Rate Balancing in Multiuser MIMO OFDM Systems

Pedro Tejera, Wolfgang Utschick, *Senior Member, IEEE,* Josef A. Nossek, *Fellow, IEEE,*
and Gerhard Bauch, *Senior Member, IEEE*

***Abstract*—Recently, the capacity region of the Gaussian broadcast channel has been characterized. For a given transmit power constraint, those points on the boundary of the capacity region can be regarded as the set of optimal operational points. The present work addresses the problem of selecting the point within this set that satisfies given constraints on the ratios between rates achieved by the different users in the network. This problem is usually known as rate balancing.**

To this end, the optimum iterative approach for general MIMO channels is revisited and adapted to an OFDM transmission scheme. Specifically, an algorithm is proposed that exploits the structure of the OFDM channel and whose convergence speed is essentially insensitive to the number of subcarriers. This is in contrast to a straightforward extension of the general MIMO algorithm to an OFDM scheme. Still, relatively high complexity and the need of a time-sharing policy to reach certain rates are at least two obstacles for a practical implementation of the optimum solution. Based on a novel decomposition technique for broadcast channels a suboptimum non-iterative algorithm is introduced that does not require time-sharing and very closely approaches the optimum solution.

***Index Terms*—MIMO systems, multiuser channels, OFDM, rate balancing, successive encoding.**

I. Introduction

INCREASING demand for broadband services calls for higher data rates in future wireless communication systems [1]. Data rates of several Mb/s for high mobility scenarios and up to 1 Gb/s in low mobility or static scenarios are expected in fourth generation systems. In the way to such transmission rates there are two major barriers to overcome. The first is scarcity of spectrum, which limits the amount of bandwidth available for transmission. The second is the wireless channel that severely distorts the signal due to multipath propagation.

The use of multiple antennas increases capacity. On the other hand, multicarrier technology effectively combats the effects of multipath propagation. Therefore, combination of multiple antennas and multicarrier technology seems to be ideal to achieve the expected rates under the mentioned constraints [2]. In the work at hand we consider the downlink of a wireless communication system with multiple antennas at the transmitter and the receivers and orthogonal frequency division multiplexing (OFDM) as transmission scheme. We assume that receivers know their respective transmission channels perfectly and the transmitter has perfect knowledge of the channel of every user. This assumption presumes perfect channel estimates, a quasi-static scenario, in which channels do not change significantly for the time between two consecutive channel estimations, and channel reciprocity in case of a time division duplex (TDD) system or a feedback link in case of a frequency division duplex (FDD) system. Under these assumptions we study approaches that aim at maximizing the sum of rates with a constraint on the ratios between the final rates achieved by the users in the network. To be more precise, if R_k is the rate obtained by user k and R is the sum of the rates of all users, the constraint can be given as a vector of relative rates $\boldsymbol{\rho} = [\ \rho_1 \quad \cdots \quad \rho_K \quad]^{\mathrm{T}}$, where $\rho_k = R_k/R$. This can be regarded as a quality of service (QoS) constraint that gives the transmitter the opportunity to choose the transmission strategy considering possible rate requirements coming from higher system layers.

In the context of OFDM and SISO channels rate balancing approaches have been proposed in [3]–[6]. In the context of multiple input multiple output (MIMO) channels, initial work reported in [7] presents a necessary condition for the optimal solution and several suboptimum algorithms. More recently, an optimal non-iterative algorithm has been presented in [8] for the case of single receive antennas and fixed encoding/decoding order. This solution is based on recently appeared duality results between the multiple access channel (MAC) and the broadcast channel (BC) [8], [9]. Essentially the same algorithm can be found in [10] applied to the minimization of power with given rate constraints. The main drawback of this approach is the limitation to single receive antennas and the non-optimization of encoding/decoding order. Moreover, time-sharing points on the boundary of the capacity region are not reachable. Also for the problem of power minimization with given rate constraints, an optimum iterative algorithm is proposed in [10] that applies to the general MIMO setting with multiple receive antennas and incorporates optimization of encoding/decoding order. There, time-sharing points can also be identified as solutions of the optimization problem. The same subgradient approach used in this algorithm can be followed in order to solve the rate balancing problem. However, convergence speed and convergence itself very strongly depend on the choice of step sizes. Recently, an algorithm has been presented in [11] that uses the ellipsoid method in order to solve the rate balancing problem. Although this method guarantees convergence in polynomial time, it is known to be very slow in practice [12]. Basically, both approaches, i.e., the subgradient approach in [10] and the ellipsoid algorithm in [11], iteratively search in the space of user priorities until

Paper approved by R. W. Heath, Jr., the Editor for MIMO Techniques of the IEEE Communications Society. Manuscript received December 1, 2006; revised June 8, 2007 and November 28, 2007.

P. Tejera, W. Utschick, and J. A. Nossek are with Technische Universität München, Theresienstr. 90, 80290 München, Germany (e-mail: pedro.tejera@mytum.de; utschick@tum.de; josef.a.nossek@nws.ei.tum.de).

G. Bauch is with DOCOMO Communication Laboratories Europe GmbH, Landsbergerstr. 312, 80687 München, Germany (e-mail: bauch@docomolab-euro.com) and with Universität der Bundeswehr München, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany.

Digital Object Identifier 10.1109/TCOMM.2009.05.060656

a vector of priorities is found such that the corresponding rate vector satisfies the given QoS constraint. To this end, within one iteration a weighted sum rate optimization problem has to be solved in order to obtain the rate vector corresponding to a given priorities vector. For this problem an iterative line search algorithm has been proposed in [13]. However, a straightforward extension of this algorithm to OFDM requires a number of iterations in order to reach convergence that seems to grow linearly with the number of subcarriers. This, of course, poses serious impediments to the applicability of this algorithm in broadband communication systems.

In this work, exploiting the block diagonal structure of the OFDM channel, an algorithm is introduced that divides the original weighted sum rate optimization problem into a set of smaller subcarrier specific optimization problems and a power allocation problem in the frequency domain. As a consequence, convergence becomes essentially insensitive to the number of subcarriers. Still, the relatively high complexity involved in the computation and implementation of the optimum solution motivates the introduction of a suboptimum non-iterative approach. This approach is based on a decomposition algorithm called cooperative zero-forcing with successive encoding and successive allocation method (CZF-SESAM) that was first introduced in [14]. The algorithm sequentially assigns non-interfering spatial dimensions to users in successive steps and ensures that the number of spectral components assigned to each user in each of these steps is consistent with the given QoS constraint. Full compliance with the QoS constraint is enforced by subsequent power loading over the set of allocated subchannels. Besides the fact that no iterations are needed, the solution obtained from application of this algorithm can be realized without resorting to time-sharing strategies. Note that practical implementation of time-sharing requires longer transmission times in order to achieve nearly error-free transmission as well as signaling multiple transmission strategies, which increases signaling overhead. Simulation results show that performance of this approach is in most cases almost as good as that of the optimum iterative approach.

The remainder of this paper is organized as follows. In Section II the system model is introduced. In Section III the optimum solution to the rate balancing problem is discussed and adapted to the OFDM setting. In Section IV a suboptimum non-iterative approach to the rate balancing problem is presented. Numerical results are provided in Section V and, finally, conclusions are drawn in Section VI.

In the following, vectors and matrices are denoted by lower case bold and capital bold letters, respectively. Random variables are represented by sans-serif characters. We use $(\bullet)^*$ for complex conjugation, $(\bullet)^{\mathrm{T}}$ for matrix transposition and $(\bullet)^{\mathrm{H}}$ for conjugate transposition. $\mathrm{E}\{\bullet\}$ and $\mathrm{Tr}\{\bullet\}$ denote the expectation and trace operators, respectively. Given a matrix $\boldsymbol{A}$, $|\boldsymbol{A}|$ represents its determinant. For Hermitian matrices, $\boldsymbol{A} \geq 0$ indicates that matrix $\boldsymbol{A}$ is positive semidefinite. Letting $\{\boldsymbol{A}_i\}_{i=1,\cdots,I}$ be the set of all matrices indexed by the variable i, $\mathrm{diag}[\boldsymbol{A}_1,\ldots,\boldsymbol{A}_I]$ represents a block diagonal matrix with matrices $\{\boldsymbol{A}_i\}_{i=1,\cdots,I}$ as blocks in the main diagonal. Finally, the identity matrix of dimension q is denoted by $\boldsymbol{I}_q$ and its sth column by $\boldsymbol{e}_s$.

