb{R}^n$ is a convex polytope limited by facets of dimension $n-1$. In particular, $\mathrm{Co}\{\mathcal{Q}\}$ is a convex polytope in $\mathbb{R}^K$ with facets of dimension $K-1$. Achievability of $\boldsymbol{r} = \gamma^* \boldsymbol{\rho}$ implies $\boldsymbol{r} \in \mathrm{Co}\{\mathcal{Q}\}$. In addition, optimality requires that $\boldsymbol{r}$ be in the boundary of $\mathrm{Co}\{\mathcal{Q}\}$, as otherwise longer rate vectors could be found within $\mathrm{Co}\{\mathcal{Q}\}$ on the straight line defined by $\boldsymbol{\rho}$. Due to convexity, there exists a hyperplane supporting $\mathrm{Co}\{\mathcal{Q}\}$ such that $\boldsymbol{r}$ is in the hyperplane. Accordingly, if the hyperplane equation is given by $\boldsymbol{v}^{\mathrm{T}} \boldsymbol{x} = d$, we have $\boldsymbol{v}^{\mathrm{T}} \boldsymbol{r} = d$ and, $\forall \boldsymbol{q} \in \mathrm{Co}\{\mathcal{Q}\}$, $\boldsymbol{v}^{\mathrm{T}} \boldsymbol{q} \leq d$. In particular, $\boldsymbol{v}^{\mathrm{T}} \boldsymbol{r}_s \leq d$, $\forall s$, and, therefore,

$$\boldsymbol{v}^{\mathrm{T}} \boldsymbol{r} = d \geq \sum_{s=1}^{|\mathcal{S}|} t_s \boldsymbol{v}^{\mathrm{T}} \boldsymbol{r}_s = \boldsymbol{v}^{\mathrm{T}} \boldsymbol{r}$$

holds. Here, equality is only achieved if $\boldsymbol{v}^{\mathrm{T}} \boldsymbol{r}_s = d$, $\forall s$, i.e., if all rate vectors lie on the supporting hyperplane.

So far, it has been shown that the rate vectors achieved in each time slot of the optimum scheduler all lie in the same hyperplane of dimension $K-1$. Now, the result follows from direct application of the Carathéodory's theorem on convex sets. This theorem states that for any point $\boldsymbol{r}$ lying in the convex hull of a set of points $\mathcal{Q}$ in a space of dimension $K-1$, there exists a subset $\mathcal{Q}' \subseteq \mathcal{Q}$ such that $|\mathcal{Q}'| \leq K$ whose convex hull also contains $\boldsymbol{r}$. *q.e.d.*

Building on the insights provided by these results, an optimum algorithm would proceed searching in the intersection of the boundary of $\mathcal{R}$ with its convex hull for a set of at most K points such that the intersection of the convex hull of this set with $\gamma \boldsymbol{\rho}$ is as far from the origin as possible. Unfortunately, since little is known about the structure of $\mathcal{R}$, such a procedure is only of theoretical interest.

4. PRACTICAL APPROACH

A practical approach to the rate balancing problem consists of choosing an appropriate set of points on the boundary of $\mathcal{R}$ and optimizing the scheduling strategy focussing on this set. This can be done by solving the following optimization problem

$$\max_{\boldsymbol{t}} \gamma \quad \text{s. t.} \quad \tilde{\boldsymbol{R}} \boldsymbol{t} = \gamma \boldsymbol{\rho},\ \|\boldsymbol{t}\|_1 = 1,\ \boldsymbol{t} \in \mathbb{R}_+^L, \qquad (5)$$

where $\tilde{\boldsymbol{R}} = [\ \boldsymbol{r}_1 \ \cdots \ \boldsymbol{r}_L\ ] \in \mathbb{R}_+^{K \times L}$ is the matrix of rate vectors selected from $\mathcal{R}$. Defining $\tilde{\boldsymbol{t}} = \boldsymbol{t}/\gamma$, (5) can be restated as a standard linear optimization problem as follows,

$$\min_{\tilde{\boldsymbol{t}}} \|\tilde{\boldsymbol{t}}\|_1 \quad \text{s. t.} \quad \tilde{\boldsymbol{R}} \tilde{\boldsymbol{t}} = \boldsymbol{\rho},\ \tilde{\boldsymbol{t}} \in \mathbb{R}_+^L.$$

Sampling the boundary of $\mathcal{R}$ very finely with a corresponding high number L of rate vectors allows to achieve practically any point in the convex hull of $\mathcal{R}$ by means of time sharing, and, hence, it assures almost optimum performance at the cost of high complexity. However, in order to keep complexity at reasonable levels, it makes sense to limit the selection to a few rate vectors that hopefully can deliver a performance close to the optimum.

