ACHIEVING QOS AND EFFICIENCY IN THE MIMO DOWNLINK WITH LIMITED POWER

Thomas Michel and Gerhard Wunder

Fraunhofer German-Sino Lab Mobile Communications, Heinrich-Hertz-Institut
Einstein-Ufer 37, D-10587 Berlin
{michel,wunder}@hhi.fhg.de

ABSTRACT

In this paper we study the problem of maximizing a weighted sum of rates under a sum power constraint in the multiple input multiple output (MIMO) downlink while rate requirements have to be met for a subset of users. This setting reflects a more sophisticated problem of resource allocation combining an efficiency objective with strict Quality of Service (QoS) constraints. It is shown that the problem can be formulated as a convex optimization problem. Unlike in the case of pure weighted rate-sum maximization (WRSM), the question of feasibility arises. Moreover, the optimum Dirty-Paper Coding (DPC) order is not given by the ordering of the initial weights, but can be derived from the Lagrangian multipliers combined with the initial weights. An efficient algorithm based on primal-dual optimization is proposed which provably converges to the global optimum. The convergence properties are illustrated by means of numerical simulations.

1. INTRODUCTION

Not only theoretical interest but also the relevance for future wireless cellular systems such as IEEE 802.16e/m (WiMAX) and 3GPP Long Term Evolution drove recent investigations of the MIMO downlink. In this context a major question of importance is how to allocate resources in an optimal fashion so that ultimate limits could be achieved under idealized assumptions and given constraints. The insights gained from these considerations can serve as a guideline for the design of real systems and development of signal processing algorithms.

From a theoretical point of view, the MIMO downlink can be modeled as a MIMO broadcast channel (BC). If the transmitter is not restricted to linear signal processing, the system is not interference limited and nonlinear Dirty Paper Coding (DPC) can be performed assuming that perfect channel state information (CSI) is available at the transmitter as well as at the receivers. It is known that this scheme achieves the entire capacity region of the MIMO BC [1]. Finding optimum resource allocations for the MIMO BC turns out to be challenging due to the complicated mathematical structure of the BC and its *non-degradedness* caused by the spatial degrees of freedom. However, it was shown in a line of work that the capacity regions of MIMO BC and its dual MIMO multiple access channel (MAC) having hermitian transposed channel matrices coincide [1–4]. Further, duality transformations relating the particular capacity achieving power allocations in the BC to the ones in the MAC were provided so that all problems can be solved in the dual MIMO MAC which has a more favorable structure [4].

The problems tackled up to now can be roughly subdivided into two classes: *fairness* and *efficiency* optimization. In fairness-based optimization problems the objective comprises the expression

$$f_{fair}(R_1, ..., R_M) = \min_m \frac{R_m}{\gamma_m}, \tag{1}$$

where R_m is the rate allocated to user m and γ_m is its QoS demand in terms of rate. If there is no additional constraint the ratio is *balanced* for all M users $\frac{R_1}{\gamma_1} = ... = \frac{R_M}{\gamma_M}$ leading to fairness independent of the current channel realization. So the *symmetric capacity* of the MIMO BC was studied in [5], which restates in fact the rate balancing problem. In [6] the fairest corner point and DPC order of the sum capacity plane was studied. In addition, the power minimization problem under rate requirements can be interpreted in this context with $\frac{R_1}{\gamma_1} = ... = \frac{R_M}{\gamma_M} = 1$ being a fairness constraint [5, 7, 8].

On the other hand, efficiency problems do not take into account the expression in (1). The objective consists of a weighted sum of rates

$$f_{eff}(R_1, ..., R_M) = \sum_m q_m R_m \tag{2}$$

where q_m is some user weighting factor. The most important example is throughput maximization. All users are equally weighted and sum capacity is taken as the efficiency

The authors are supported in part by the *Bundesministerium für Bildung und Forschung (BMBF)* under grant FK 01 BU 350

measure; algorithms were presented for the MIMO MAC with individual power constraints in [9] and for the MIMO BC with a sum power constraint later in [10]. Algorithms for maximizing a weighted sum of rates were proposed in [11] and very recently in [12] for the MISO case. Moreover sum power minimization under a throughput constraint can be seen as an efficiency optimization problem [13].

In this paper, we bring together both criteria. The objective has the form (2) but additionally using (1) strict QoS constraints in terms of rates are taken into account

$$f_{fair}(R_1, ..., R_M) \geq 1.$$

This problem was studied already for the single antenna orthogonal frequency division multiplexing (OFDM) BC in [14] and first steps towards the MIMO case were made in [8], where the ideas from [14] using the concept of marginal utility functions was applied to the MIMO-OFDM BC. However, the convergence to the global optimum could not be proven. Furthermore, the presented algorithm is quite demanding. Here we solve the problem exploiting concepts from optimization theory for convex non-differentiable functions and provide a provably convergent algorithm based on the ellipsoid method originally introduced by Khachiyan [15].

The remainder of this paper is organized as follows. Section 2 describes the MIMO BC and MAC system model and introduces the uplink-downlink duality. In Section 3 the two antagonistic resource allocation strategies are presented, while in Section 4 the joint problem is studied. Numerical examples are illustrated in Section 5 and Section 6 concludes this paper.