II. System Model

We consider the downlink of a cellular wireless communication system. The base station is equipped with t transmit antennas. Each user $k \in \{1,\ldots,K\}$ has r_k receive antennas. An OFDM transmission scheme is employed with a cyclic prefix that is assumed to be longer than the length of the power delay profile of the channel so that no intersymbol interference occurs. The channel is assumed to be invariant for the duration of an OFDM symbol so that orthogonality between subcarriers is preserved during transmission. According to these assumptions the relationship between the vector of transmitted signals $\boldsymbol{x}_n \in \mathbb{C}^{t\times 1}$ and the vector $\boldsymbol{y}_{n,k} \in \mathbb{C}^{r_k\times 1}$ of received signals for user k at subcarrier $n \in \{1,\ldots,N\}$ can be expressed as

$$\boldsymbol{y}_{n,k} = \boldsymbol{H}_{n,k}\boldsymbol{x}_n + \boldsymbol{w}_{n,k},$$

where $\boldsymbol{H}_{n,k} \in \mathbb{C}^{r_k\times t}$ is the channel matrix of user k at subcarrier n and $\boldsymbol{w}_{n,k} \in \mathbb{C}^{r_k\times 1}$ a realization of a zero-mean circularly symmetric complex Gaussian distributed random variable $\mathsf{w}_{n,k}$ representing noise with covariance matrix $\mathrm{E}\{\mathsf{w}_{n,k}\mathsf{w}_{n,k}^{\mathrm{H}}\} = \boldsymbol{I}_{r_k}$. Noise processes of different subcarriers are assumed to be uncorrelated. The transmitter is assumed to perfectly know all matrices $\boldsymbol{H}_{n,k}$ and the average transmit power over the whole spectrum is limited, i.e.,

$$\frac{1}{N}\sum_{n=1}^{N}\mathrm{Tr}\left\{\mathrm{E}\{\mathsf{x}_n\mathsf{x}_n^{\mathrm{H}}\}\right\} \leq P_{\mathrm{Tx}}. \tag{1}$$

A MIMO OFDM system can be viewed as a MIMO non frequency selective system where blocks of transmit and receive antennas are decoupled from each other. Specifically, if we define

$$\tilde{\boldsymbol{H}}_k = \mathrm{diag}\left[\begin{array}{ccc}\boldsymbol{H}_{1,k} & \cdots & \boldsymbol{H}_{N,k}\end{array}\right],$$

$\tilde{\boldsymbol{y}}_k = \left[\begin{array}{ccc}\boldsymbol{y}_{1,k}^{\mathrm{T}} & \cdots & \boldsymbol{y}_{N,k}^{\mathrm{T}}\end{array}\right]^{\mathrm{T}}$, $\tilde{\boldsymbol{n}}_k = \left[\begin{array}{ccc}\boldsymbol{n}_{1,k}^{\mathrm{T}} & \cdots & \boldsymbol{n}_{N,k}^{\mathrm{T}}\end{array}\right]^{\mathrm{T}}$ and $\tilde{\boldsymbol{x}} = \left[\begin{array}{ccc}\boldsymbol{x}_1^{\mathrm{T}} & \cdots & \boldsymbol{x}_N^{\mathrm{T}}\end{array}\right]^{\mathrm{T}}$, we can write

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{n}}, \tag{2}$$

with $\tilde{\boldsymbol{y}} = \left[\begin{array}{ccc}\tilde{\boldsymbol{y}}_1^{\mathrm{T}} & \cdots & \tilde{\boldsymbol{y}}_K^{\mathrm{T}}\end{array}\right]^{\mathrm{T}}$, $\tilde{\boldsymbol{n}} = \left[\begin{array}{c}\tilde{\boldsymbol{n}}_1^{\mathrm{T}}\cdots\tilde{\boldsymbol{n}}_K^{\mathrm{T}}\end{array}\right]^{\mathrm{T}}$ and $\tilde{\boldsymbol{H}} = \left[\begin{array}{ccc}\tilde{\boldsymbol{H}}_1^{\mathrm{T}} & \ldots & \tilde{\boldsymbol{H}}_K^{\mathrm{T}}\end{array}\right]^{\mathrm{T}}$. Equation (2) corresponds to the usual MIMO BC model. Therefore, at least conceptually, every algorithm applicable to the MIMO BC can straightforwardly be extended to a MIMO OFDM setting. The essential difference consists of the high dimensionality of OFDM systems and the special structure of matrix $\tilde{\boldsymbol{H}}$ that, as we shall see in the next section, can be exploited to improve performance of iterative optimization algorithms. The vector of transmitted signals in (2), i.e., $\tilde{\boldsymbol{x}}$, results from the superposition of user specific signals $\tilde{\boldsymbol{s}}_k$ with covariance matrices $\tilde{\boldsymbol{\Sigma}}_k$, i.e., $\tilde{\boldsymbol{x}} = \sum_{k=1}^{K}\tilde{\boldsymbol{s}}_k$. Assuming statistical independence of the signals transmitted to different users, (1) can be rewritten as

$$\frac{1}{N}\sum_{k=1}^{K}\mathrm{Tr}\left\{\tilde{\boldsymbol{\Sigma}}_k\right\} \leq P_{\mathrm{Tx}}. \tag{3}$$

III. Optimum Approach

A. Problem Formulation

The rate balancing problem in the MIMO BC can be mathematically formulated as

$$\max_{\boldsymbol{r},\gamma} \gamma \quad \text{s. t.} \quad \boldsymbol{r} = \gamma\boldsymbol{\rho},\ \boldsymbol{r} \in \mathcal{C}\left(\tilde{\boldsymbol{H}}, P_{\mathrm{Tx}}\right). \tag{4}$$

Here, maximization is performed over the choice of the scalar γ and the rate vector $\boldsymbol{r} = [\ R_1 \ \cdots \ R_K\ ]^{\mathrm{T}}$, which is constrained to belong to the capacity region denoted by $\mathcal{C}\left(\tilde{\boldsymbol{H}}, P_{\mathrm{Tx}}\right)$ and to lie on the straight line defined by $\boldsymbol{\rho}$ and the origin. Obviously, the solution to the rate balancing problem is given by the point on the boundary of the capacity region that lies on this straight line.

From a practical point of view this problem formulation is important since it incorporates system requirements, represented by the QoS constraint $\boldsymbol{\rho}$, into the design of the transmission strategy. This is in contrast with a pure sum rate maximizing strategy [15]–[17] that completely adapts transmission to the channel, possibly switching off some of the users for the sake of total throughput. Contrary to weighted sum rate maximization [13], rate balancing determines the share of throughput finally achieved by each user.

B. Iterative Algorithms

Iterative algorithms for the solution of (4) can be found by exploiting Lagrangian duality. Due to convexity of the capacity region [9], [18], (4) is convex. Furthermore, as rate vectors in the interior of $\mathcal{C}\left(\tilde{\boldsymbol{H}}, P_{\mathrm{Tx}}\right)$ can always be found that lie on the straight line defined by the QoS constraint, strong duality holds [19]. As a consequence, (4) can be alternatively solved by solving the dual problem. The dual objective function can be written as

$$g(\boldsymbol{\lambda}) = \max_{\boldsymbol{r},\gamma} \gamma \left(1 - \sum_{k=1}^{K} \lambda_k\right) + \sum_{k=1}^{K} \lambda_k \frac{R_k}{\rho_k},$$

with $\boldsymbol{r} \in \mathcal{C}\left(\tilde{\boldsymbol{H}}, P_{\mathrm{Tx}}\right)$ and $\boldsymbol{\lambda} = [\ \lambda_1 \ \cdots \ \lambda_K\ ]^{\mathrm{T}}$. This function is equal to ∞ unless $\sum_{k=1}^{K} \lambda_k = 1$, therefore, the corresponding dual problem can be stated as

$$\min_{\boldsymbol{\lambda}} \max_{\boldsymbol{r}} \sum_{k=1}^{K} \lambda_k \frac{R_k}{\rho_k}, \tag{5}$$

subject to $\sum_{k=1}^{K} \lambda_k = 1$. Problem (5) can be rewritten as

$$\min_{\tilde{\boldsymbol{\lambda}}} \max_{\boldsymbol{r}} \frac{R_K}{\rho_K} + \sum_{k=1}^{K-1} \lambda_k \left(\frac{R_k}{\rho_k} - \frac{R_K}{\rho_K}\right),$$

where the constraint has been incorporated into the objective function and $\tilde{\boldsymbol{\lambda}} = [\ \lambda_1 \ \cdots \ \lambda_{K-1}\ ]^{\mathrm{T}}$. Let $\boldsymbol{r}^*_{\tilde{\boldsymbol{\lambda}}} = [\ R^*_{\tilde{\boldsymbol{\lambda}},1} \ \cdots \ R^*_{\tilde{\boldsymbol{\lambda}},K}\ ]^{\mathrm{T}}$ be the optimum rate vector for

$$g(\tilde{\boldsymbol{\lambda}}) = \max_{\boldsymbol{r}} \frac{R_K}{\rho_K} + \sum_{k=1}^{K-1} \lambda_k \left(\frac{R_k}{\rho_k} - \frac{R_K}{\rho_K}\right).$$

It can be easily observed that $\tilde{\boldsymbol{r}}_{\tilde{\boldsymbol{\lambda}}} = [\ \tilde{R}_1 \ \cdots \ \tilde{R}_{K-1}\ ]^{\mathrm{T}}$ with

$$\tilde{R}_k = \frac{R^*_{\tilde{\boldsymbol{\lambda}},k}}{\rho_k} - \frac{R^*_{\tilde{\boldsymbol{\lambda}},K}}{\rho_K}$$

is a subgradient of $g(\tilde{\boldsymbol{\lambda}})$ at $\tilde{\boldsymbol{\lambda}}$, i.e., $g(\tilde{\boldsymbol{\lambda}} + \boldsymbol{\Delta}\tilde{\boldsymbol{\lambda}}) \geq g(\tilde{\boldsymbol{\lambda}}) + \tilde{\boldsymbol{r}}^{\mathrm{T}}_{\tilde{\boldsymbol{\lambda}}} \boldsymbol{\Delta}\tilde{\boldsymbol{\lambda}}$. Hence, in order to minimize $g(\tilde{\boldsymbol{\lambda}})$ a subgradient approach can be followed moving in each iteration a step in the direction $\boldsymbol{\Delta}\tilde{\boldsymbol{\lambda}} = -\boldsymbol{A}\tilde{\boldsymbol{r}}_{\tilde{\boldsymbol{\lambda}}}$ with $\boldsymbol{A} \geq 0$. This is essentially the approach chosen in [10] to solve the power minimization problem. Alternatively, this subgradient can be used as an oracle that, at each iteration, allows the computation of a new ellipsoid of smaller volume than the previous one but still containing the minimizing $\tilde{\boldsymbol{\lambda}}$. This is the approach followed in [11]. In any case, at each iteration, it is necessary to solve the weighted sum rate problem corresponding to the maximization step in (5) in order to compute the subgradient.