A good compromise between complexity and performance is achieved by doing the selection of rate vectors as follows. First, only subsets $\mathcal{G}_i \subset \mathcal{U}$ are considered such that $|\mathcal{G}_i| \leq M$, i.e., no rate vectors are considered resulting from serving more than M users in one time slot as these vectors are usually in the interior of the convex hull of $\mathcal{R}$. From all subsets with cardinality less or equal to M, two rate vectors are computed. A rate vector that satisfies the rate constraints imposed by $\boldsymbol{\rho}$ for the users in the subset. A rate vector obtained by uniformly allocating power in the dual uplink [4] over the users of the subset. The first vector is a natural choice that provides the optimum solution in cases where $K \leq M$ and no time sharing is needed to achieve the optimum. The second vector is likely to be close to a sum rate maximizing point for a given subset as observed in [7]. The choice of sum rate optimum vectors is purposeful due to Theorem 1 and the following result.

Theorem 3 *Let $\mathcal{G} \subseteq \mathcal{U}$ be any group of users and let $\boldsymbol{r} \in \mathbb{R}^{K \times 1}$ be the sum rate maximizing vector for this group. Vector $\boldsymbol{r}$ lies in the boundary of $\mathrm{Co}\{\mathcal{R}\}$.*

Proof: Assume that $\boldsymbol{r}$ lies in the interior of $\mathrm{Co}\{\mathcal{R}\}$. This means that there exists $\alpha > 1$ such that $\alpha \boldsymbol{r} \in \mathrm{Co}\{\mathcal{R}\}$. As a result a number I of vectors $\boldsymbol{r}_i \in \mathcal{R}$, $1 \leq i \leq I$, can be found such that $\alpha \boldsymbol{r} = \sum_i \mu_i \boldsymbol{r}_i$ with $\sum_i \mu_i = 1$. Let $\boldsymbol{e} \in \{0,1\}^{K \times 1}$ be a vector with entries $e_k = 1$ if $k \in \mathcal{G}$, $e_k = 0$ if $k \notin \mathcal{G}$. Choose

$$j = \arg\max_{1 \leq i \leq I} \{\boldsymbol{e}^{\mathrm{T}} \boldsymbol{r}_i\}.$$

Putting all together the following relations hold

$$\boldsymbol{e}^{\mathrm{T}} \boldsymbol{r} \leq \alpha \boldsymbol{e}^{\mathrm{T}} \boldsymbol{r} = \sum_i \mu_i \boldsymbol{e}^{\mathrm{T}} \boldsymbol{r}_i \leq \boldsymbol{e}^{\mathrm{T}} \boldsymbol{r}_j.$$

This shows that there exists a vector $\boldsymbol{r}_j \in \mathcal{R}$ achieving a sum rate larger than $\boldsymbol{r}$ as far as the users in $\mathcal{G}$ are concerned. This clearly contradicts the initial assumption of $\boldsymbol{r}$ being a sum rate maximizing vector for the elements of $\mathcal{G}$. *q.e.d.*

Due to the restriction $|\mathcal{G}_i| \leq M$, the number of sets that must be considered is given by

$$\sum_{m=1}^{M} \binom{K}{m} \approx \mathcal{O}(K^M),$$

i.e., the complexity is polynomial in the number of users. This is in contrast with the exponential complexity that would result if no restriction were imposed on the number of users served in one time slot.

This approach is illustrated in Fig. 1 for $M = 2$ transmit antennas, $K = 3$ users and $\mathrm{SNR} = 20$ dB. All subsets of $\mathcal{U} = \{1, 2, 3\}$ with 1 and 2 elements are considered. For each set, the constraint fulfilling vector and the uniform power allocation vector have been computed and plotted in the figure. Note that in case of subsets with only one user the boundary of $\mathcal{R}$ (represented by dotted lines) is achieved by a unique vector. The constraint is represented by a straight line departing from the origin. The optimum scheduler assigns a portion of time to each of the sets with two users, in particular, to the points at the vertices of the dashed triangle. For two of the sets the uniform power allocation points are chosen, for the other set, with users 1 and 3, the constraint fulfilling vector is chosen. The point represented by the square is achieved by switching between these three points.

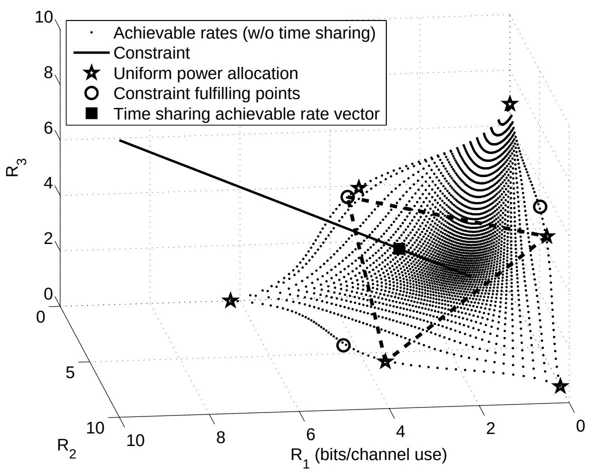

Fig. 1. Achievable rate region without time sharing and achievable rate vector for $\rho = [1, 1, 1]$.

