Nonconcave Utility Maximization in the MIMO Broadcast Channel

Johannes Brehmer and Wolfgang Utschick
Associate Institute for Signal Processing
Technische Universität München
{brehmer, utschick}@tum.de

***Abstract*—The problem of determining an optimal parameter setup at the physical layer in multi-user, multi-antenna downlink is considered. An aggregate utility, which is assumed to depend on the users' rates, is used as performance metric. It is not assumed that the utility function is concave, allowing for more realistic utility models of applications with limited scalability. Due to the structure of the underlying capacity region, a two step approach is necessary: First, an optimal rate vector is determined. Second, the optimal parameter setup is derived from the optimal rate vector. Two methods for computing an optimal rate vector are proposed: First, based on the differential manifold structure offered by the boundary of the MIMO BC capacity region, a gradient projection method on the boundary is developed. Being a local algorithm, the method converges to a rate vector which is not guaranteed to be a globally optimal solution. Second, the monotonic structure of the rate space problem is exploited to compute a globally optimal rate vector with an outer approximation algorithm. While the second method yields the global optimum, the first method is shown to provide an attractive trade-off between utility performance and computational complexity.**

I. Introduction

The majority of current wireless communication systems is based on the principle of *orthogonal multiple access*. Simply speaking, multiple users compete for a set of shared channels, and access to the channels is coordinated such that each channel is used by a single user only. The decision which user accesses which channel is made at the *medium access* (MAC) layer, with the result that at the *physical* (PHY) layer, transmission is over single-user channels. Based on recent advances in physical layer techniques such as MIMO signal processing and multi-user coding, it has been shown that significant performance gains can be achieved by allowing one channel to be used by multiple users at once [1], [2], [3], [4], [5]. In other words, the physical layer paradigm is shifting from single-user channels to multi-user channels. This change also dissolves the strict distinction between MAC and PHY layers, as the question which users access which channels can only be answered in a joint treatment of both layers.

In this work, a multi-user, multi-antenna downlink in a single-cell wireless system is considered, which, from the viewpoint of information theory, corresponds to a *MIMO broadcast channel* (MIMO BC) [3], [6]. While the aforementioned shift to multi-user channels is motivated by the potential gains in system performance, an evident drawback of this shift is the increased design complexity. In other words, multi-antenna, multi-user channels significantly increase the set of design parameters and degrees of freedom at the PHY layer. Clearly, strategies for tuning these parameters in an optimal manner are of great interest.

The desire for maximum system performance leads immediately to the question of optimality criteria. While voice and best effort data applications have been predominant, future wireless systems are expected to provide a multitude of heterogeneous applications, ranging from best effort data to low-delay gaming applications, from low-rate messaging to high-rate video. The heterogeneity of these applications requires application-aware optimality criteria, i.e., it is no longer sufficient to optimize PHY and MAC layer with respect to criteria such as average throughput or proportional rate fairness.

Utility functions have been widely used as a model for the properties of upper layers. In this work, the focus is on the optimization of the PHY layer parameters, and a generic utility model in terms of a function that is monotone in the user's rates is employed. For a wide range of applications, utility models can be found in the literature. In [7], applications are classified based on their elasticity with respect to the allocated rate. Best effort applications can be modeled with a concave utility [7]. On the other hand, less elastic applications result in a nonconcave utility model [7], [8]. While most works on

utility maximization in wireless systems assume concave utilities, the nonconcave setup has received relatively little attention [8], [9], [10]. Based on the premise that some relevant application classes can be more precisely modeled by nonconcave utilities, this work proposes a solution strategy that provides at least locally optimal performance in the nonconcave case.