1.1. Notation

Sets are represented by calligraphic letters. Lower case bold letters represent vectors and upper case bold letters denote matrices. $\mathbf{A} \succeq 0$ means that $\mathbf{A}$ is a positive semidefinite matrix, $|\cdot|$ is the determinant and $\mathrm{tr}(\cdot)$ denotes the trace operator. Applied to sets $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$. All logarithms are to the base e. A complex variable $c = a + jb$ is said to be circular symmetric complex Gaussian distributed $n \sim \mathcal{CN}(0, 1)$ if its real and imaginary part are independently distributed with $a \sim \mathcal{N}(0, 1/2)$ and $b \sim \mathcal{N}(0, 1/2)$.

2. SYSTEM MODEL AND UPLINK-DOWNLINK DUALITY

2.1. System model

Consider a frequency flat discrete MIMO broadcast channel with M receivers, each owning n_R antennas and a base station equipped with n_T transmit antennas.

Then the signal received by user m is given by

$$\mathbf{y}_m(t) = \mathbf{H}_m(t) \sum_{n=1}^{M} \mathbf{x}_n(t) + \mathbf{n}_m(t)$$

where $\mathbf{H}_m(t) \in \mathbb{C}^{n_R \times n_T}$ is the mth user's channel, $\mathbf{x}_m(t) \in \mathbb{C}^{n_T \times 1}$ is the signal dedicated to user m and $\mathbf{n}_m \in \mathbb{C}^{n_R \times 1}$ denotes the additive noise having i.i.d. circular symmetric complex Gaussian entries with unit variance.

The dual MIMO multiple access channel (MAC) is given by

$$\mathbf{r}(t) = \sum_{m=1}^{M} \mathbf{H}_m^H(t) \mathbf{s}_m(t) + \mathbf{z}(t)$$

where $\mathbf{r}(t) \in \mathbb{C}^{n_T \times 1}$ is the signal received at the base station, $\mathbf{s}_m(t) \in \mathbb{C}^{n_R \times 1}$ denotes the signal transmitted by user m and $\mathbf{z}(t) \in \mathbb{C}^{n_T \times 1}$ is again an i.i.d. circular symmetric complex Gaussian additive noise process.

In both scenarios it is assumed that the channel $\mathbf{H}(t) = [\mathbf{H}_1(t), ..., \mathbf{H}_M(t)]^T$ is known perfectly at both ends of the link. Then the base station can perform DPC in the downlink, rendering harmless interference of previously encoded users. Equivalently, the base station can apply successive interference cancellation (SIC) in the dual uplink. Then it is known that the capacity can be achieved in the downlink as well as in the uplink using codebooks with i.i.d. Gaussian entries as the number of channels uses goes to infinity.

2.2. Duality of MIMO MAC and MIMO BC

Assuming that the number of channel uses is sufficiently large during the coherence time of each fading block, the block index t can be omitted in the following assuming that capacity can be achieved asymptotically.

The capacity region of the MIMO MAC is given by

$$\mathcal{C}_{MAC}(\mathbf{H}, \bar{P}) = \bigcup_{\substack{\sum_m \mathrm{tr}(\mathbf{Q}_m) \leq \bar{P} \\ \mathbf{Q}_m \succeq 0}} \Bigg\{ \mathbf{R} : \sum_{m \in \mathcal{S}} R_m \leq \log \left| \mathbf{I} + \sum_{m \in \mathcal{S}} \mathbf{H}_m^H \mathbf{Q}_m \mathbf{H}_m \right| \quad \forall \mathcal{S} \subseteq \{1, ..., M\} \Bigg\} \tag{3}$$

where

$$\mathbf{Q}_m = \mathbb{E}\left\{\mathbf{s}_m \mathbf{s}_m^H\right\}$$

is the covariance matrix of the signal transmitted by user m and $\bar{P}$ is a sum power constraint. Assuming the same sum power constraint $\bar{P}$ for both channels, it is known from [1] that the capacity region of the Gaussian MIMO BC coincides with the capacity region of a dual MIMO multiple access channel (MAC) with hermitian transposed channel realizations

$$\mathcal{C}_{BC}(\mathbf{H}^H, \bar{P}) \equiv \mathcal{C}_{MAC}(\mathbf{H}, \bar{P})$$

where $\mathcal{C}_{MAC}(\mathbf{H}, \bar{P})$ is defined in (3). Moreover it is known that any point achievable with a certain DPC order and a given power allocation in the MIMO BC can be achieved with the reverse SIC order and a different power allocation in the MIMO MAC. The transformations relating the two transmission strategies to each other were derived in [4].

Thus all problems can equivalently be studied in the dual MIMO MAC and the subscripts $\cdot_{MAC}$ and $\cdot_{BC}$ will be omitted from now on.

3. TWO CONFLICTIVE RESOURCE ALLOCATION PRINCIPLES

In general, any point on the boundary of $\mathcal{C}(\mathbf{H}, \bar{P})$ is a solution to a weighted rate sum maximization problem for a certain set of weights $\mathbf{q} = [q_1, ..., q_M]^T$:

$$\max \sum_{m=1}^{M} q_m R_m \quad \text{subj. to} \quad \mathbf{R} \in \mathcal{C}(\mathbf{H}, \bar{P}). \tag{4}$$

The optimization problem in (4) yields a point on the boundary of $\mathcal{C}(\mathbf{H}, \bar{P})$ corresponding to a tangent hyperplane with normal vector $\mathbf{q}$. It can be formulated as a convex program and solved efficiently. Choosing $\mathbf{q} = [1, ..., 1]$ refers to the important special case of throughput maximization. This is illustrated in Figure 1). It is well known that for any fixed power allocation, i.e. for any set of fixed transmit covariance matrices $\{\mathbf{Q}_1, ..., \mathbf{Q}_M\}$, the achievable rate region is a polymatroid limited by $2^M - 1$ constraints. Due to the polymatroid structure it can shown that the optimum DPC order (BC) is given by π such that

$$q_{\pi(M)} \geq ... \geq q_{\pi(1)}$$

where user $\pi(1)$ is encoded first followed by user $\pi(2)$. This insight is crucial because it ensures the convexity of the program.