C. Weighted Sum Rate Optimization

In this and the next sections we turn our attention to the weighted sum rate maximization problem. Although this is a nonconvex problem in the BC, the following equivalent problem in the dual MAC can be stated which turns out to be convex provided that the decoding order π is optimally chosen [20],

$$\max_{\pi, \{\tilde{\boldsymbol{Q}}_k\}_{k=1,\cdots,K}} \sum_{k=1}^{K} \mu_k R_k, \tag{6}$$

$$\text{s. t.} \quad \frac{1}{N} \sum_{k=1}^{K} \mathrm{Tr}\left\{\tilde{\boldsymbol{Q}}_k\right\} \leq P_{\mathrm{Tx}},\ \tilde{\boldsymbol{Q}}_k \geq 0,\ \forall k.$$

In the dual MAC the achievable rates are given by

$$R_{\pi(k)} = \log \frac{\left|\boldsymbol{I}_{tN} + \sum_{i=k}^{K} \tilde{\boldsymbol{H}}^{\mathrm{H}}_{\pi(i)} \tilde{\boldsymbol{Q}}_{\pi(i)} \tilde{\boldsymbol{H}}_{\pi(i)}\right|}{\left|\boldsymbol{I}_{tN} + \sum_{i=k+1}^{K} \tilde{\boldsymbol{H}}^{\mathrm{H}}_{\pi(i)} \tilde{\boldsymbol{Q}}_{\pi(i)} \tilde{\boldsymbol{H}}_{\pi(i)}\right|} \quad \forall k, \tag{7}$$

where $\pi(i)$ denotes the user whose information is decoded in ith place. Optimization is performed over the set $\{\tilde{\boldsymbol{Q}}_k\}_{k=1,\cdots,K}$ of transmit covariance matrices and the decoding order π. The optimum π is such that users with a certain priority μ_k are decoded before users with higher priority and later than users with lower priority. The optimum MAC covariance matrices can be converted into optimum BC covariance matrices by means of the transformations given in [9].

In the following, for the sake of notational simplicity we shall assume that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_K$, i.e. π corresponds to a reverse map. Under this assumption, substitution of (7) in (6) yields,

$$\max_{\{\tilde{\boldsymbol{Q}}_k\}_{k=1,\ldots,K}} \sum_{k=1}^{K} \eta_k \log \left|\boldsymbol{I}_{tN} + \sum_{i=1}^{k} \tilde{\boldsymbol{H}}^{\mathrm{H}}_i \tilde{\boldsymbol{Q}}_i \tilde{\boldsymbol{H}}_i\right|, \tag{8}$$

$$\text{s. t.} \quad \frac{1}{N} \sum_{k=1}^{K} \mathrm{Tr}\{\tilde{\boldsymbol{Q}}_k\} \leq P_{\mathrm{Tx}},\ \tilde{\boldsymbol{Q}}_k \geq \boldsymbol{0}\ \forall k,$$

where $\eta_k = \mu_k - \mu_{k+1}$ and $\mu_{K+1} = 0$. In the Appendix it is shown that, optimally, the covariance matrices $\tilde{\boldsymbol{Q}}_k$ have a block diagonal structure matching the structure of their respective channels $\tilde{\boldsymbol{H}}_k^{\mathrm{H}}$, i.e., $\tilde{\boldsymbol{Q}}_k = \mathrm{diag}\,[\,\boldsymbol{Q}_{1,k} \quad \cdots \quad \boldsymbol{Q}_{N,k}\,] \in \mathbb{C}^{r_k N \times r_k N}$. This result dramatically reduces the dimensionality of the input space, thereby decreasing the complexity required by standard convex optimization methods to solve this problem. Consequently, the above optimization problem can be rewritten as

$$\max_{\substack{\{\boldsymbol{Q}_{n,k}\}_{n=1,\ldots,N} \\ k=1,\ldots,K}} \sum_{\substack{k=1 \\ n=1}}^{K,N} \eta_k \log \left| \boldsymbol{I}_t + \sum_{i=1}^{k} \boldsymbol{H}_{n,i}^{\mathrm{H}} \boldsymbol{Q}_{n,i} \boldsymbol{H}_{n,i} \right|, \quad (9)$$

subject to $\frac{1}{N}\sum_{k=1,n=1}^{K,N} \mathrm{Tr}\{\boldsymbol{Q}_{n,k}\} \leq P_{\mathrm{Tx}}$ and $\boldsymbol{Q}_{n,k} \geq \boldsymbol{0}\ \forall n,k$. In order to compute optimum covariance matrices corresponding to boundary points of the capacity region, an algorithm has been proposed in [13] that at each step improves the choice of covariance matrices by searching on the line defined by the eigenvector associated to the largest eigenvalue of the gradients of the objective function. This algorithm can readily be applied to the multicarrier formulation of the problem given by (9) as follows. The gradient obtained by deriving the objective function with respect to any covariance matrix $\boldsymbol{Q}_{n,k}$ can be written as,

$$\boldsymbol{G}_{n,k} = \sum_{j=k}^{K} \eta_j \boldsymbol{H}_{n,k} \left(\boldsymbol{I}_t + \sum_{i=1}^{j} \boldsymbol{H}_{n,i}^{\mathrm{H}} \boldsymbol{Q}_{n,i} \boldsymbol{H}_{n,i} \right)^{-1} \boldsymbol{H}_{n,k}^{\mathrm{H}}.$$

Let $\lambda_{n,k}^{\ell}$ denote the principal eigenvalue of the gradient matrix $\boldsymbol{G}_{n,k}^{\ell}$ obtained in the ℓth iteration. Then, similar to [13], we consider the one-dimensional subspace defined by the unit norm eigenvector $\boldsymbol{v}_{n',k'}^{\ell}$ associated with the maximum principal eigenvalue $\lambda_{n',k'}^{\ell} = \max_{n,k}\{\lambda_{n,k}^{\ell}\}_{\substack{n=1,\ldots,N \\ k=1,\ldots,K}}$ in order to search for an improved set of covariance matrices. Accordingly, the new set of covariance matrices are computed as

$$\boldsymbol{Q}_{n,k}^{\ell+1} = \xi \boldsymbol{Q}_{n,k}^{\ell} + (1-\xi) N P_{\mathrm{Tx}} \boldsymbol{v}_{n',k'}^{\ell} \boldsymbol{v}_{n',k'}^{\ell,\mathrm{H}} \delta_{n,n'} \delta_{k,k'}, \quad (10)$$

where $0 \leq \xi \leq 1$, and $\delta_{s,s'} = 1$ if $s = s'$ and $\delta_{s,s'} = 0$ otherwise. As indicated in [13], the optimum value of ξ along this segment can be found applying bisection. Henceforth, this algorithm will be referred to as line search (LS) algorithm.

Although, theoretically, the LS algorithm converges to the optimum, in practice, the number of iterations required to achieve convergence appears to increase linearly with the number of subcarriers. This is related to the fact that the step size per iteration disminishes as the number of subcarriers increases, i.e., ξ in (10) approaches 1. This is, in turn, a consequence of the fact that, at each iteration, only the structure of the covariance matrix of one user on one subcarrier is updated. As a result, the algorithm becomes very inefficient and eventually impracticable if applied to systems with a large number of subcarriers.

D. Divide and Conquer

In order to speed up computation of covariance matrices, we propose to divide problem (9) into a number of smaller problems. To this end, for each subcarrier, we factorize $\boldsymbol{Q}_{n,k} = p_n \bar{\boldsymbol{Q}}_{n,k}$ such that $\sum_{k=1}^{K} \mathrm{Tr}\{\bar{\boldsymbol{Q}}_{n,k}\} \leq 1$ and $\sum_{n=1}^{N} p_n \leq N P_{\mathrm{Tx}}$. Taking this factorization into account, optimum covariance matrices are found iterating the following two steps.