There exists a significant amount of literature on utility maximization for wireless networks, see, e.g., [11], [12], [13], [10] and references therein. The network-oriented works usually consider a large number of nodes with a simple physical layer setup, and focus on computationally efficient and distributed resource allocation strategies for large networks. In contrast, this work focuses on the optimization of a limited-size infrastructure network with a complex multi-antenna, multi-user PHY/MAC layer configuration. Utility maximization in the MIMO BC is also investigated in [14]. The authors solve the utility maximization problem based on Lagrange duality, under the assumption of concave utility functions. *Dual methods* are frequently used in network utility maximization [10], but rely on the assumption of problem convexity. This work makes the following contributions: First, a primal gradient-based method for addressing the utility maximization problem in the MIMO BC is developed. The proposed method does not rely on a convexity assumption and can provide convergence to local optima in the nonconvex case. The quality of such local solutions depends on the specific problem instance and can only be evaluated if the global optimum is known. The second contribution of this work is the application of methods from the field of deterministic global optimization to the nonconcave utility maximization problem. It is shown that the utility maximization problem in the MIMO BC can be cast as a monotonic optimization problem [15]. The monotonicity structure can be exploited to efficiently find the global optimum by an outer approximation algorithm.

Notation: Vectors and vector-valued functions are denoted by bold lowercase letters, matrices by bold uppercase letters. The transpose and the Hermitian transpose of $\boldsymbol{Q}$ are denoted by $\boldsymbol{Q}^{\mathrm{T}}$ and $\boldsymbol{Q}^{\mathrm{H}}$, respectively. The identity matrix is denoted by $\mathbf{1}$. Concerning boldface, an exception is made for gradients: The gradient of a function u evaluated at $\boldsymbol{x}$ is a vector $\nabla u(\boldsymbol{x})$, the gradient of a function $\boldsymbol{f}$ evaluated at $\boldsymbol{x}$ is a matrix $\nabla \boldsymbol{f}(\boldsymbol{x})$ whose i-th column is the gradient at $\boldsymbol{x}$ of the i-th component function of $\boldsymbol{f}$ [16]. The following definitions of order relations between vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^K$, with $K > 1$, are used:

$$\begin{aligned} \boldsymbol{x} \geq \boldsymbol{y} &\Leftrightarrow \forall k : x_k \geq y_k, \\ \boldsymbol{x} > \boldsymbol{y} &\Leftrightarrow \boldsymbol{x} \geq \boldsymbol{y}, \exists k : x_k > y_k, \\ \boldsymbol{x} \gg \boldsymbol{y} &\Leftrightarrow \forall k : x_k > y_k. \end{aligned}$$

Order relations $\leq, <, \ll$ are defined in the same manner.

II. Problem Setup

At the physical layer, a MIMO broadcast channel with K receivers is considered. The transmitter has N transmit antennas, while receiver k is equipped with M_k receive antennas. The transmitter sends independent information to each of the receivers.

The received signal at receiver k is given by

$$\boldsymbol{y}_k = \boldsymbol{H}_k \sum_{i=1}^{K} \boldsymbol{x}_i + \boldsymbol{\eta}_k,$$

where $\boldsymbol{H}_k \in \mathbb{C}^{M_k \times N}$ is the channel to receiver k and $\boldsymbol{x}_k \in \mathbb{C}^N$ is the signal transmitted to receiver k. Furthermore, $\boldsymbol{\eta}_k$ is the circularly symmetric complex Gaussian noise at receiver k, with $\boldsymbol{\eta}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{1}_{M_k})$.

Let $\boldsymbol{Q}_k$ denote the transmit covariance matrix of user k. The total transmit power has to satisfy the power constraint $\mathrm{tr}\left(\sum_{k=1}^{K} \boldsymbol{Q}_k\right) \leq P_{\mathrm{tr}}$. Accordingly, with $\boldsymbol{Q} = (\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_K)$ the set of feasible transmit covariance matrices is given by

$$\mathcal{Q} = \left\{ \boldsymbol{Q} : \boldsymbol{Q}_k \in \mathbb{H}_+^N, \mathrm{tr}\left(\sum_{k=1}^{K} \boldsymbol{Q}_k\right) \leq P_{\mathrm{tr}} \right\}.$$

where $\mathbb{H}_+^N$ denotes the set of positive semidefinite Hermitian $N \times N$ matrices.

As proved in [6], capacity is achieved by *dirty paper coding* (DPC). Let π denote the encoding order, i.e., $\pi : \{1, \ldots, K\} \rightarrow \{1, \ldots, K\}$ is a permutation and $\pi(i)$ is the index of the user which is encoded at the i-th position. Moreover, let $\mathit{\Pi}$ denote the set of all possible permutations on $\{1, \ldots, K\}$.