It is of considerable interest to solve (4) since the boundary of the capacity region represents the set of *efficient* rate tuples (i.e. the Pareto-optimal rate tuples). Moreover, the weights can be interpreted as *rate rewards* and may be chosen according to any weighting criterion. In particular it is known that the scheduling strategy following (4) and using as weights $\mathbf{q}$ the instantaneous queue lengths under quite mild conditions achieves the maximum region of stabilizable arrival rates [16,17]. Motivated by the relevance of (4), different algorithms have been developed (see e.g. [12] and references therein). However, note that varying fading realizations may result in very unfair rate allocations even for throughput maximization.

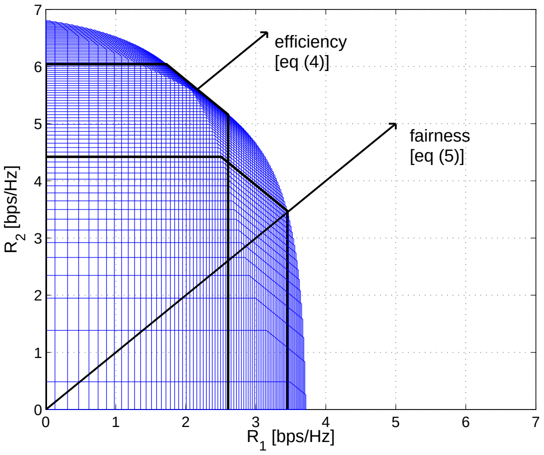

Fig. 1. Illustration of exemplary efficiency $\mathbf{q} = [1, ..., 1]^T$ and fairness $\boldsymbol{\gamma} = [1, ..., 1]^T$ optimization for random 2 user channel. Solving polymatroids highlighted.

On the other hand the minimum rate margin can be maximized leading to a *fair* rate allocation:

$$\max \min_{m} \frac{R_m}{\gamma_m} \quad \text{subj. to} \quad \mathbf{R} \in \mathcal{C}(\mathbf{H}, \bar{P}). \tag{5}$$

This approach yields a rate tuple on the boundary of $\mathcal{C}(\mathbf{H}, \bar{P})$ with allocated rates proportional to the desired set of QoS $\boldsymbol{\gamma} = [\gamma_1, ..., \gamma_M]^T$. As (4) likewise (5) can be stated as a convex program [5]. An algorithmic optimization is a bit more demanding and relies on the saddle point property the optimization of the (non-differentiable) dual function. Choosing $\boldsymbol{\gamma} = [1, ..., 1]^T$ leads to the equal rate tuple, i.e. the *symmetric capacity*.

Both approaches are illustrated exemplarily in Figure 1. It can be observed that depending on the channel realizations the solutions to (4) and (5) choosing $\mathbf{q}$ and $\boldsymbol{\gamma}$ to be the all-ones vector may differ drastically.

4. WEIGHTED RATE-SUM OPTIMIZATION WITH MINIMUM RATES

Both resource allocation principles - (4) and (5) - have certain drawbacks. In (4) instantaneous QoS-requirements being independent of the current channel realization are not considered which might lead to very unfair rate allocations for users with bad channels even if their rate reward is high. On the other hand in (5) the statistics of the fading process do not play any role: the current channel realizations have no influence on the relation of attained rates, which prevents to use forms of temporal/multiuser diversity.

In order to resolve these shortcomings, we combine the two principles and consider the following problem:

$$\begin{aligned} &\max \sum_{m=1}^{M} q_m R_m \\ &\text{subj. to } \min_{m\in\mathcal{S}} \frac{R_m}{\gamma_m} \geq 1 \quad \mathcal{S} \subseteq \{1,...,M\} \\ &\mathbf{R} \in \mathcal{C}(\mathbf{H}, \bar{P}) \end{aligned} \tag{6}$$

The set $\mathcal{S}$ comprises all users m with a strict QoS requirement γ_m, which might be caused e.g. by the service currently provided. The transmit power is limited to $\bar{P}$. Unfortunately, (6) can not be stated explicitly in terms of the transmit covariance matrices $\{\mathbf{Q}_1, ..., \mathbf{Q}_M\}$.

4.1. Convexity and feasibility

For ease of notation we introduce the QoS vector $\boldsymbol{\gamma} \in \mathbb{R}^M$ with

$$\gamma_m = 0 \quad \forall m \notin \mathcal{S}$$

in the following. Taking into account the constraints the feasible set can be made explicit yielding

$$\mathcal{R}_f(\mathbf{H}, \bar{P}, \boldsymbol{\gamma}) = \mathcal{C}(\mathbf{H}, \bar{P}) \bigcap_{i\in\mathcal{S}} \{\mathbf{R} \in \mathbb{R}^M : R_i \geq \gamma_i\} \tag{7}$$

where $\mathbf{H}$ is a set of channel realizations, $\bar{P}$ a sum power budget and γ_i are QoS requirements. The capacity region $\mathcal{C}(\mathbf{H}, \bar{P})$ is a convex set per definition and the QoS constraints define half spaces. Thus $\mathcal{R}_f(\mathbf{H}, \bar{P}, \boldsymbol{\gamma})$ is a convex set, since it comprises the intersection of convex sets. For an illustration of $\mathcal{R}_f(\mathbf{H}, \bar{P}, \boldsymbol{\gamma})$ see Figure 2.