Computation of rate vectors obtained with uniform power loading in the dual uplink involves computing the optimum MMSE beamforming vectors in the uplink. The corresponding beamforming vectors in the downlink have the direction of the MMSE beamforming vectors in the dual uplink and their norm is obtained by computing the principal eigenvector of the extended downlink coupling matrix [4].

Computation of the constraint fulfilling rate vectors can be performed as follows. Without loss of generality consider a set $\mathcal{G}$ including the first N users of $\mathcal{U}$. Let $\boldsymbol{\rho}_{\mathcal{G}} = [\ \rho_1 \ \cdots \ \rho_N\ ]^{\mathrm{T}}$ be the constraint vector corresponding to the users in this set. First, initialize $r_n^{(0)} = \rho_n$ and $\beta_n^{(1)} = 2^{r_n^{(0)}} - 1,\ \forall n$, and solve

$$\max_{\{\boldsymbol{p}_n\}_{n=1,\ldots,N}} \eta \quad \text{s. t.} \quad \mathrm{SINR}_n^{(1)} = \eta \beta_n^{(1)}\ \forall n, \qquad \sum_{n=1}^{N} \|\boldsymbol{p}_n\|_2^2 \leq P_{\mathrm{Tx}} \tag{6}$$

by using the SINR balancing algorithm in [4]. Here, $\mathrm{SINR}_n^{(1)}$ is given by (2) with appropriate indexing. Next, compute $\xi_n = \log_2(1 + \mathrm{SINR}_n^{(1)}),\ \forall n$, and

$$\gamma^{(1)} = \min_n \left\{ \frac{\xi_n}{r_n^{(0)}} \right\}. \tag{7}$$

Define $r_n^{(1)} = \gamma^{(1)} r_n^{(0)}$ and $\beta_n^{(2)} = 2^{r_n^{(1)}} - 1,\ \forall n$. We observe that $\beta_n^{(2)} \leq \mathrm{SINR}_n^{(1)},\ \forall n$, where equality holds for the index that minimizes (7). That is, $\boldsymbol{\beta}^{(2)} = [\ \beta_1^{(2)} \ \cdots \ \beta_N^{(2)}\ ]^{\mathrm{T}}$ is a vector of feasible SINR values. As a result, solving (6) with $\boldsymbol{\beta}^{(2)}$ as a constraint yields $\eta \geq 1$. That means that the SINR of every user can be improved or maintained with respect to $\boldsymbol{\beta}^{(2)}$ and, hence, also the rate of every user can be improved or maintained with respect to $\boldsymbol{r}^{(1)} = [\ r_1^{(1)} \ \cdots \ r_N^{(1)}\ ]^{\mathrm{T}}$. In other words, if $\xi_n = \log_2(1 + \mathrm{SINR}_n^{(2)})$,

$$\gamma^{(2)} = \min_n \left\{ \frac{\xi_n}{r_n^{(1)}} \right\} \geq 1.$$

Iterating this procedure a sequence of feasible rate vectors $\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(2)}, \boldsymbol{r}^{(3)}, \ldots$ is obtained, all of them being colinear with $\boldsymbol{\rho}_{\mathcal{G}}$ and having increasing norms, i.e., $\|\boldsymbol{r}^{(1)}\| \leq \|\boldsymbol{r}^{(2)}\| \leq \|\boldsymbol{r}^{(3)}\| \leq \cdots$. Since, due to the power constraint, this sequence is bounded, the algorithm converges. This algorithm is summarized in Table 1.

5. SUBOPTIMUM APPROACHES

5.1. Suboptimum Beamforming

Preselected rate vectors on the boundary of $\mathcal{R}$ require optimum beamforming vectors, i.e., MMSE vectors. As we have just seen in the previous section, computation of the optimum beamformers for a given constraint $\boldsymbol{\rho}_{\mathcal{G}}$ involves a doubly iterative algorithm. Indeed, problem (6) must be repeteadly solved and its solution is computed by the iterative SINR balancing algorithm in [4]. In order to reduce computational

`initialization :` $\ell = 1,\ r_n^{(0)} = \rho_n,\ \beta_n^{(1)} = 2^{r_n^{(0)}} - 1,\ \forall n$
`repeat :`
1. $\max_{\{\boldsymbol{p}_n\}_{n=1,\dots,N}} \eta$ s. t. $\mathrm{SINR}_n^{(\ell)} = \eta\beta_n^{(\ell)}\ \forall n$, $\sum_{n=1}^{N} \|\boldsymbol{p}_n\|_2^2 \leq P_{\mathrm{Tx}}$
2. $\gamma^{(\ell)} = \min_n \left\{ \frac{\xi_n}{r_n^{(\ell-1)}} \right\}, \quad \xi_n = \log_2(1 + \mathrm{SINR}_n^{(\ell)})$
3. $r_n^{(\ell)} = \gamma^{(\ell)} r_n^{(\ell-1)},\ \beta_n^{(\ell+1)} = 2^{r_n^{(\ell)}} - 1,\ \forall n,\ \ \ell = \ell + 1$
`until` $1 - \epsilon < \gamma^{(\ell)} < 1 + \epsilon$

Table 1. Rate balancing algorithm

complexity, instead of optimum MMSE beamformers, ZF beamforming vectors can be employed. Doing so, the preselected rate vectors will generally not lie on the boundary of $\mathcal{R}$. However, as we shall see, some of the preselected rate vectors will always be good enough to achieve a solution almost as good as that obtained with MMSE beamforming vectors.