First, for given $\boldsymbol{p} = [\,p_1 \quad \cdots \quad p_N\,]^{\mathrm{T}}$, solve

$$\max_{\{\bar{\boldsymbol{Q}}_{n,k}\}_{k=1,\ldots,K}} \sum_{k=1}^{K} \eta_k \log \left| \boldsymbol{I}_t + p_n \sum_{i=1}^{k} \boldsymbol{H}_{n,i}^{\mathrm{H}} \bar{\boldsymbol{Q}}_{n,i} \boldsymbol{H}_{n,i} \right|, \quad (11)$$

subject to $\sum_{k=1}^{K} \mathrm{Tr}\{\bar{\boldsymbol{Q}}_{n,k}\} \leq 1$ and $\bar{\boldsymbol{Q}}_{n,k} \geq \boldsymbol{0}\ \forall k$, for every n.

Second, for a given set $\{\bar{\boldsymbol{Q}}_{n,k}\}_{\substack{n=1,\ldots,N \\ k=1,\ldots,K}}$, solve

$$\max_{\boldsymbol{p}} \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_k \log \left| \boldsymbol{I}_t + p_n \sum_{i=1}^{k} \boldsymbol{H}_{n,i}^{\mathrm{H}} \bar{\boldsymbol{Q}}_{n,i} \boldsymbol{H}_{n,i} \right|, \quad (12)$$

subject to $\sum_{n=1}^{N} p_n \leq N P_{\mathrm{Tx}}$ and $p_n \geq 0$.

Both problems are convex. In the second step, an optimum power allocation over subcarriers $\boldsymbol{p}$ is found for a given set of normalized covariance matrices. In the first, given the optimum power allocation $\boldsymbol{p}$ obtained in the previous iteration, an optimum set of normalized covariance matrices is found for every subcarrier. It is clear that each step improves the value of the objective function in (9) and hence convergence is guaranteed.

In the first step, optimization of normalized covariance matrices can be done applying the algorithm presented in [13]. In the second step, the Karush-Kuhn-Tucker (KKT) conditions of the optimization problem [19] yield the following set of equations,

$$\sum_{k=1}^{K} \eta_k \,\mathrm{Tr}\left\{ \left(\boldsymbol{I}_t + p_n \boldsymbol{A}_{n,k}\right)^{-1} \boldsymbol{A}_{n,k} \right\} - \nu + \xi_n = 0\ \forall n, \quad (13)$$

$$N P_{\mathrm{Tx}} - \sum_{n=1}^{N} p_n \geq 0,\ \nu \geq 0,\ p_n \geq 0,\ \xi_n \geq 0\ \forall n,$$

$$\nu \left(N P_{\mathrm{Tx}} - \sum_{n=1}^{N} p_n \right) = 0,\ \xi_n\ p_n = 0\ \forall n,$$

where $\boldsymbol{A}_{n,k} = \sum_{i=1}^{k} \boldsymbol{H}_{n,i}^{\mathrm{H}} \bar{\boldsymbol{Q}}_{n,i} \boldsymbol{H}_{n,i}$. Considering the eigenvalues $\{\lambda_{n,k}^{s}\}_{s=1,\cdots,t}$ of matrix $\boldsymbol{A}_{n,k}$, (13) can be rewritten as

$$\sum_{k=1}^{K} \sum_{s=1}^{t} \frac{\eta_k \lambda_{n,k}^{s}}{1 + p_n \lambda_{n,k}^{s}} - \nu + \xi_n = 0\ \forall n.$$

An efficient algorithm can be implemented that computes the power allocation $\boldsymbol{p}$ satisfying these conditions based on the following two observations.

Observation 1: For a given ν, $p_n \neq 0$ if and only if $\sum_{k=1}^{K} \sum_{s=1}^{t} \eta_k \lambda_{n,k}^{s} > \nu$. In that case, $\xi_n = 0$ and

$$\sum_{k=1}^{K} \sum_{s=1}^{t} \frac{\eta_k \lambda_{n,k}^{s}}{1 + p_n \lambda_{n,k}^{s}} - \nu \quad (14)$$

is a monotonically decreasing function of the transmit power p_n.

Observation 2: The optimum ν is a monotonically decreasing function of the transmit power P_{Tx}. In addition,

$$\nu < \max_{n} \left\{ \sum_{k=1}^{K} \sum_{s=1}^{t} \eta_k \lambda_{n,k}^{s} \right\},$$

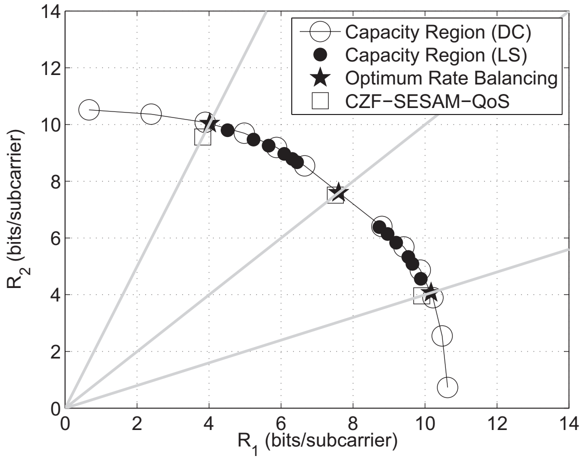

Fig. 1. Capacity region and rate balancing points for a BC with $K = 2$, $t = 4$, $r_k = 2$ and $N = 16$.

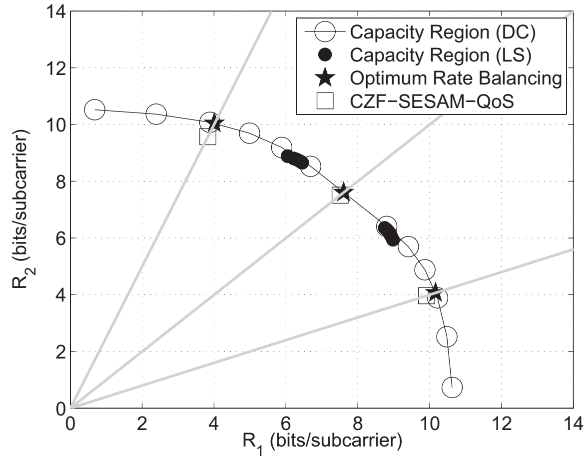

Fig. 2. Capacity region and rate balancing points for a BC with $K = 2$, $t = 4$, $r_k = 2$ and $N = 64$.

i.e., at least one subcarrier gets some power.

From observation 1 it becomes clear that for a given ν there is a unique power allocation $\boldsymbol{p}$ which can be efficiently computed. On the other hand, according to observation 2, if this power allocation exceeds the available transmit power, ν should be increased, otherwise it should be decreased. In this way, bisection can be used in order to compute ν corresponding to the particular transmit power constraint.

The number of iterations per tone required by the divide and conquer (DC) algorithm to reach convergence is essentially independent of the total number of subcarriers. Indeed, the number of subcarriers does not appear in the formulation of problem (11). Thus, for a given power allocation vector, computation of the optimum normalized covariance matrices is independent of this parameter. In turn, for fixed normalized covariance matrices the power allocation on a certain subcarrier n is computed by equating (14) to zero and solving for p_n. We also note that this computation does not essentially depend on the number of subcarriers. Figs. 1 and 2 show the boundary of the capacity regions of two randomly chosen BCs with $N = 16$ and $N = 64$, respectively. Each circle represents a point on the boundary of the capacity region corresponding to a particular choice of priorities, $\mu_1 = 0.1m$, $m \in \{0, 1, \ldots, 10\}$, and $\mu_2 = 1 - \mu_1$. These points have been computed with the DC algorithm presented in this section. As starting point a uniform power allocation over the frequency and scaled identity covariance matrices have been chosen. Table I shows the number of iterations that the DC and the LS algorithms need in order to reach some of these points. For the DC algorithm, outer iterations means the number of times that the power allocation problem (12) needs to be solved. Inner iterations refers to the average number of iterations needed to solve (11) accumulated over the total number of outer iterations. This number is equivalent to the average number of gradient computations per user and subcarrier that is required to reach the final solution. It can be observed that these numbers are almost invariant with respect to the number of subcarriers. By contrast, the number of iterations needed by the LS algorithm in order to reach the same performance as the DC algorithm is observed to increase by approximately factor 4 when passing from $N = 16$ to $N = 64$ subcarriers. In the LS algorithm the number of iterations are equivalent to the number of gradient computations performed per user and subcarrier. Limiting this number to that required by the DC algorithm, the rate vectors achieved by the LS algorithm are represented by the black dots in Figs. 1 and 2. For an approximately constant number of operations per subcarrier, we observe that points computed by the LS algorithm tend to accumulate around the starting point for increasing number of subcarriers.[1]

IV. Non-Iterative Algorithm

Despite the significant complexity reduction obtained by introducing the DC algorithm in the previous section, the optimum approach to the rate balancing problem remains certainly involved. On the one hand subgradient based methods, in general, and the ellipsoid method, in particular, are known to converge very slowly. Furthermore, convergence speed of the LS algorithm in the inner loop of the DC algorithm has been observed to be very sensitive to the number of transmit antennas [21]. A further practical drawback of the optimum approach consists of the existence of solutions that are only achievable by means of time-sharing. This is the case of the optimum points corresponding to the QoS constraint in the middle in Figs. 1 and 2. These rate vectors are only achievable by switching between transmission strategies yielding points at the border of the time-sharing segment. For K users, switching between K different transmission strategies might be required, which increases signaling overhead and the time needed to effectively realize nearly error-free transmission at the desired rates.