Using (7) the main problem in (6) can be rewritten as

$$\max_{\mathbf{R}\in\mathcal{R}_f(\mathbf{H},\bar{P},\boldsymbol{\gamma})} \sum_{m=1}^{M} q_m R_m. \tag{8}$$

Since (8) consists of the maximization of an affine function over a convex set it is a convex program. As already pointed out in [8], the optimization is nevertheless nontrivial since the formulation is only implicit.

Necessary and sufficient for feasibility is the non-emptyness of the feasible set $\mathcal{R}_f(\mathbf{H}, \bar{P}, \boldsymbol{\gamma})$. Thus the problem in (6) is feasible if and only if

$$\bar{P} \geq P_{min} \tag{9}$$

where

$$\begin{aligned} P_{min} = \quad &\min \quad P \\ &\text{subj. to} \quad \boldsymbol{\gamma} \in \mathcal{C}(\mathbf{H}, P). \end{aligned} \tag{10}$$

since $\bar{P} < P_{min}$ immediately leads to $\mathcal{R}_f(\mathbf{H}, \bar{P}, \boldsymbol{\gamma}) = \emptyset$. The condition in (9) can be easily checked solving (10).

A direct formulation for (10) was given in [8] yielding $2^M - 1$ constraints each corresponding to one of the partial sums in the definition of $\mathcal{C}(\mathbf{H}, \bar{P})$ given in (3). An alternative approach was presented in [7], where the authors optimized the Lagrangian dual function. This method will turn out to be useful in the following.

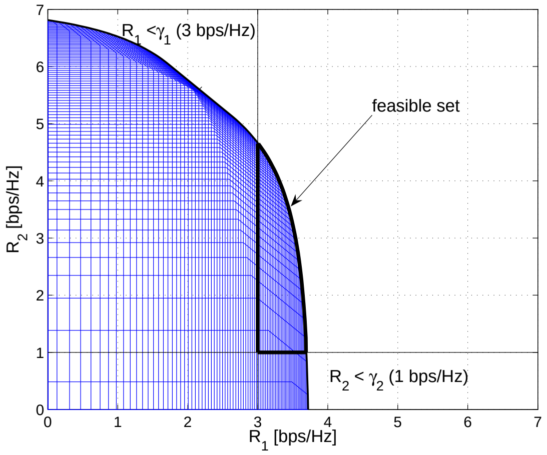

Fig. 2. Illustration of $\mathcal{R}_f(\mathbf{H}, \bar{P}, \boldsymbol{\gamma})$ for $\boldsymbol{\gamma} = [3 \;\; 1]^T$ and $\mathcal{C}(\mathbf{H}, P)$ as in Figure 1.

4.2. Optimization in the dual domain

Similar to [5, 7] the crucial idea is to solve (6) in the dual domain. To this end consider the Lagrangian of (6)

$$\mathcal{L}(\mathbf{R}, \boldsymbol{\mu}) = \sum_{m=1}^{M} q_m R_m + \sum_{i\in\mathcal{S}} \mu_i (R_i - \gamma_i).$$

Note that it is assumed that $\mathbf{R} \in \mathcal{C}(\mathbf{H}, \bar{P})$, which keeps the sum power constraint implicit. Maximizing with respect to the primal variables (i.e. $\mathbf{R}$) yields the dual function

$$g(\boldsymbol{\mu}) = \sup_{\mathbf{R}\in\mathcal{C}(\mathbf{H},\bar{P})} \mathcal{L}(\mathbf{R}, \boldsymbol{\mu})$$

where the supremum can be replaced by the maximum since $\mathcal{C}(\mathbf{H}, \bar{P})$ is compact. Now define coefficients

$$\tilde{q}_m = \begin{cases} q_m + \mu_m & m \in \mathcal{S} \\ q_m & \text{otherwise} \end{cases} \tag{11}$$

summing up the Lagrangian multipliers $\boldsymbol{\mu}$ and the initial weights $\mathbf{q}$. Then the dual function can be rewritten as

$$g(\boldsymbol{\mu}) = \max_{\mathbf{R}\in\mathcal{C}(\mathbf{H},\bar{P})} \sum_{m=1}^{M} \tilde{q}_m R_m - \sum_{i\in\mathcal{S}} \mu_i \gamma_i. \tag{12}$$

Obviously (12) is an affine version of (4) which can be efficiently solved. However, the dual function may be nondifferentiable at least for some $\boldsymbol{\mu}$. Thus optimization methods which do not rely on derivatives or gradients have to be used. In general any cutting plane method is suitable for minimizing (12).

In the following we make use of the ellipsoid method, which has been applied several times recently [5, 7]. The ellipsoid method generates a sequence of shrinking ellipses containing the solution. It can be interpreted as a kind of cutting plane method, which rules out half spaces according to the evaluation of a subgradient[1]. At an arbitrary point $\hat{\boldsymbol{\mu}}$ a subgradient can be found using the definition of the dual function: Let $\hat{\mathbf{R}}$ be the solution to

$$\hat{\mathbf{R}} = \arg \max_{\mathbf{R} \in \mathcal{C}(\mathbf{H}, \bar{P})} \mathcal{L}(\mathbf{R}, \hat{\boldsymbol{\mu}}).$$