Without loss of generality, let again $\mathcal{G}$ be a set including the first N users of $\mathcal{U}$ and $N \leq M$. The unit norm ZF beamforming vector for user n is given by

$$\boldsymbol{v}_n = \frac{\left[\boldsymbol{H}_{\mathcal{G}}^{\mathrm{H}} \left(\boldsymbol{H}_{\mathcal{G}} \boldsymbol{H}_{\mathcal{G}}^{\mathrm{H}}\right)^{-1}\right]_{\bullet,n}}{\sqrt{\left[\left(\boldsymbol{H}_{\mathcal{G}} \boldsymbol{H}_{\mathcal{G}}^{\mathrm{H}}\right)^{-1}\right]_{n,n}}}.$$

where $\boldsymbol{H}_{\mathcal{G}} \in \mathbb{C}^{N\times M}$ is a matrix formed by the channel vectors of the users in group $\mathcal{G}$. Specifically, the nth row of $\boldsymbol{H}_{\mathcal{G}}$ is vector $\boldsymbol{h}_n^{\mathrm{H}}$. The channel gains of the scalar channels resulting from applying unit-norm ZF beamformers at the transmitter are given by

$$g_n = \left(\left[\left(\boldsymbol{H}_{\mathcal{G}} \boldsymbol{H}_{\mathcal{G}}^{\mathrm{H}}\right)^{-1} \right]_{n,n} \right)^{-1}.$$

Having computed the unit-norm ZF beamforming vectors, sum rate maximizing rate vectors can be obtained performing a waterfilling power allocation over the resulting decoupled scalar channels. In case a rate constraint $\boldsymbol{\rho}_{\mathcal{G}}$ has to be fulfilled the optimum power allocation can be computed as follows. The constraint $r_n = \gamma\rho_n,\ \forall n$, implies $p_n = (2^{\gamma\rho_n} - 1)/g_n,\ \forall n$. Due to the monotonicity of $\sum_n p_n$ as a function of γ, a bisection search can be applied in order to look for the value γ for which $\sum_n p_n = P_{\mathrm{Tx}}$. Starting with a certain value for γ, if the total power exceeds the power constraint, γ is decreased. Otherwise, γ is increased.

5.2. Suboptimum Group Selection

Even though the restriction $|\mathcal{G}_i| \leq M$ deacreases complexity from exponential to polynomial order, further complexity reduction may be convenient if, for instance, the number of users in the network is very high. Some work on low complexity grouping algorithms have been done in recent years, e.g., [5] [6]. Roughly speaking, algorithms of this kind group users together with nearly orthogonal channels. Here, the following simple grouping algorithm is considered, which is similar to those presented in [5] and [6].

Let $\mathcal{W}^\ell$ be the set of users that have still not been assigned to any group at step ℓ in the execution of the algorithm. Initially, $\mathcal{W}^0 = \mathcal{U}$. Aiming at the construction of the first set $\mathcal{G}_1$ we define $\boldsymbol{T}_{\mathcal{G}_1}^{(0)} = \boldsymbol{I}_M$ and $C_{\mathcal{G}_1}^{(0)} = 0$. Construction starts by choosing the user with the largest channel norm, i.e.,

$$k_0 = \arg\max_k \{\boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{T}_{\mathcal{G}_1}^{(0)} \boldsymbol{h}_k\}, \forall k \in \mathcal{W}^0. \tag{8}$$

Then, the following updates are made, $\mathcal{W}^1 = \mathcal{W}^0 \setminus \{k_0\}$,

$$\boldsymbol{T}_{\mathcal{G}_1}^{(1)} = \boldsymbol{T}_{\mathcal{G}_1}^{(0)} - \frac{\boldsymbol{T}_{\mathcal{G}_1}^{(0)} \boldsymbol{h}_{k_0} \boldsymbol{h}_{k_0}^{\mathrm{H}} \boldsymbol{T}_{\mathcal{G}_1}^{(0)}}{\boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{T}_{\mathcal{G}_1}^{(0)} \boldsymbol{h}_k}$$