In this section a non-iterative algorithm is introduced that requires a complexity similar to that involved in one inner iteration of the DC algorithm and very closely approaches the optimum solution. Furthermore, any solution can be readily implemented without resorting to time-sharing. In the following, this algorithm will be referred to as CZF-SESAM-QoS.

[1]Note that the initial covariance matrices in these examples yield a solution close to both ends of the time-sharing segment.

TABLE I
NUMBER OF ITERATIONS NEEDED BY DC AND LS ALGORITHMS FOR CONVERGENCE WITH $N = 16$ ($N = 64$).

μ_1	0	0.2	0.4	0.6	0.8	1
inner iterations (DC)	38.4 (38.4)	30.0 (29.7)	15.1 (15.7)	10.1 (10.0)	23.6 (23.4)	38.6 (37.6)
outer iterations (DC)	3 (3)	9 (9)	3 (3)	2 (2)	7 (7)	3 (2)
iterations (LS)	895 (3665)	965 (3852)	607 (2986)	415 (2755)	604 (2525)	897 (3602)

Figs. 1 and 2 show the solutions achieved by CZF-SESAM-QoS for the three particular QoS constraints depicted in these examples.

A. CZF-SESAM

CZF-SESAM-QoS is based on the CZF-SESAM algorithm that was first introduced in [14]. On each subcarrier, this algorithm decomposes the broadcast channel into a set of virtually decoupled scalar subchannels by performing a successive allocation of spatial dimensions to users. Here, a subchannel is characterized by a unit norm transmit weighting vector (beamformer), and a unit norm receive weighting vector. The procedure is recalled in Table II. In order to allocate the jth spatial dimension, first, channel matrices are projected into a subspace that is orthogonal to the subspace spanned by beamformers of previously established subchannels. In the second step, a singular value decomposition of the projected channel matrices is performed. Every pair of right and left singular vectors represents a subchannel that may be allocated to the respective user in the third step of the algorithm. Due to the projection in the first step, it is guaranteed that signals sent over any of these spatial subchannels do not interfere with previously assigned subchannels. In the third step, one subchannel, i.e., a pair of right and left singular vectors, is selected according to a given selection rule $\mathcal{O}_n$ and is assigned to the corresponding user. In step 4, the projector is updated by removing the spatial dimension allocated in step 3. These four steps are repeated until no more spatial dimensions are available. Note that the number of available spatial dimensions is upperbounded by $\min\{t, \sum_k r_k\}$. It was shown in [22] that CZF-SESAM practically achieves the Sato bound on sum rate of the broadcast channel. An algorithm for weighted sum rate maximization based on CZF-SESAM was presented in [23], where it was shown that CZF-SESAM can achieve a large fraction of the capacity region. Further aspects of the CZF-SESAM algorithm have been discussed in [24], [25] and [22].

TABLE II
CZF-SESAM ALGORITHM

```
initialization :   j = 1,  T_{n,1} = I_t  ∀n
repeat :
   1.  H^j_{n,k} = H_{n,k} T_{n,j}   ∀n, k
   2.  H^j_{n,k} = U^j_{n,k} Λ^j_{n,k} V^{j,H}_{n,k}   ∀n, k
   3.  (k_0, s_0) = O_n{λ^{s,j}_{n,k}},
       v^j_n = V^j_{n,k_0} e_{s_0},  u^j_n = U^j_{n,k_0} e_{s_0}  ∀n
   4.  T_{n,j+1} = T_{n,j} − v^j_n v^{j,H}_n   ∀n,  j = j + 1
until   j > Σ_{k=1}^{K} r_k  or  T_{n,j} = 0 ∀n
```

B. CZF-SESAM-QoS

Incorporation of QoS constraints into CZF-SESAM uniquely affects step 3 of this algorithm, i.e., the selection rule $\mathcal{O}_n$, where index n indicates that this selection rule might be frequency dependent. Aiming at the allocation of the jth spatial component on every subcarrier, the first and second steps of the CZF-SESAM algorithm are independently executed on all frequency dimensions. Let $\boldsymbol{\Lambda}^j_{n,k}$ be the matrix of singular values of user k, on subcarrier n, in the jth execution of the repeat loop (jth layer), and let $\lambda^{s,j}_{n,k}$ be the sth eigenvalue of this matrix. In the following a method is described for the selection of the jth spatial subchannel on every subcarrier. In order to determine the particular subchannel to be allocated on each subcarrier this method considers the set of all the singular values $\{\lambda^{s,j}_{n,k}\}$ and the given QoS constraint. The procedure consists of three basic steps.

1) Selection of Largest Singular Values: First, for each user, the largest singular value on each subcarrier is selected, i.e.,

$$\lambda^j_{n,k} = \max_s \{\lambda^{s,j}_{n,k}\} \quad \forall n, k,$$

and only these subchannels are considered in the following steps.

2) Determination of Spectral Shares: Second, the number of frequency components is determined that shall be assigned to each user taking into account a given QoS constraint. To this end, first, the capacity is computed that each user could achieve in this layer should all frequency components be assigned to that user. As an example, capacity of user k is computed as

$$C^j_k = \frac{1}{N}\sum_{n=1}^{N} \log\left(1 + p_{n,k}(\lambda^j_{n,k})^2\right),$$

where N is the number of subcarriers and $p_{n,k}$ is obtained from a waterfilling power allocation over the singular values $\lambda^j_{n,k}$. In order to compute capacities at layer j, it is assumed that the average power is limited to P_{Tx}/j. This is merely a heuristic that permits computation of capacity in a particular layer without considering subchannels assigned in previous or subsequent layers. The reason for the division by j is that channel gains become smaller in each layer and so does the power finally allocated to each layer.

Now, we consider the plane defined by the rate vectors $\boldsymbol{R}_k = C^j_k \boldsymbol{e}_k$, $\forall k$, and compute the intersection point of this plane and the straight line defined by the given QoS constraint $\boldsymbol{\rho}$. The equation of the plane is given by $\boldsymbol{R} = \sum_{k=1}^{K} \beta_k \boldsymbol{R}_k$ with $\sum_{k=1}^{K} \beta_k = 1$, and that of the straight line by $\boldsymbol{R} = \gamma\boldsymbol{\rho}$. The intersection point is obtained solving the following linear system of equations,

$$\begin{aligned}\gamma\boldsymbol{\rho} &= \beta_1\boldsymbol{R}_1+\beta_2\boldsymbol{R}_2+\ldots+\beta_K\boldsymbol{R}_K\\ 1 &= \beta_1+\beta_2+\ldots+\beta_K.\end{aligned}$$

The resulting weight β_k is interpreted as the fraction of subcarriers that should be allocated to user k at layer j to comply with QoS constraint $\boldsymbol{\rho}$. Correspondingly, the number of subcarriers assigned to that user in that layer is given by $N_k = \beta_k N$, which can be rounded and readjusted to obtain integers adding up to the total number of subcarriers. This procedure and interpretation of the parameters β_k is optimum if there is only one spatial dimension, e.g., $t = 1$, the channels are non frequency selective, each subcarrier is exclusively assigned to a unique user and the same amount of power is allocated on every subcarrier. Only in such a case the plane represents the boundary of the set of achievable rates and the intersection of this plane with the QoS constraint is the optimum operational point. Even though in all other cases this method is suboptimum, it shall be seen that it delivers excellent results.

3) Effective Subchannel Allocation: In the third step, allocation of subcarriers to users is performed such that compliance with the subcarrier numbers obtained in the previous step is guaranteed. To this end, first, on each subcarrier the subchannel is selected with largest gain, i.e.,

$$\lambda_n^j = \max_k \left\{\lambda_{n,k}^j\right\}\ \forall n. \tag{15}$$

This selection is optimum with respect to sum capacity but it might not be in agreement with the numbers of subcarriers computed in the previous section. If this is the case the selection must be modified in order to match these numbers. This can be done as follows.

Let $\tilde{N}_k$ be the number of subchannels of user k selected according to (15) and define the following sets: The set of users to which additional subchannels should be assigned, $\mathcal{R} = \{k|N_k - \tilde{N}_k > 0\}$, the set of users from which subchannels should be removed, $\mathcal{D} = \{k|N_k - \tilde{N}_k < 0\}$, the set of subcarriers on which subchannel selection could be modified, i.e., subcarriers on which a user of set $\mathcal{D}$ has been assigned a subchannel, $\mathcal{C} = \{n|\lambda_n^j = \lambda_{n,k}^j,\ k \in \mathcal{D}\}$ and, finally, a set with the differences between gains of selected subchannels and gains of non-selected subchannels, $\mathcal{S} = \{\Delta\lambda_{n,k}|k \in \mathcal{R},\ n \in \mathcal{C}\}$, where $\Delta\lambda_{n,k} = \lambda_n^j - \lambda_{n,k}^j$. With these definitions the following procedure is repeated until the sets $\mathcal{D}$ and $\mathcal{R}$ are empty, i.e., until the number of subcarriers assigned to each user coincides with the number N_k computed in the previous section.