Then

$$\begin{aligned} g(\boldsymbol{\mu}) &= \max_{\mathbf{R} \in \mathcal{C}(\mathbf{H}, \bar{P})} \mathcal{L}(\mathbf{R}, \boldsymbol{\mu}) \\ &\geq \mathcal{L}(\hat{\mathbf{R}}, \boldsymbol{\mu}) \\ &= g(\hat{\boldsymbol{\mu}}) + \sum_{i \in \mathcal{S}} (\mu_i - \hat{\mu}_i)(\hat{R}_i - \gamma_i). \end{aligned} \tag{13}$$

Defining the vector

$$\boldsymbol{\nu} = \hat{\mathbf{R}} - \boldsymbol{\gamma} \tag{14}$$

it can be easily seen using (13) that

$$\boldsymbol{\nu} \in \partial g(\hat{\boldsymbol{\mu}}),$$

where $\partial g(\hat{\boldsymbol{\mu}})$ is the subdifferential of $g(\boldsymbol{\mu})$ at $\hat{\boldsymbol{\mu}}$. Thus, given $\hat{\boldsymbol{\mu}}$, the half space of dual parameters corresponding to

$$\boldsymbol{\mu} : (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^T \boldsymbol{\nu} \geq 0$$

can be ruled out.

The entire procedure is summarized in Algorithm 1. In the following subsection the algorithm initialization is studied, which is not trivial due to the possibly very limited size of the feasible set. Furthermore we comment on the fifth step of Algorithm1.

4.3. Algorithm initialization: bounds on the optimum Lagrangian multipliers $\boldsymbol{\mu}^*$

In order to apply the ellipsoid or any other cutting plane method a necessity is that an initial set containing the solution can be specified. To this end upper bounds on the Lagrangian multipliers are needed. In many cases this is not a major problem, since a feasible point may be easy to find. Unlike in [5] where the search can be limited to the unit sphere, it is not trivial to find an initial ellipsoid covering the optimal Lagrangian multipliers $\boldsymbol{\mu}^*$ in the first step of Algorithm 1. The following proposition provides a bound on the optimum dual variables $\boldsymbol{\mu}^*$:

Algorithm 1 MIMO WRSM with Minimum Rates

```
(0) check feasibility by solving (10)
if problem is feasible then
  (1) initialize $\boldsymbol{\mu}^{(0)}$ according to
      $\mu_m^{(0)} = \begin{cases} \theta_m/2 & m \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}$
  with $\theta_m$ defined in (20) and choose an initial ellipse $\mathbf{M}^{(0)}$ such that
      $||\boldsymbol{\Gamma}^{(0)^{1/2}}(\mathbf{x} - \boldsymbol{\mu}^{(0)})|| \leq 1$
  for all $\mathbf{x}$ with $0 \leq\ x_m \leq\ \theta_m \quad \forall m = 1, ..., M.$
  while desired accuracy not reached do
    (2) with $\tilde{\mathbf{q}}^{(n)}$ defined in (11) solve
        $\mathbf{R}^{(n)} = \arg \max_{\mathcal{C}(\mathbf{H}, \bar{P})} \sum_{m=1}^{M} \tilde{q}_m^{(n)} R_m \qquad (15)$
    (3) determine subgradient $\boldsymbol{\nu}^{(n)}$ according to
        $\boldsymbol{\nu}^{(n)} = \mathbf{R}^{(n)} - \boldsymbol{\gamma} \qquad (16)$
    (4) update ellipse
        $\boldsymbol{\Gamma}^{(n+1)} = \frac{|\mathcal{S}|^2 - 1}{|\mathcal{S}|^2} \left( \boldsymbol{\Gamma}^{(n)} + \frac{2}{|\mathcal{S}| - 1} \frac{\boldsymbol{\nu}^{(n)} \boldsymbol{\nu}^{(n)^T}}{\boldsymbol{\nu}^{(n)^T} \boldsymbol{\Gamma}^{(n)^{-1}} \boldsymbol{\nu}^{(n)}} \right) \qquad (17)$
    with new centroid
        $\boldsymbol{\mu}^{(n+1)} = \boldsymbol{\mu}^{(n)} + \frac{1}{|\mathcal{S}|} \frac{\boldsymbol{\Gamma}^{(n)^{-1}} \boldsymbol{\nu}^{(n)}}{\sqrt{\boldsymbol{\nu}^{(n)^T} \boldsymbol{\Gamma}^{(n)^{-1}} \boldsymbol{\nu}^{(n)}}} \qquad (18)$
    (5) assure that $\boldsymbol{\mu}^{(n+1)} \in \mathbb{R}_+^M$
  end while
end if
```

Proposition 1. *Let*

$$\boldsymbol{\mu}^* = \arg \min_{\boldsymbol{\mu} \succeq \mathbf{0}} g(\boldsymbol{\mu})$$

be the optimum dual variables, with $g(\boldsymbol{\mu})$ defined in (12). Assume that there exists a rate tuple $\mathbf{R} \in \operatorname{relint} \mathcal{R}_f(\mathbf{H}, \bar{P}, \boldsymbol{\gamma})$. Then

$$0 \leq \mu_m^* \leq \theta_m, \quad m \in \mathcal{S} \tag{19}$$

where

$$\theta_m = \frac{\sum_{n=1}^{M} q_n \left(R_n^{su}(\bar{P}) - R_n \right)}{R_m - \gamma_m} \tag{20}$$

with $R_n^{su}(\bar{P})$ being the maximum single user rate corre-

[1] For further details the interested reader is referred to [18, 19].

sponding to the water-filling solution

$$R_n^{su}(\bar{P}) = \max_{\mathbf{Q}\succeq\mathbf{0}:\mathrm{tr}(\mathbf{Q})\leq\bar{P}} \log\left|\mathbf{I}+\mathbf{H}_n^H\mathbf{Q}\mathbf{H}_n\right|. \qquad (21)$$

Proof. An upper bound on $g(\boldsymbol{\mu}^*)$ can be found as follows.

$$\begin{aligned} g(\boldsymbol{\mu}^*) =& \mathcal{L}(\mathbf{R}^*, \boldsymbol{\mu}^*) \\ =& \sum_{n=1}^{M} q_n R_n^* + \sum_{n=1}^{M} \mu_n^*(R_n^* - \gamma_n) \qquad (22) \\ \leq& \sum_{n=1}^{M} q_n R_n^{su} \end{aligned}$$

The last inequality follows from the fact that the second addend in (22) is zero and $R_n^{su} \geq R_n^*$ for all n.