For fixed $\boldsymbol{Q}$ and π, an achievable rate vector is given by $\boldsymbol{r}(\boldsymbol{Q}, \pi) = (r_1(\boldsymbol{Q}, \pi), \ldots, r_K(\boldsymbol{Q}, \pi))$, with

$$r_{\pi(i)} = \log \frac{\det\left(\mathbf{1} + \boldsymbol{H}_{\pi(i)} \left(\sum_{j \geq i} \boldsymbol{Q}_{\pi(j)}\right) \boldsymbol{H}_{\pi(i)}^{\mathrm{H}}\right)}{\det\left(\mathbf{1} + \boldsymbol{H}_{\pi(i)} \left(\sum_{j > i} \boldsymbol{Q}_{\pi(j)}\right) \boldsymbol{H}_{\pi(i)}^{\mathrm{H}}\right)}.$$

Let $\mathcal{R}$ denote the set of rate vectors achievable by feasible $\boldsymbol{Q}$ and π:

$$\mathcal{R} = \left\{ \boldsymbol{r}(\boldsymbol{Q}, \pi) : \boldsymbol{Q} \in \mathcal{Q}, \pi \in \mathit{\Pi} \right\}.$$

The capacity region of the MIMO BC is defined as the convex hull of $\mathcal{R}$ [3]:

$$\mathcal{C} = \mathrm{co}(\mathcal{R}).$$

Accordingly, each element of $\mathcal{C}$ can be written as a convex combination of elements of $\mathcal{R}$, i.e., for each $\boldsymbol{r} \in \mathcal{C}$, there exists a set of coefficients $\{\alpha_w\}$, a set of transmit covariance matrices $\{\boldsymbol{Q}^{(w)}\}$, and a set of encoding orders $\{\pi^{(w)}\}$ such that

$$\boldsymbol{r} = \sum_{w=1}^{W} \alpha_w \boldsymbol{r}(\boldsymbol{Q}^{(w)}, \pi^{(w)}), \tag{1}$$

with $\alpha_w \geq 0, \sum_{w=1}^{W} \alpha_w = 1, \boldsymbol{Q}^{(w)} \in \mathcal{Q}$, and $\pi^{(w)} \in \mathit{\Pi}$. In other words, $\boldsymbol{r}$ is achieved by time-sharing between rate vectors $\boldsymbol{r}(\boldsymbol{Q}^{(w)}, \pi^{(w)}) \in \mathcal{R}$.

Each $\boldsymbol{r} \in \mathcal{C}$ can be achieved by time-sharing between at most K rate vectors $\boldsymbol{r}(\boldsymbol{Q}^{(w)}, \pi^{(w)}) \in \mathcal{R}$, thus $W \leq K$. Accordingly, the physical layer parameter vector can be defined as follows:

$$\boldsymbol{x}_\mathrm{P} = (\alpha_w, \boldsymbol{Q}^{(w)}, \pi^{(w)})_{w=1}^{K}.$$

Moreover, the set of feasible PHY parameter setups is given by

$$\mathcal{X}_\mathrm{P} = \left\{ \boldsymbol{x}_\mathrm{P} : \alpha_w \geq 0, \sum_{w=1}^{W} \alpha_w = 1, \boldsymbol{Q}^{(w)} \in \mathcal{Q}, \pi^{(w)} \in \mathit{\Pi} \right\}.$$

Given the set $\mathcal{X}_\mathrm{P}$, an obvious problem is finding a parameter setup $\boldsymbol{x}_\mathrm{P}^*$ that is, in a desired sense, optimal.

In this work, it is assumed that the properties of the upper layers are summarized in a system utility function $u : \mathbb{R}_+^K \to \mathbb{R}$, whose value depends only on the rate vector provided by the physical layer. The parameter optimization problem is then given by

$$\max_{\boldsymbol{x}_\mathrm{P}} u(\boldsymbol{r}(\boldsymbol{x}_\mathrm{P})) \quad \text{s.t.} \quad \boldsymbol{x}_\mathrm{P} \in \mathcal{X}_\mathrm{P}, \tag{2}$$

where $\boldsymbol{r}(\boldsymbol{x}_\mathrm{P})$ follows from Eq. (1). Concerning the function u, it is assumed that larger rates result in higher utility, i.e., it is assumed that u is strictly monotonically increasing.[1] Moreover, it is assumed that u is continuous, and differentiable on $\mathbb{R}_{++}^K$. The function u is not assumed to be concave.