First, find the user of set $\mathcal{R}$ and carrier of set $\mathcal{C}$ corresponding to the smallest gain difference with respect to a selected subchannel,

$$(k', n') = \underset{n,k}{\operatorname{argmin}} \left\{\Delta\lambda_{n,k}\right\}, \Delta\lambda_{n,k} \in \mathcal{S}.$$

Then, find the user to which initially the subchannel on subcarrier n' has been assigned,

$$k'' = \underset{k}{\operatorname{argmax}}\{\lambda_{n',k}^j\}.$$

Next, change the assignment on the selected subcarrier, i.e. $\lambda_{n'}^j = \lambda_{n',k'}^j$. Finally, update subchannel counters, $\tilde{N}_{k''} = \tilde{N}_{k''}-1$, $\tilde{N}_{k'} = \tilde{N}_{k'}+1$, and redefine sets accordingly. Though suboptimal, this procedure yields a good performance and has a clear rationale. It departs from the sum capacity optimum subchannel selection and modifies at each step the allocation so that the incurred channel gain loss is minimized.

Once allocation at layer j has been completed, projectors are correspondingly updated on each subcarrier (see step 4 in Table II) and allocation of the $(j+1)$th spatial dimension starts.

C. Waterfilling With QoS Constraint

After the allocation process has been concluded, for each user, a set of scalar mutually decoupled subchannels is obtained over which power loading can be applied so as to maximize sum rate under consideration of the given QoS constraint. A suboptimum algorithm for this problem has been previously proposed in [6]. An optimum algorithm is derived in this section. Let $g_{k,\ell}$ represent the channel gain of the ℓth subchannel assigned to user k and L_k the total amount of subchannels assigned to that user. The optimization problem to be solved in order to find the power loading that maximizes sum rate subject to a QoS constraint $\boldsymbol{\rho}$ can be stated as follows,

$$\max_{\{\boldsymbol{p}_k\}_{k=1,\ldots,K}} \frac{1}{\rho_1}\sum_{\ell=1}^{L_1}\log(1+p_{1,\ell}g_{1,\ell}^2),$$

subject to

$$\frac{1}{\rho_k}\sum_{\ell=1}^{L_k}\log(1+p_{k,\ell}g_{k,\ell}^2)-\frac{1}{\rho_1}\sum_{\ell=1}^{L_1}\log(1+p_{1,\ell}g_{1,\ell}^2) = 0,\ \forall k>1,$$

and $p_{k,\ell} \geq 0\ \forall k,\ell$, $NP_{\mathrm{Tx}} - \sum_{k=1}^K\sum_{\ell=1}^{L_k} p_{k,\ell} \geq 0$, where $\boldsymbol{p}_k = [\ p_{k,1}\ \ \ldots\ \ p_{k,L_k}\ ]^{\mathrm{T}}$ and $p_{k,\ell}$ is the power allocated on the ℓth subchannel of user k. The Lagrangian of this optimization problem can be written as

$$\begin{aligned}&\mathcal{L}\left(\{p_{k,\ell}\}_{\substack{k=1,\ldots,K\\ \ell=1,\ldots,L_k}}, \eta, \{\mu_{k,\ell}\}_{\substack{k=1,\ldots,K\\ \ell=1,\ldots,L_k}}, \{\nu_k\}_{k=1,\ldots,K}\right) =\\ &= \sum_{k=1}^K \frac{\nu_k}{\rho_k}\sum_{\ell=1}^{L_k}\log(1+p_{k,\ell}g_{k,\ell}^2)+\\ &+\eta\left(NP_{\mathrm{Tx}} - \sum_{k=1}^K\sum_{\ell=1}^{L_k}p_{k,\ell}\right)+\sum_{k=1}^K\sum_{\ell=1}^{L_k}\mu_{k,\ell}p_{k,\ell},\end{aligned}$$

where $\nu_1 = 1 - \sum_{k=2}^K \nu_k$. The corresponding relevant KKT conditions read

$$\frac{\nu_k}{\rho_k}\frac{g_{k,\ell}^2}{1+p_{k,\ell}g_{k,\ell}^2} - \eta + \mu_{k,\ell} = 0,\ \forall k, \tag{16}$$

$$\begin{aligned}&\frac{1}{\rho_k}\sum_{\ell=1}^{L_k}\log(1+p_{k,\ell}g_{k,\ell}^2)-\\ &-\frac{1}{\rho_1}\sum_{\ell=1}^{L_1}\log(1+p_{1,\ell}g_{1,\ell}^2) = 0,\ \forall k>1,\end{aligned} \tag{17}$$

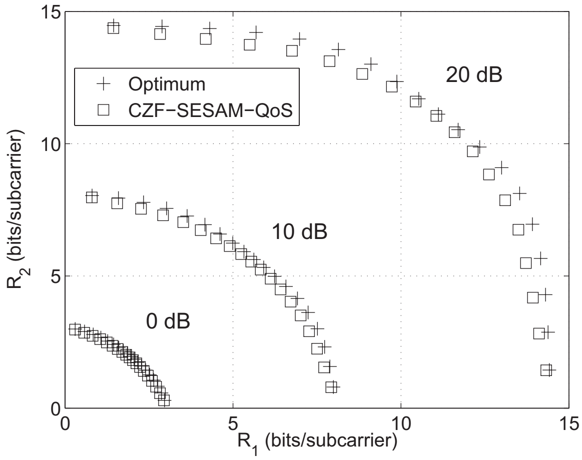

Fig. 3. Average optimum and suboptimum rate balancing points for a balanced BC with $K = 2$, $t = 4$, $r_k = 2$ and $N = 16$.

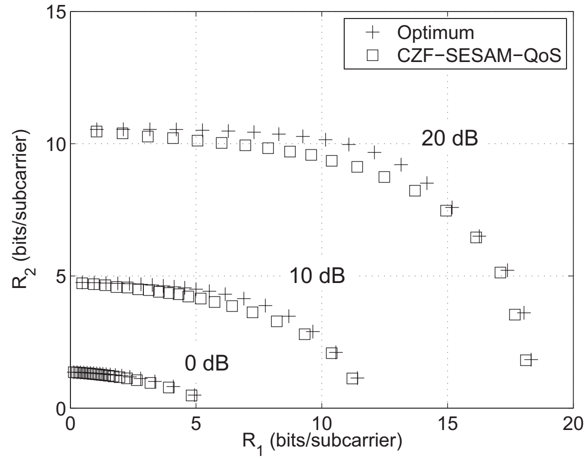

Fig. 4. Average optimum and suboptimum rate balancing points for an unbalanced BC with $K = 2$, $t = 4$, $r_k = 2$ and $N = 16$.

$$NP_{\mathrm{Tx}} - \sum_{k=1}^{K}\sum_{\ell=1}^{L_k} p_{k,\ell} \geq 0,\ p_{k,\ell} \geq 0,\ \forall k,\ell, \tag{18}$$

$$\eta \geq 0,\ \mu_{k,\ell} \geq 0,\ \forall k,\ell,$$

$$\eta\left(NP_{\mathrm{Tx}} - \sum_{k=1}^{K}\sum_{\ell=1}^{L_k} p_{k,\ell}\right) = 0, \quad \mu_{k,\ell}p_{k,\ell} = 0,\ \forall k,\ell.$$

Considering $p_{k,\ell} \geq 0$, $\mu_{k,\ell}p_{k,\ell} = 0$ and (16) we can write

$$p_{k,\ell} = \left[\xi_k - \frac{1}{g_{k,\ell}^2}\right]^+ \ \forall k,\ell, \tag{19}$$

where $\xi_k = \nu_k/\rho_k\eta$. This result has the form of a waterfilling solution with a user dependent water level ξ_k. These levels are optimally chosen so that the $K-1$ equalities in (17) hold and (18) is satisfied with equality. A bisection procedure can be used to determine these parameters in the following way. First, the water level of a certain user is arbitrarily chosen, e.g., ξ_1. The water levels of all other users are then determined so that equations (17) are satisfied. The total power is subsequently computed using (19). If the resulting total power is smaller than the available transmit power NP_{Tx} the level ξ_1 is increased. Otherwise, the level ξ_1 is decreased. These computations are repeated until the required power is approximately equal to the available power. Note that the optimization problem is non convex. However, in this case the KKT conditions are sufficient as there is only one allocation that fulfils them.

V. Numerical Results

In order to avoid misleading conclusions from particular examples, average results over a number of different channel realizations are required. To this end, simulations have been carried out of a system with $N = 16$, $K = 2$, $t = 4$ and $r_k = 2$. Entries of channel matrices on different subcarriers have been assumed to be independently and identically distributed according to a zero-mean complex-valued Gaussian distribution. Fig. 3 shows average results for a balanced BC channel where the entries of the channel matrices of both users are assumed to have unit variance. QoS constraints have been considered such that $R_1/R_2 = 0.1n$ and $R_2/R_1 = 0.1n$ with $1 \leq n \leq 10$, $n \in \mathbb{N}$. Fig. 4 shows results for the case of an unbalaced BC channel, where the variance of the entries of the channel matrices of user 1 has been set to 4 and the variance of the entries of the channel matrices of user 2 has been decreased to 1/4. In both cases rate vectors have been plotted for three different SNR values, which is here defined as $\mathrm{SNR} = P_{\mathrm{Tx}}/\sigma^2$, being σ^2 the variance of the noise at each receive antenna. It can be observed that CZF-SESAM-QoS almost achieves the performance of the optimum solution in both plots. This is specially true for the range of points achieving the maximum sum rate as well as for points close to the axes. For points in between some rate loss can be noticed. However, in any case this loss is observed to be below 7% of the optimum rate per user.