and $C_{\mathcal{G}_1}^{(1)}$ is set equal to the sum rate obtained by the users in group $\mathcal{G}_1$ using ZF beamforming vectors and waterfilling power allocation. At this stage $C_{\mathcal{G}_1}^{(1)} = \log_2(1 + \|\boldsymbol{h}_{k_0}\|^2 P_{\mathrm{Tx}})$. Selection of the second user is made according to (8) using $\boldsymbol{T}_{\mathcal{G}_1}^{(1)}$ instead of $\boldsymbol{T}_{\mathcal{G}_1}^{(0)}$ and considering only users in $\mathcal{W}^1$. Thereafter, $\boldsymbol{T}^{(2)}$ and $C_{\mathcal{G}_1}^{(2)}$ are computed as explained above. Let k_1 be the user selected at this stage. In order to decide whether this second user is included in the first group, $C_{\mathcal{G}_1}^{(2)}$ is compared with $C_{\mathcal{G}_1}^{(1)}$. If the sum rate obtained with both users is greater than that obtained with only the first user the second user is included in $\mathcal{G}_1$ and after making the update $\mathcal{W}^2 = \mathcal{W}^1 \setminus \{k_1\}$ the same steps are repeated to look for a third canditate. Otherwise, the second user is excluded from group $\mathcal{G}_1$ and this set is considered to be complete. In this case, construction of the second group, $\mathcal{G}_2$, starts in the next step by initializing $\boldsymbol{T}_{\mathcal{G}_2}^{(0)} = \boldsymbol{I}_M$ and $C_{\mathcal{G}_2}^{(0)} = 0$ and following the same steps as in the construction of group $\mathcal{G}_1$. The algorithm finishes once $\mathcal{W}^\ell = \emptyset$. A general outline of this algorithm is given in Table 2.

This is essentially the same algorithm described in [5] but using sum capacity as criterion for the inclusion of an additional user in a group rather than a more or less arbitrary orthogonality factor. The procedure has also some similarities to the tree search proposed in [6]. Here, the metric function is sum rate. However, the search, which starts from the bottom, does not necessarily covers all the tree and does not follow an order determined beforehand. Instead it is carried out dynamically and finishes as soon as all users have been assigned to groups.

6. NUMERICAL RESULTS

Figs. 2 and 3 show the average value γ obtained from application of approaches discussed in previous sections to a `initialization` : $\ell = 0, \quad g = 1, \quad \mathcal{W}^0 = \mathcal{U}, \quad \mathcal{G}_1 = \emptyset$

$r_n^{(0)} = \rho_n, \quad \boldsymbol{T}_{\mathcal{G}_1}^{(0)} = \boldsymbol{I}_M, \quad C_{\mathcal{G}_1}^{(0)} = 0$

`repeat` :

1. $k_0 = \arg\max_k \{\boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{T}_{\mathcal{G}_g}^{(\ell)} \boldsymbol{h}_k\}, \ \forall k \in \mathcal{W}^\ell, \ \mathcal{G}_g = \mathcal{G}_g \cup \{k_0\}$
2. `compute` $C_{\mathcal{G}_g}^{(\ell+1)}$
3. `if` $C_{\mathcal{G}_g}^{(\ell+1)} > C_{\mathcal{G}_g}^{(\ell)}$
   1. $\boldsymbol{T}_{\mathcal{G}_g}^{(\ell+1)} = \boldsymbol{T}_{\mathcal{G}_g}^{(\ell)} - \dfrac{\boldsymbol{T}_{\mathcal{G}_g}^{(\ell)} \boldsymbol{h}_{k_0} \boldsymbol{h}_{k_0}^{\mathrm{H}} \boldsymbol{T}_{\mathcal{G}_g}^{(\ell)}}{\boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{T}_{\mathcal{G}_g}^{(\ell)} \boldsymbol{h}_k}$
   2. $\mathcal{W}^{\ell+1} = \mathcal{W}^\ell \setminus \{k_0\}, \ \ell = \ell + 1$

   `else`
   1. $\mathcal{G}_g = \mathcal{G}_g \setminus \{k_0\}, \ g = g + 1$
   2. $\mathcal{G}_g = \emptyset, \ \boldsymbol{T}_{\mathcal{G}_g}^{(\ell)} = \boldsymbol{I}_M, \ C_{\mathcal{G}_g}^{(\ell)} = 0$

`until` $\mathcal{W}^\ell = \emptyset$

Table 2. Suboptimum grouping algorithm

scenario with $M = 2$ transmit antennas and $K = 5$ users. Note that, in both plots, the value of γ coincides with the rate of user $k = 1$ as $\rho_1 = 1$. Averaging has been done over a number of channel realizations whose coefficients have been independently drawn according to a complex Gaussian distribution of unit variance. The label "combinatorial grouping" designates the approach discussed in Section 4, which considers all possible groups that can be built with no more users than antennas. The label "suboptimum grouping" refers to the grouping approach discussed in Section 5.2.

It can be observed that suboptimum ZF beams perform nearly as well as optimum MMSE beams with both grouping approaches. The largest gap can be noticed between combinatorial grouping and suboptimum grouping at moderate to high SNR values. The gap is, at least in part, due to the fact that the suboptimum grouping approach is unable to allocate a certain user in more than one time slot. In this scenario, this means that even at high SNR there exists a time slot in which only one user is served. In turn, this means that, at least in one time slot, the potential multiplexing gain remains unexploited. Note that this is a drawback of the suboptimum algorithm proposed here as well as the simple grouping algorithms proposed in [5] and [6].