[1] Strict monotonicity implies that

$$\boldsymbol{r} > \boldsymbol{r}' \Rightarrow u(\boldsymbol{r}) > u(\boldsymbol{r}'). \tag{3}$$

III. Nonconcave utilities

One of the premises of this work is that nonconcave utilities are of high practical relevance in future communication systems. Consider the case $K = 1$. A strictly monotone function $u : r \mapsto u(r)$ is concave if the gain in utility obtained from increasing r decreases with increasing r, for all $r \in \mathbb{R}_+$. A common example for such a behaviour are best effort data applications, where any increase in rate is good, but a saturation effect leads to a decreasing gain for larger r [7]. Such elastic applications are perfectly scalable. On the other extreme, applications that have fixed rate requirements (such as traditional voice service) are not scalable at all (inelastic), and are more precisely modeled by a nonconcave utility: Below a certain rate threshold, utility is zero, above the threshold utility takes on its maximum value (step function) [7].

Based on recent advances in multimedia coding, future multimedia applications can be expected to lie between these two extremes: They are scalable to some extent, but do not provide the perfect scalability of best effort services. As an example, the scalable video coding extension of the H.264/AVC standard [17] provides support of scalability based on a layered video codec. Due to the finite number of layers, the decoded video's quality only increases at those rates where an additional layer can be transmitted. Moreover, if the gain between layers is not incremental (such as experienced when switching between low and high spatial resolution), such a behaviour can be more precisely modeled by a nonconcave utility, which, in contrast to a concave utility, does not require a steady decrease of the gain over the whole range of feasible rates. To summarize, the flexibility offered by nonconcave utilities allows for more precise models of multimedia applications, which only have a finite number of operation modes and show a non-monotone behaviour of the gains experienced by an increase in rate.

IV. Direct Approach

Based on (2), a first approach may be to directly optimize the composite function $u \circ r$ with respect to the PHY parameters $\boldsymbol{x}_\mathrm{P}$. In general, however, this approach will fail, due to the discrete nature of $\mathit{\Pi}$ and the nonconvexity of problem (2), even for a concave utility function u.

In contrast, the capacity region is convex by definition, thus the problem

$$\max_{\boldsymbol{r}} u(\boldsymbol{r}) \quad \text{s.t.} \quad \boldsymbol{r} \in \mathcal{C} \tag{4}$$

is convex for concave u. This motivates solution approaches that operate in the rate space and not in the physical layer parameter space.

A special case for which the direct approach succeeds is given by the utility $u(\boldsymbol{r}) = \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{r}$, i.e., *weighted sum rate maximization* (WsrMax). In this case, time sharing is not required, i.e., $\alpha_w^* = 0, w > 1$. Moreover, the gradient ∇u is independent of $\boldsymbol{r}$, and an optimal encoding order π^* can be directly inferred from $\boldsymbol{\lambda}$ [18], [3], [4]. As a result, the problem is reduced to finding the optimal transmit covariance matrices, which can be solved as a convex problem in the dual MAC [4]. Denote by $\boldsymbol{r}_{\text{wsr}}(\boldsymbol{\lambda}, \pi^*)$ the rate vector that maximizes weighted sum rate for a given weight $\boldsymbol{\lambda}$ and a corresponding optimal encoding order π^*, i.e.,

$$\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{r}_{\text{wsr}}(\boldsymbol{\lambda}, \pi^*) = \max_{\boldsymbol{Q}\in\mathcal{Q}} \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{Q}, \pi^*). \tag{5}$$

For general utility functions, the optimal solution may require time-sharing. In particular, if no further assumptions concerning the properties of u are made, the loss incurred by approximating a time-sharing solution by a rate vector $\boldsymbol{r} \in \mathcal{R}$ may be significant. Moreover, even if the optimal solution does not require time-sharing, it is not clear how to find the optimal encoding order.