On the other hand we have

$$\begin{aligned} g(\boldsymbol{\mu}^*) =& \mathcal{L}(\mathbf{R}^*, \boldsymbol{\mu}^*) \\ \geq& \mathcal{L}(\mathbf{R}, \boldsymbol{\mu}^*) \\ =& \sum_{n=1}^{M} q_n R_n + \sum_{n=1}^{M} \mu_n^*(R_n - \gamma_n) \qquad (23) \\ \geq& \sum_{n=1}^{M} q_n R_n + \mu_m^*(R_m - \gamma_m) \end{aligned}$$

Combining (22) and (23) and solving for μ_m^* leads to (19). □

Note that (21) is very easy to solve and $R_n^{su}(\bar{P})$ can be calculated using water-filling over the inverse eigenvalues of $\mathbf{H}_n^H\mathbf{H}_n$. In principle any upper bound on $\sum_{n=1}^{M} q_n R_n^*$ can be used: For example

$$\begin{aligned} \sum_{n=1}^{M} q_n R_n^* \leq& \max_n q_n \sum_{n=1}^{M} R_n^* \qquad (24) \\ \leq& \max_n q_n R_{sum} \end{aligned}$$

can be used, where R_{sum} is the sum capacity:

$$R_{sum} = \max_{\mathbf{Q}:\sum_{n=1}^{M}\mathrm{tr}(\mathbf{Q}_n)\leq\bar{P}} \log\left|\mathbf{I}+\sum_{n=1}^{m}\mathbf{H}_n^H\mathbf{Q}_n\mathbf{H}_n\right|.$$

However, using the single user water-filling solutions $\mathbf{R}^{su}$ is very appealing due to its simplicity.

It remains to find an interior point $\mathbf{R}$ within the relative interior of the feasible set $\mathcal{R}_f(\mathbf{H}, \bar{P}, \boldsymbol{\gamma})$. Defining

$$\mathbf{R}_\epsilon = [\gamma_1 + \epsilon, ..., \gamma_M + \epsilon]^T$$

this can be done solving

$$\begin{aligned} P_\epsilon = & \quad \min \;\; P \\ & \text{subj. to} \;\; \mathbf{R}_\epsilon \in \mathcal{C}(\mathbf{H}, P) \end{aligned} \qquad (25)$$

while choosing ϵ such that $P_\epsilon < \bar{P}$. Note that the volume of the initial ellipse and thus the number of iterations till convergence of Algorithm 1 depends on the tightness of the derived bounds on $\boldsymbol{\mu}^*$.

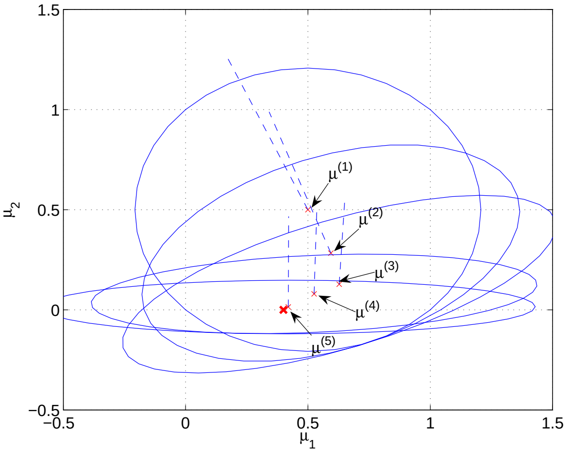

Fig. 3. Exemplary convergence behavior for a random 2 user example with $\boldsymbol{\mu}^* = [0.4 \quad 0]$ (major red x). Dashed lines represent subgradients.

4.4. Convergence Issues

It can be proven that the sequence of centroids $\boldsymbol{\mu}^{(n)}$ generated by Algorithm 1 converges to the optimum Lagrangian multipliers:

$$\lim_{n\to\infty} \boldsymbol{\mu}^{(n)} = \boldsymbol{\mu}^*. \qquad (26)$$

The proof is standard. See e.g. [19].

Some words have to be said about step 5 of Algorithm 1, which is not specified in detail up to now. It is not clear from the beginning which QoS constraints will turn out to be active or not (see e.g. user 3 in Figure 4 and user 2 in Figure 3). If a QoS constraint γ_m is not active we consequently get $\mu_m^* = 0$. The dual (12) is defined on $\boldsymbol{\mu} \in \mathbb{R}_+^M$. However, the ellipsoid algorithm may generate iterates $\boldsymbol{\mu}^{(n)}$ with negative components $\mu_m^{(n)} < 0$ such that $g(\boldsymbol{\mu}^{(n)})$ does not exist. In this case choose a subgradient $\boldsymbol{\nu}$ with

$$\nu_m = \begin{cases} -1 & \text{if } \mu_m^{(n)} < 0 \\ 0 & \text{otherwise} \end{cases} \qquad (27)$$

and proceed with step 4 of the algorithm. This procedure is repeated until $\boldsymbol{\mu}^{(n+1)} \in \mathbb{R}_+^M$ and thus the negative half-space is ruled out. Note that no evaluation of the dual is needed while shrinking the ellipse further. However, this may occur various times since in contrast to pure cutting plane methods the ellipsoid method adds new parts to the remaining feasible set (see Figure 3).