Deeper insights can be gained by looking at scenarios with two users. A particular example of such a scenario with $M = 2$ antennas is shown in Figs. 4, 5 and 9. The points achieved by each of the four approaches are plotted for all constraints vectors of the form $\boldsymbol{\rho} = [1, \alpha]^{\mathrm{T}}$ with $\alpha = n/10$ or $\alpha = 10/n$ and $n = 0, 1, \ldots, 10$. In addition to the curves of the approaches previously discussed, we have also included the boundary of the achievable region without time sharing and that of the region that can be achieved by use of ZF beamforming vectors. The first region is the set $\mathcal{R}$ defined in (4).

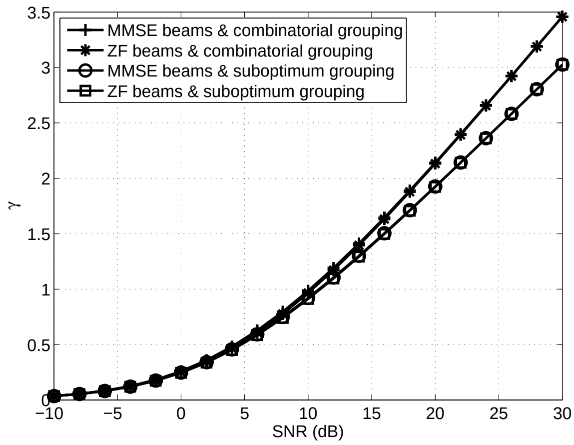

Fig. 2. Average rate per channel use achieved by any of the users in a system with $M = 2$, $K = 5$ and $\rho = [1, 1, 1, 1, 1]$.

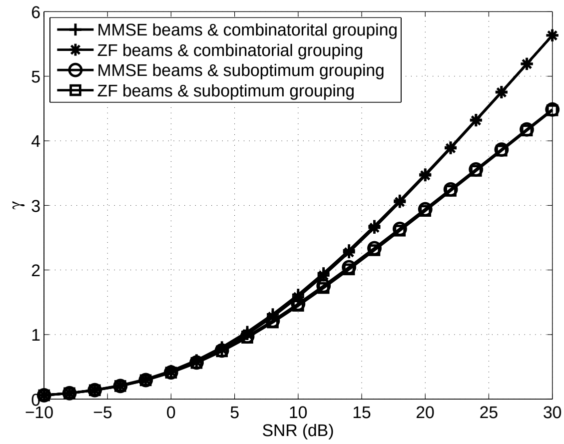

Fig. 3. Average rate per channel use achieved by the first user in a system with $M = 2$, $K = 5$ and $\rho = [1, 0.8, 0.6, 0.4, 0.2]$.

The ZF region is formed by the surface under the dotted curve in each figure and the intervals comprised between 0 and the respective single user capacity in both axes. Although the ZF region is certainly much smaller in the three figures, it can be easily seen that the convex hull of the ZF region is in all cases, at least, almost as large as $\mathrm{Co}\{\mathcal{R}\}$. In particular, for $\mathrm{SNR} = 0$ dB, the convex hull of the ZF region is equal to $\mathrm{Co}\{\mathcal{R}\}$. Almost equality can also be claimed for $\mathrm{SNR} = 20$ dB. The most significant difference can be appreciated for $\mathrm{SNR} = 10$ dB. However, this difference hardly represents a loss of more than one tenth of one bit for each user. Even if this is a particular example, it seems to be generally true that the convex hull of the ZF region is almost as large as $\mathrm{Co}\{\mathcal{R}\}$. At least, it is extraordinarily difficult to find examples for which this statement is not true. This somehow extends the observation made in [7], concerning the nearly sum rate optimality of ZF beamforming vectors, to any point of the time sharing achievable region obtained with independently coded Gaussian inputs. This observation is also along the lines of observations made in the context of successive encoding regarding the nearly optimality of ZF beamformers [3]. The main conclusion that may be drawn from these observations is that, in both independent encoding and successive encoding approaches, scheduling and resources allocation are crucial and, provided that these are appropriately implemented, no significant losses can be expected from the use of ZF beamforming vectors.

Note that in this setting the combinatorial grouping approaches consider three sets and a total of four points out of $\mathcal{R}$ in order to compute any of the resulting rate vectors represented by markers in the figures. For $\mathrm{SNR} = 0$ dB and $\mathrm{SNR} = 10$ dB, these four points are enough to achieve practically all points of $\mathrm{Co}\{\mathcal{R}\}$. On the contrary, for $\mathrm{SNR} = 20$ dB, we observe that the set of rate vectors achievable with these four points is clearly non-convex. That is, in this case, either more points or better points should be preselected in order to achieve all rate vectors within $\mathrm{Co}\{\mathcal{R}\}$.