Fig. 5 shows average rate per user obtained in a BC with $N = 16$, $t = 4$, $r_k = 2$ and a "maximum" fairness constraint, i.e., $\rho_1 = \rho_2 = \cdots = \rho_K$, for $K = 2$, $K = 5$ and $K = 10$ users. As above, all entries of channel matrices in every subcarrier have been independently drawn from a zero-mean complex-valued Gaussian distribution with unit variance. CZF-SESAM-QoS practically achieves the performance of the optimum solution for 2 and 5 users. However, for the case of 10 users the gap between the optimum solution and CZF-SESAM-QoS is noticeable. The reason for that might be the high number of users per subcarrier in the system. As the number of users per subcarrier increases, the optimum solution tends to split the users in groups that are served in separate OFDM symbols as part of a time-sharing strategy. By contrast, CZF-SESAM-QoS tries to comply with the QoS constraint in each single OFDM symbol. This strategy becomes increasingly inefficient for growing number of users. In Table III average numbers are given concerning computation and implementation of the optimum solution in Fig. 5. As stop condition for the ellipsoid method we require that the maximum radius of the ellipsoid at a certain iteration become smaller that $\epsilon = 0.01$ or, alternatively, that

$$\max_k \left| \frac{R_k^\ell}{\rho_k} - \frac{1}{K}\sum_{i=1}^{K}\frac{R_i^\ell}{\rho_i} \right| \leq \epsilon,$$

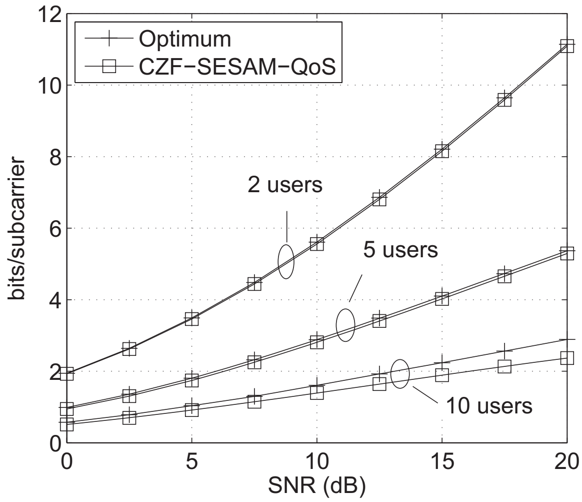

Fig. 5. Average optimum and suboptimum rates per user with a "maximum" fairness QoS constraint. $N = 16$, $t = 4$ and $r_k = 2$.

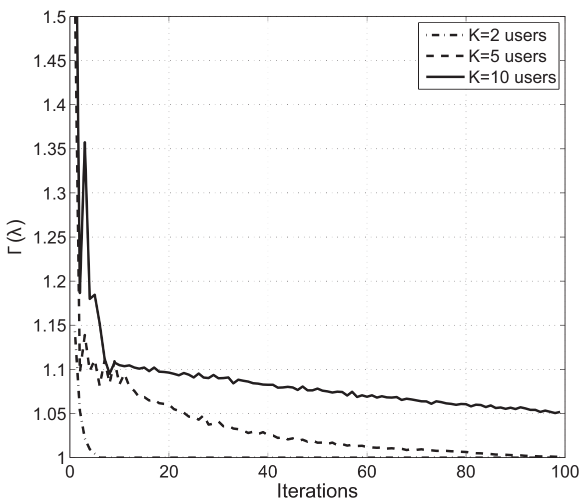

Fig. 6. Convergence of the rate balancing optimum iterative algorithm based on the ellipsoid method. Averaged curves over 100 channel realizations. $N = 16$, $t = 4$, $r_k = 2$, $\mathrm{SNR} = 10$ dB. $\Gamma(\boldsymbol{\lambda}) = g(\boldsymbol{\lambda})/\gamma^{\mathrm{opt}}$.

i.e., the vector of rates obtained at a certain iteration ℓ is almost parallel to the constraint vector $\boldsymbol{\rho}$. As stop conditions for (11) and (12), we require the increment in the value of the respective objective function at a certain iteration to be smaller than 0.1% and 1% of the value achieved in the previous iteration, respectively. Specially significant is the degradation in convergence speed of the ellipsoid method as the number of users increases. This can also be observed in Fig. 6, where the convergence behavior of the ellipsoid algorithm is shown for $\mathrm{SNR} = 10$ dB. As already discussed in the previous section, specially troublesome for practical implementation is the number of time-sharing corner points between which switching is required to actually achieve the optimum rates. In Table III we observe that, for a maximum fairness QoS constraint, the average number of necessary time-sharing points approaches the actual number of users.

VI. Conclusions

The rate balancing problem and details around the computation of the optimum solution have been discussed in a MIMO OFDM context. In particular, an efficient algorithm has been proposed to solve the weighted sum rate problem in a multicarrier setting, which constitutes the most costly step of the optimum rate balancing approach. The main merit of the new algorithm is that the number of iterations per subcarrier involved in the search of the optimum rate vector appears to be independent of the total number of subcarriers in the system. This is in contrast with a direct application of the state-of-the-art solution to an OFDM setting. Furthermore, a suboptimum non-iterative algorithm has been presented that nearly reaches the optimum solution while showing important advantages concerning computational complexity as well as implementation of the resulting transmission strategy.

Appendix
Optimality of Block Diagonal Covariance Matrices

Let $\{\mathbf{a}_n\}_{n=1,\dots,N}$ be a set of random vectors with $\mathbf{a}_n \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{A}_n)$, $\forall n$. In addition, let $\mathbf{a} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{A})$ be the random vector defined as $\mathbf{a} = [\ \mathbf{a}_1^{\mathrm{T}} \ \ \cdots \ \ \mathbf{a}_n^{\mathrm{T}}\ ]^{\mathrm{T}}$. It holds

$$\begin{aligned} n \log \pi e + \log |\boldsymbol{A}| &= h(\mathbf{a}_1, \dots, \mathbf{a}_N) \\ &= \sum_{n=1}^{N} h(\mathbf{a}_n | \mathbf{a}_1, \dots, \mathbf{a}_{n-1}) \\ &\leq \sum_{n=1}^{N} h(\mathbf{a}_n) \\ &= n \log \pi e + \log \prod_{n=1}^{N} |\boldsymbol{A}_n|, \end{aligned} \tag{20}$$

where the inequality follows from the fact that conditioning reduces entropy. Let $\{\tilde{\boldsymbol{Q}}_k^o\}_{k=1,\dots,K}$ be the set of covariance matrices achieving the optimum in (8), and let $\{\tilde{\boldsymbol{Q}}_k^b\}_{k=1,\dots,K}$ be a set of block diagonal matrices obtained out of the optimum matrices by setting the off-diagonal elements to zero. Certainly, the set $\{\tilde{\boldsymbol{Q}}_k^b\}_{k=1,\dots,K}$ satisfies the constraints of (8). Moreover, using (20) we can write

$$\begin{aligned} &\log \left| \boldsymbol{I}_{tN} + \sum_{i=1}^{k} \tilde{\boldsymbol{H}}_i^{\mathrm{H}} \tilde{\boldsymbol{Q}}_i^o \tilde{\boldsymbol{H}}_i \right| \\ &\qquad \leq \log \left| \boldsymbol{I}_{tN} + \sum_{i=1}^{k} \tilde{\boldsymbol{H}}_i^{\mathrm{H}} \tilde{\boldsymbol{Q}}_i^b \tilde{\boldsymbol{H}}_i \right|, \forall k. \end{aligned}$$

This contradicts the initial assumption of $\{\tilde{\boldsymbol{Q}}_k^o\}_{k=1,\dots,K}$ being optimum unless these matrices are all block diagonal.

References

[1] K. Tachikawa, "A perspective on the evolution of mobile communications," *IEEE Commun. Mag.*, pp. 66–73, Oct. 2003.

[2] H. Sampath, S. Talwar, J. Tellado, V. Erceg, and A. Paulraj, "A fourth-generation MIMO-OFDM broadband wireless system: design, performance, and field trial results," *IEEE Commun. Mag.*, pp. 143–149, Sept. 2002.

[3] W. Rhee and J. M. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," in *Proc. Veh. Technol. Conf.*, May 2000, pp. 1085–1089.

TABLE III
AVERAGE NUMBERS INVOLVED IN THE COMPUTATION AND IMPLEMENTATION OF THE OPTIMUM RATE BALANCING WITH FAIRNESS QOS CONSTRAINT FOR $N = 16$, $t = 4$, $r_k = 2$ AND $K = 2/5/10$.