4.5. Optimum DPC order

As already studied in [8], the *sum* of Lagrangian multipliers $\boldsymbol{\mu}^*$ and initial weights $\mathbf{q}$ yielding $\tilde{\mathbf{q}}$ as defined in (11) reveals the optimal Dirty-Paper encoding order $\pi(\cdot)$:

$$\tilde{q}_{\pi(1)} \geq ... \geq \tilde{q}_{\pi(M)}. \qquad (28)$$

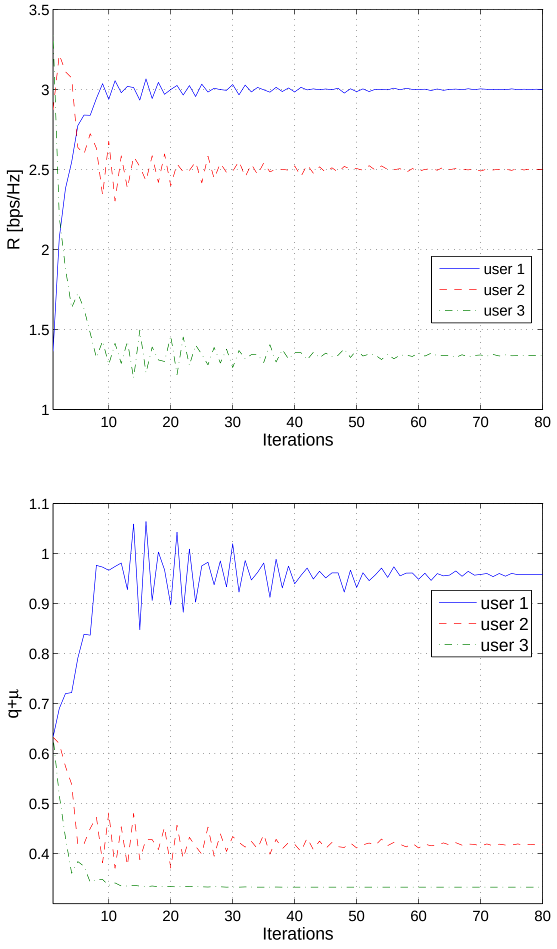

Fig. 4. Simulation example for a three user system with $n_T = 3$, $n_R = 1$, $\mathbf{q} = [1/3 \;\; 1/3 \;\; 1/3]^T$ and QoS parameters $\boldsymbol{\gamma} = [3 \;\; 2.5 \;\; 0.5]^T$.

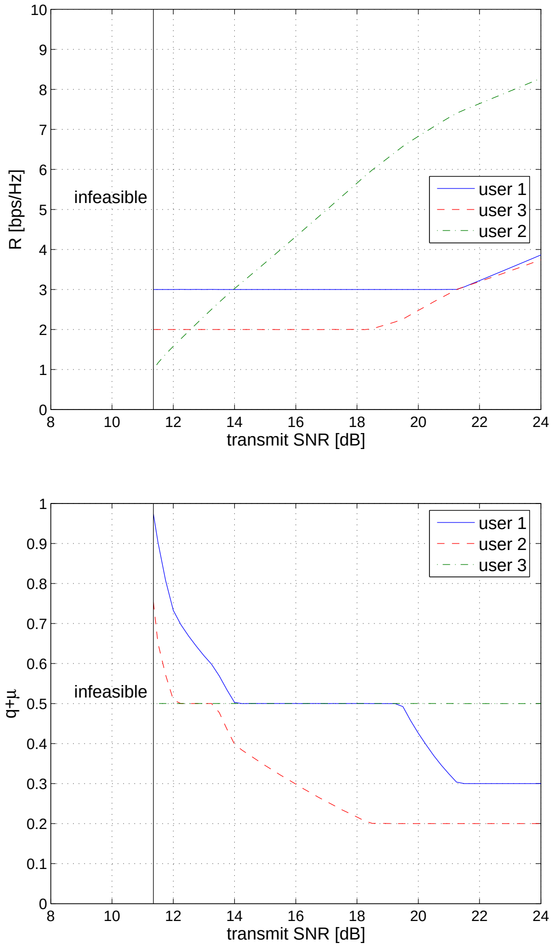

Fig. 5. Rate and weights over transmit SNR for an exemplary three user system with $n_T = 3$, $n_R = 1$, $\mathbf{q} = [0.3 \;\; 0.2 \;\; 0.5]^T$ and QoS parameters $\boldsymbol{\gamma} = [3 \;\; 2 \;\; 1]^T$.

This is a direct consequence of the polymatroid structure of the capacity region for a fixed resource allocation. So in contrast to (4), where the optimal encoding order depends only on the initial weights $\mathbf{q}$, the DPC order is *not* known in advance. This result is similar to [14].

5. SIMULATIONS

In Figure 3 the ellipsoid method is illustrated for a two user example. It can be seen that the algorithm generates a sequence of ellipses with shrinking volume containing the solution. It can be observed that the sequence of centroids $\boldsymbol{\mu}^{(n)}$ converges to $\boldsymbol{\mu}^*$. The dashed lines represent the subgradients $\boldsymbol{\nu}^{(n)}$ determining the half-space to be ruled out.