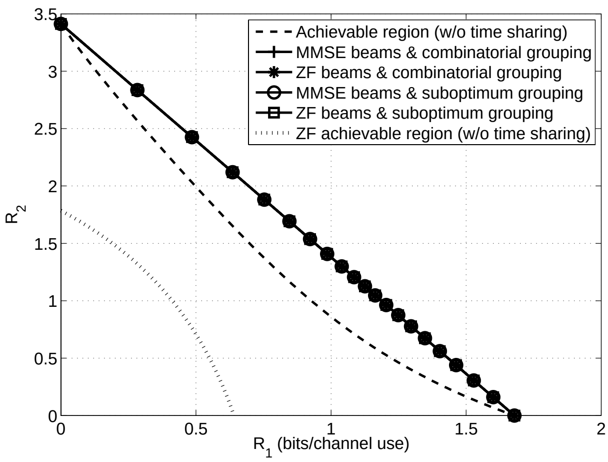

Fig. 4. Example of regions for $M = 2$, $K = 2$ and $\mathrm{SNR} = 0$ dB.

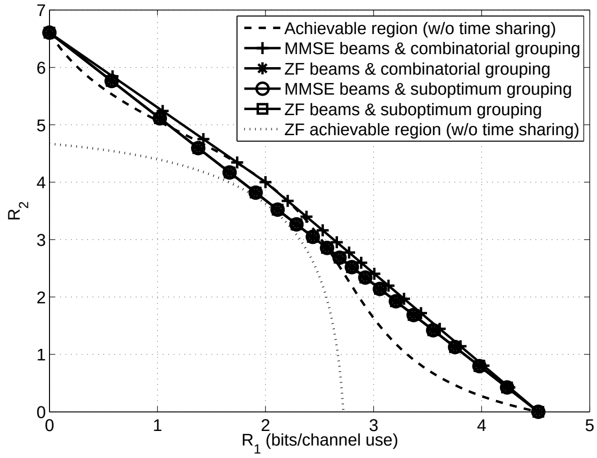

Fig. 5. Example of regions for $M = 2$, $K = 2$ and $\mathrm{SNR} = 10$ dB.

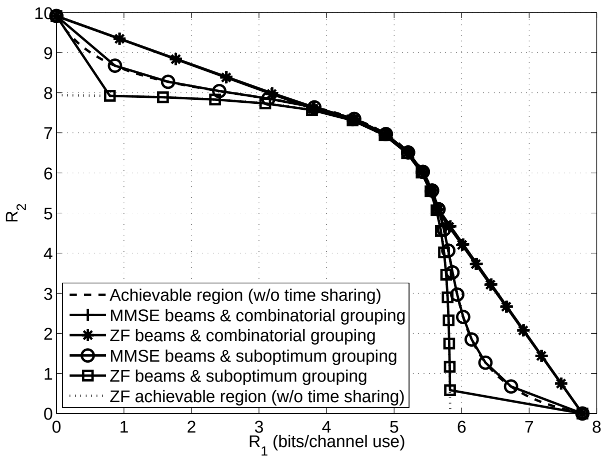

Fig. 6. Example of regions for $M = 2$, $K = 2$ and $\mathrm{SNR} = 20$ dB.

Average performance results are presented for the two-user setting with $M = 2$ in Figs. 7, 8 and 9. Averaging has been performed over a number of channel realizations whose coefficients have been independently drawn according to a complex Gaussian distribution of unit variance. The rate constraints are the same as those considered in Figs. 4, 5 and 6. The figures confirm the nearly optimality of ZF beamforming vectors for all rate constraints and SNR values. For $\mathrm{SNR} = 0$ dB and $\mathrm{SNR} = 10$ dB suboptimum grouping performs almost as well as combinatorial grouping. This is due to the fact that, at these SNR values, serving both users separately is nearly optimum for many channels. As separate scheduling of users can be done by the suboptimum grouping approaches, no large suboptimality is incurred in these cases. The most sensitive performance gaps are appreciated in Fig. 6, especially in the region between the axes and the middle part of the curve. In this areas the optimum scheduling happens frequently to be switching between a time slot with both users and a time slot with one of the users alone. This form of scheduling can not be performed by the suboptimum grouping approaches. This limitation is particularly visible in Fig. 6. There, we observe that the suboptimum grouping approaches opt for serving both users in the same time slot all over the range of rate constraints except for $n = 0$. This scheduling is optimum in the middle of the curve, where these approaches perform as well as the combinatorial grouping approaches. However, towards the sides of the curves performance worsens as time sharing is not exploited.