SNR (dB)	0	5	10	15	20
inner iterations (DC)	11.3/17.8/19.9	9.9/19.2/21.7	7.0/17.8/20.3	4.0/15.3/18.0	2.5/13.2/16.0
iterations (ellipsoid method)	6.4/85.7/440.6	6.8/85.1/437.4	7.0/84.3/422.2	7.0/83.8/418.4	7.0/83.3/411.6
time-sharing corner points	1.4/3.1/7.9	1.6/4.4/7.9	1.9/4.5/8.1	2.0/4.7/8.4	2.0/4.7/8.8

[4] I. C. Wong, Z. Shen, B. L. Evans, and J. G. Andrews, “A low complexity algorithm for proportional resource allocation,” in *Proc. IEEE Workshop Signal Processing Systems*, 2004.

[5] C. Mohanram and S. Bhashyam, “A sub-optimal joint subcarrier power allocation algorithm for multiuser OFDM,” *IEEE Commun. Lett.*, vol. 9, pp. 685–687, 2005.

[6] Z. Shen, J. G. Andrews, and B. L. Evans, “Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints,” *IEEE Trans. Wireless Commun.*, vol. 4, pp. 2726–2737, 2005.

[7] E. Jorswieck and H. Boche, “Rate balancing for the multi-antenna gaussian broadcast channel,” in *Proc. IEEE 7th Int. Symp. Spread-Spectrum Tech. Appl.*, Sept. 2002.

[8] M. Schubert and H. Boche, “Iterative multiuser uplink and downlink beamforming under SINR constraints,” *IEEE Trans. Signal Processing*, vol. 53, pp. 2324–2333, 2005.

[9] 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.

[10] W. Fung, C.-H. F. Yu, and T. J. Lim, “Multi-antenna downlink precoding with individual rate constraints: power minimization and user ordering,” in *Proc. Conf. Information Sciences Systems*, 2004.

[11] J. Lee and N. Jindal, “Symmetric capacity of MIMO downlink channels,” in *Proc. IEEE Internatinal Symposium Inform. Theory (ISIT)*, 2006.

[12] A. Nemirovski, “Advances in convex optimization: conic programming,” plenary lecture at the International Congress of Mathematicians, Aug. 2006.

[13] H. Viswanathan, S. Venkatesan, and H. Huang, “Downlink capacity evaluation of cellular networks with known-interference cancellation,” *IEEE J. Select. Areas Commun.*, vol. 21, pp. 802–811, June 2003.

[14] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek, “A novel decomposition technique for multiuser MIMO,” in *Proc. IEEE/ITG Workshop Smart Antennas*, Apr. 2005.

[15] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. J. Goldsmith, “Sum power iterative water-filling for multi-antenna gaussian broadcast channels,” *IEEE Trans. Inform. Theory*, vol. 51, pp. 1570–1580, 2005.

[16] W. Yu, “Sum-capacity computation for the gaussian vector broadcast channel via dual decomposition,” *IEEE Trans. Inform. Theory*, vol. 52, pp. 754–759, 2006.

[17] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek, “Subchannel allocation in multiuser multiple input multiple output systems,” *IEEE Trans. Inform. Theory*, vol. 52, pp. 4721–4732, Oct. 2006.

[18] H. Weingarten, Y. Steinberg, and S. Shamai, “The capacity region of the gaussian multiple-input multiple-output broadcast channel,” *IEEE Trans. Inform. Theory*, vol. 52, pp. 3936–3964, 2006.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambrige University Press, 2004.

[20] A. J. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, “Capacity limits of MIMO channels,” *IEEE J. Select. Areas Commun.*, vol. 21, pp. 684–702, June 2003.

[21] 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, pp. 1640–1646, 2006.

[22] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek, “Analysis of the impact of channel estimation errors on the decomposition of multiuser MIMO channels,” in *Proc. IEEE International Conf. Acoustics, Speech Signal Processing*, 2006.

[23] J. Brehmer, A. Molin, P. Tejera, and W. Utschick, “Low complexity approximation of the MIMO broadcast channel capacity region,” in *Proc. IEEE International Conf. Commun.*, 2006.

[24] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek, “Sum-rate maximizing decomposition approaches for multiuser MIMO-OFDM,” in *Proc. IEEE International Symposium Personal, Indoor Mobile Radio Commun.*, Berlin, Sept. 2005.

[25] ——, “Efficient implementation of successive encoding schemes for the MIMO OFDM broadcast channel,” in *Proc. IEEE International Conf. Commun.*, 2006.

Pedro Tejera received the telecommunication engineering degree from the Universidad Politécnica de Madrid (UPM) in 1999. In 1999 he worked as a network manager for the mobile communications provider Viag Interkom GmbH in Munich, Germany. From 2000 to 2002 he worked at the Einrichtung für Systeme der Kommunikationstechnik (Fraunhofer Gesellschaft) as a researcher in the field of digital subscriber lines. In 2002 he joined the Fachgebiet Methoden der Signalverarbeitung at the Technische Universität München (TUM), where his main topic of research was transmission schemes for MIMO-OFDM wireless systems. In 2008 he received the Dr.-Ing. degree in electrical engineering from the Technische Universität München (TUM) and joined Celestrius, a spin-off company of ETH Zürich, where he co-designed a IEEE 802.11n transceiver. Currently, he is working as a functional manager responsible for the integration of receiver technology in automotive infotainment systems at Alpine Electronics Europe GmbH in Munich. Dr. Tejera is a recipient of the literature award 2007 of the German Information Technology Society (ITG).

Wolfgang Utschick completed several industrial education programs before he received the diploma ('93) and doctoral degrees ('98) both with honors in electrical engineering from the Technische Universität München (TUM). In this period he held a scholarship of the Bavarian Government for exceptional students. From '98 - '02 he co-directed the Signal Processing Group at the Institute of Circuit Theory and Signal Processing at the TUM, and from '00 to '02 he was consulting on 3 GPP standardization in the field of multi-element antenna systems. He was a visiting researcher at the ETH Zürich in 2000 and guest professor at the University of Edinburgh in 2005. In 2002 he was appointed Professor at the TUM where he is head of the Fachgebiet Methoden der Signalverarbeitung. His research interests include signal processing, MIMO systems, stochastic processes and optimization theory in the field of wireless communications. Dr. Utschick is senior member of the VDE and of the IEEE where he is currently instrumental as an associate editor for the TRANSACTIONS ON CIRCUITS AND SYSTEMS - PART I. In 2006 Dr. Utschick was awarded for his excellent teaching records at TUM and in 2007 he received the Literature Award of the German Information Technology Society (ITG).

Josef A. Nossek (S'72-M'74-SM'81-F'93) received the Dipl.-Ing. and the Dr. techn. degrees in electrical engineering from the University of Technology in Vienna, Austria in 1974 and 1980, respectively. In 1974 he joined Siemens AG in Munich, Germany as a member of the technical staff, in 1978 he became supervisor, and from 1980 on he was the head of the department. In 1987 he was promoted head of all radio systems design. Since 1989 he has been a Full Professor for circuit theory and signal processing at the Munich University of Technology where he teaches undergraduate and graduate courses on circuit and systems theory and signal processing and leads research on signal processing algorithms for communications, especially multiantenna systems. He was President Elect, President and Past President of the IEEE Circuits and Systems Society in 2001, 2002 and 2003 respectively. He was Vicepresident of VDE (Verband der Elektrotechnik, Elektronik und Informationstechnik e.V.) 2005 and 2006 and is President of VDE 2007 and 2008. His awards include the ITG Best Paper Award 1988, the Mannesmann Mobilfunk (now Vodafone) Innovationsaward 1998, the Award for Excellence in Teaching from the Bavarian Ministry for Science, Research and Art in 1998. From the IEEE Circuits and Systems Society he received the Golden Jubilee Medal for 'Outstanding Contributions to the Society' in 1999 and the Education Award in 2008.

Gerhard Bauch received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from Munich University of Technology in 1995 and 2001, respectively, and the Diplom-Volkswirt degree from FernUniversitaet Hagen in 2001. In 1996, he was with the German Aerospace Center (DLR), Oberpfaffenhofen, Germany. From 1996-2001 he was a member of the scientific staff at Munich University of Technology (TUM). In 1998 and 1999 he was a visiting researcher at AT&T Labs Research, Florham Park, NJ, USA. In 2002 he joined DOCOMO Euro-Labs, Munich, Germany, where he is currently manager of the Advanced Radio Transmission Group. In 2007 he was also appointed Research Fellow at DOCOMO Euro-Labs. Since October 2003 he has also been an adjunct professor at Munich University of Technology. Furthermore, he lectured as a guest professor at the University of Udine, Italy, and the Alpen-Adria University Klagenfurt, Austria. He received the best paper award of the European Personal Mobile Communications Conference (EPMCC) 1997, the Texas Instruments Award 2001 of TUM, the Award of the German Information Technology Society (ITG in VDE) 2002 (ITG Foerderpreis) and the Literature Award 2007 of the German Information Technology Society (ITG in VDE). He has (co-)authored a textbook *Contemporary Communications Systems* as well as more than 100 scientific papers in major journals and international conferences. His research interests include channel coding and modulation, turbo processing, multihop transmission, multiple access and various aspects of signal processing in multi-antenna systems (MIMO).