Figure 4 shows the convergence of Algorithm 1 for a three user system with an exemplary random channel. The base station has $n_T = 3$ antennas while the users own $n_R = 1$ each. The weights are given by $\mathbf{q} = [1/3 \;\; 1/3 \;\; 1/3]^T$ (thus the objective is throughput) and the QoS constraints are $\boldsymbol{\gamma} = [3 \;\; 2.5 \;\; 0.5]^T$ bps/Hz. The transmit SNR is 10dB. The upper plot depicts the achieved rates and the lower plot the corresponding sum of weights and Lagrangian multipliers $\mathbf{q} + \boldsymbol{\mu}$. Obviously the QoS constraint of user 3 is not active resulting in $\mu_3^* = 0$, while users 1 and 2 achieve their QoS requirements of 3 and 2.5 bps/Hz having active constraints $\mu_i^* > 0, \;\; i = 1, 2$. In Figure 5, the development of rates and $\mathbf{q} + \boldsymbol{\mu}$ over SNR is depicted for an exemplary system with random channel realizations, $\mathbf{q} = [0.3 \;\; 0.2 \;\; 0.5]^T$ and $\boldsymbol{\gamma} = [3 \;\; 2 \;\; 1]^T$. Below a transmit SNR of 11.35 dB the problem is infeasible. The Lagrangian multipliers decrease monotonously with increasing SNR. As the constraints become inactive, the corresponding users' rate starts to increase. Note that there exist two fractions (users 2,3 and 1,3) within the SNR range where $\tilde{q}_m = \tilde{q}_n$ which corresponds to time-sharing among users m and n.

6. CONCLUSIONS

In this paper, we studied the problem of maximizing a certain efficiency objective given by a weighted sum of rates in the MIMO BC while the sum power is limited and QoS constraints have to be met for a subset of users. This allows a certain tradeoff between efficiency and fairness. We presented an algorithm based on optimization of the Lagrangian dual which provably converges to the global optimum. The question of feasibility was studied and bounds on the optimum Lagrangian multipliers were derived. These are necessary for the initialization of the algorithm. The convergence properties were illustrated and numerical simulations validated the theoretical analysis.

7. REFERENCES

[1] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian MIMO broadcast channel," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 3936–3964, Sept. 2006.

[2] G. Caire and S. Shamai, "On achievable rates in a multiantenna gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[3] P. Vishwanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inform. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.

[4] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.

[5] J. Lee and N. Jindal, "Symmetric capacity of MIMO downlink channels," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Seattle, July 2006.

[6] M. Maddah-Ali, A. Mobasher, and A. Khandani, "Fairness in multiuser systems with polymatroid capacity region," *IEEE Trans. Inform. Theory*, 2006, submitted.

[7] M. Mohseni, R. Zhang, and J. Cioffi, "Optimized transmission for fading multiple-access and broadcast channels with multiple antennas," *IEEE Journal on Selected Areas in Comm.*, vol. 24, no. 8, pp. 1627–1639, Aug. 2006.

[8] G. Wunder and T. Michel, "Minimum rates scheduling for MIMO-OFDM broadcast systems," in *Proc. IEEE Int. Symp. on Spread Spectrum Techniques and Applications (ISSSTA)*, Manaus, Aug. 2006.

[9] W. Yu, W. Rhee, S. Boyd, and J. Cioffi, "Iterative water-filling for Gaussian vector multiple access channels," *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 145–151, Jan. 2004.

[10] N. Jindal, W. Rhee, S. Vishwanath, S. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.

[11] H. Viswanathan, S. Venkatesan, and H. Huang, "Downlink capacity evaluation of cellular networks with known interference cancellation," *IEEE Journal on Selected Areas in Comm.*, vol. 21, no. 5, pp. 802–811, May 2003.

[12] M. Kobayashi and G. Caire, "An Iterative Water-Filling Algorithm for Maximum Weighted Sum-Rate of Gaussian MIMO-BC," *IEEE Journal on Selected Areas in Comm.*, vol. 24, no. 8, pp. 1640–1646, Aug. 2006.

[13] T. Michel and G. Wunder, "Sum rate iterative water-filling for Gaussian MIMO broadcast channels," in *Proc. Intern. Symp. On Wireless Personal Multimedia Communications (WPMC)*, San Diego, Sept. 2006.

[14] ——, "Minimum rates scheduling for OFDM broadcast channels," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, Toulouse, May 2006.

[15] L. Khachiyan, "A polynomial algorithm in linear programming," in *Soviet Mathematics Doklady*, 1979, pp. 191–194.

[16] M. Neely, E. Modiano, and C. Rohrs, "Power allocation and routing in multibeam satellites with time-varying channels," *IEEE/ACM Trans. on Networking*, vol. 11, pp. 138–152, Feb. 2003.

[17] H. Boche and M. Wiczanowski, "Stability-optimal transmission policy for multiple antenna multiple access channel in the geometric view," *EURASIP Signal Processing Journal, Special Issue on Advances in Signal Processing-assisted Cross-layer Designs*, 2006, to appear.

[18] R. Freund and C. Roos, "Lecture Notes, TU Delft, Course WI4 060," 2004, available at *www.isa.ewi.tudelft.nl/ roos/courses/wi485/ellips.pdf.*

[19] S. Boyd, "Lecture Notes, Stanford University, Class EE392o," 2006, available at *www.stanford.edu/class/ee392o/elp.pdf.*