An optimization algorithm operating in the rate space of course still requires a means to compute points from $\mathcal{C}$. WsrMax over $\mathcal{C}$ can be cast as a convex problem. Moreover, efficient algorithms for solving the WsrMax problem in the MIMO BC have been proposed recently [19], [20]. Based on this observation, the proposed algorithm is formulated such that iterates on $\mathcal{C}$ are obtained as solutions of WsrMax problems.

V. Iterative Efficient Set Approximation

To solve problem (2), a two-step procedure is followed: First, determine a (possibly locally) optimal solution $\boldsymbol{r}^*$ of problem (4) by operating in the rate space. Second, given $\boldsymbol{r}^*$, determine a parameter setup $\boldsymbol{x}_{\text{p}}^*$ such that

$$\boldsymbol{r}(\boldsymbol{x}_{\text{p}}^*) = \boldsymbol{r}^*.$$

Due to the assumed strict monotonicity of the function u, all candidate solutions to problem (2) lie on the Pareto efficient boundary of $\mathcal{C}$. The Pareto efficient set is defined as

$$\mathcal{E} = \left\{\boldsymbol{r} \in \mathcal{C} : \nexists \boldsymbol{r}' \in \mathcal{C} : \boldsymbol{r}' > \boldsymbol{r}\right\}. \tag{6}$$

Knowing that $\boldsymbol{r}^* \in \mathcal{E}$, a gradient projection method is proposed that generates iterates on $\mathcal{E}$. Note that there exist different flavors of gradient projection methods: A gradient projection on arbitrary convex sets [16], requiring a Euclidean projection, and a gradient projection on sets equipped with a differential manifold structure [21], [22], [23]. In this work, the second approach is followed.

In the classical gradient projection method of Rosen [24], it is assumed that the feasible set is described by a set of constraint functions $\boldsymbol{h}, \boldsymbol{m}$ such that the set of feasible $\boldsymbol{r}$ is given by $\boldsymbol{h}(\boldsymbol{r}) \leq \boldsymbol{0}, \boldsymbol{m}(\boldsymbol{r}) = \boldsymbol{0}$, with $\boldsymbol{h}, \boldsymbol{m}$ differentiable. For the capacity region of the MIMO BC, such a description in terms of constraint functions in $\boldsymbol{r}$ is not available (basically, all that is available is a method to compute points on its efficient boundary, by means of WsrMax). The key for a gradient-based optimization in the rate space is to recognize the differentiable manifold structure offered by the efficient boundary of the capacity region. By exploiting this structure, a gradient ascent on $\mathcal{E}$ that does not rely on a description in terms of constraint functions is possible.

A. Gradient Ascent on $\mathcal{E}$

The following problem is considered:

$$\max_{\boldsymbol{r}\in\mathcal{E}} u(\boldsymbol{r}). \tag{7}$$

The efficient set $\mathcal{E}$ is a $K-1$ dimensional manifold with boundary [25], where the boundary of $\mathcal{E}$ corresponds to rate vectors $\boldsymbol{r} \in \mathcal{E}$ with at least one user having zero rate. Furthermore, it is assumed that for the MIMO BC, the interior of the efficient set, defined by

$$\tilde{\mathcal{E}} = \left\{\boldsymbol{r} \in \mathcal{E} : \boldsymbol{r} \gg \boldsymbol{0}\right\},$$

is smooth up to first order, i.e., $\tilde{\mathcal{E}}$ is a C^1 differentiable [25], $K-1$ dimensional manifold. Based on this assumption, there exists a set $\{\boldsymbol{\phi}_{\boldsymbol{r}}\}_{\boldsymbol{r}\in\tilde{\mathcal{E}}}$ of differentiable local parameterizations $\boldsymbol{\phi}_{\boldsymbol{r}} : \mathcal{U}_{\boldsymbol{r}} \subset \mathbb{R}^{K-1} \to \tilde{\mathcal{E}}$, with $\mathcal{U}_{\boldsymbol{r}}$ open and $\boldsymbol{\phi}_{\boldsymbol{r}}(\boldsymbol{0}) = \boldsymbol{r}$ [25].