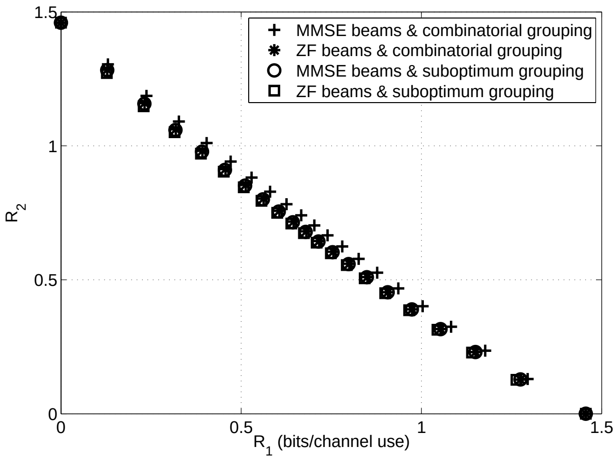

Fig. 7. Average rates for $M = 2$, $K = 2$ and $\mathrm{SNR} = 0$ dB.

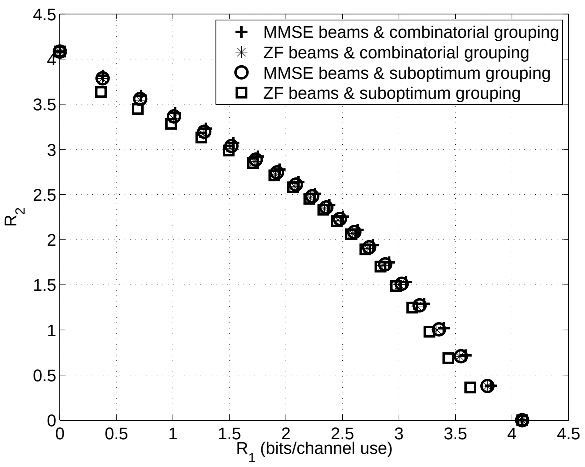

Fig. 8. Average rates for $M = 2$, $K = 2$ and $\mathrm{SNR} = 10$ dB.

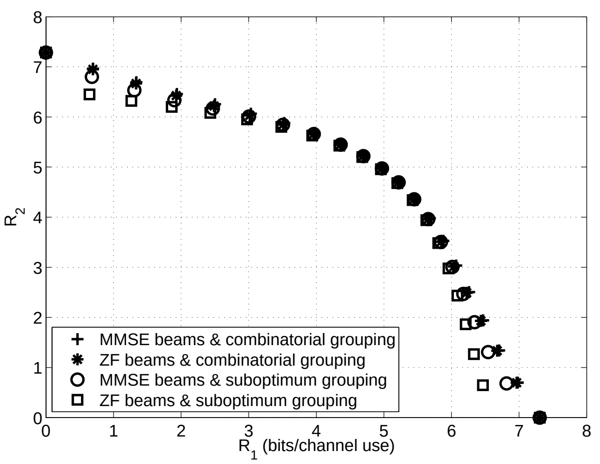

Fig. 9. Average rates for $M = 2$, $K = 2$ and $\mathrm{SNR} = 20$ dB.

7. CONCLUSION

The rate balancing problem with independent encoding of information streams has been introduced and discussed. An algorithm has been proposed to solve this problem that has the potential to reach optimality if no complexity constraints are considered. An efficient implementation of this algorithm has been introduced that is based on optimum MMSE beamforming vectors and consideration of all possible user groups with no more users than transmit antennas. This implementation has been compared with some other implementations based on ZF beamforming vectors and simpler grouping approaches. While application of ZF beamforming vectors does not incur significant performance loss, simple scheduling approaches that allocate users in no more than one time slot are clearly suboptimum especially at high SNR values and for unbalanced rate balancing constraints.

8. REFERENCES

[1] M. Schubert and H. Boche, "Iterative Multiuser Uplink and Downlink Beamforming Under SINR Constraints," *IEEE Trans. Signal Processing*, vol. 53, pp. 2324–2333, 2005.

[2] C.-H. F. Fung, W. Yu, and T. J. Lim, "Multi-antenna downlink precoding with individual rate constraints: power minimization and user ordering," in *Conference on Information Sciences and Systems*, 2004.

[3] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek, "Rate Balancing in Multiuser MIMO OFDM Systems," *Submitted to IEEE Trans. on Communications*, 2007.

[4] M. Schubert and H. Boche, "Solution of the Multiuser Downlink Beamforming Problem With Individual SINR Constraints," *IEEE Transactions on Vehicular Technology*, vol. 53, pp. 18–28, Jan. 2004.

[5] T. Yoo and A. Goldsmith, "On the Optimality of Multi-Antenna Broadcast Scheduling Using Zero-Forcing Beamforming," *IEEE Journal on Selected Topics in Communications*, pp. 528–541, 2006.

[6] M. Fuchs, G. Del Galdo, and M. Haardt, "A novel tree-based scheduling algorithm for the downlink of multi-user mimo systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2005.

[7] D. Schmidt, M. Joham, R. Hunger, and W. Utschick, "Near Maximum Sum-Rate Non-Zero-Forcing Linear Precoding with Successive User Selection," in *Proc. 40th Asilomar Conference on Signals, Systems and Computers*, November 2006.