For simplicity, it is first assumed that $\boldsymbol{r}^* \in \tilde{\mathcal{E}}$. Based on this assumption, starting at $\boldsymbol{r}^{(0)}$, a sequence of iterates $\boldsymbol{r}^{(n)} \in \tilde{\mathcal{E}}$ is generated. At each $\boldsymbol{r}^{(n)}$, a parameterization $\boldsymbol{\phi}_{\boldsymbol{r}^{(n)}}$ is available. Composing parameterization and utility function results in a function $f_{\boldsymbol{r}} = u \circ \boldsymbol{\phi}_{\boldsymbol{r}}$, which maps an open subset of $\mathbb{R}^{K-1}$ into $\mathbb{R}$. The composite function $f_{\boldsymbol{r}}$ is amenable to standard methods for unconstrained optimization. Based on this observation, a gradient ascent is carried out on the set of functions $f_{\boldsymbol{r}} = u \circ \boldsymbol{\phi}_{\boldsymbol{r}}$. Let $\boldsymbol{r}^{(n)}$ denote the n-th iterate, and $\boldsymbol{\mu}^{(n)}$ its coordinates in the parameterization $\boldsymbol{\phi}_{\boldsymbol{r}^{(n)}}$, i.e., $\boldsymbol{\mu}^{(n)} = \boldsymbol{\phi}_{\boldsymbol{r}^{(n)}}^{-1}(\boldsymbol{r}^{(n)}) = \boldsymbol{0}$. By definition of $f_{\boldsymbol{r}}$, $u(\boldsymbol{r}) = f_{\boldsymbol{r}}(\boldsymbol{0})$. The composite

function $f_{\boldsymbol{r}}$ is differentiable at $\mathbf{0}$, with gradient $\nabla f_{\boldsymbol{r}}$ at $\mathbf{0}$ given by

$$\nabla f_{\boldsymbol{r}}(\mathbf{0}) = \nabla \boldsymbol{\phi}_{\boldsymbol{r}}(\mathbf{0}) \nabla u(\boldsymbol{r}), \tag{8}$$

where $\nabla \boldsymbol{\phi}_{\boldsymbol{r}}^{\mathrm{T}}$ is the Jacobian of $\boldsymbol{\phi}_{\boldsymbol{r}}$. If $\nabla f_{\boldsymbol{r}}(\mathbf{0}) \neq \mathbf{0}$, then $\nabla f_{\boldsymbol{r}}(\mathbf{0})$ is an ascent direction of $f_{\boldsymbol{r}}$ at $\mathbf{0}$, i.e., there exists a $\beta > 0$ such that for all t, $0 < t \leq \beta$

$$t \nabla f_{\boldsymbol{r}}(\mathbf{0}) \in \mathcal{U}_{\boldsymbol{r}}, \tag{9}$$

$$f_{\boldsymbol{r}}(t \nabla f_{\boldsymbol{r}}(\mathbf{0})) > f_{\boldsymbol{r}}(\mathbf{0}), \tag{10}$$

where (9) follows from the fact that $\mathcal{U}_{\boldsymbol{r}}$ is open and (10) from the differentiability of $f_{\boldsymbol{r}}$, see, e.g., Theorem 2.1 in [26]. This gives rise to the following iteration:

$$\boldsymbol{\mu}^{(n)} = \boldsymbol{\phi}_{\boldsymbol{r}^{(n)}}^{-1}(\boldsymbol{r}^{(n)}) = \mathbf{0}, \tag{11}$$

$$\boldsymbol{\mu}^{(n+1)} = t \nabla f_{\boldsymbol{r}^{(n)}}(\mathbf{0}), \tag{12}$$

$$\boldsymbol{r}^{(n+1)} = \boldsymbol{\phi}_{\boldsymbol{r}^{(n)}}(\boldsymbol{\mu}^{(n+1)}), \tag{13}$$

with $t > 0$ chosen such that properties (9) and (10) are fulfilled. The algorithm defined in Eqs. (11)-(13) is a so-called *varying parameterization approach* to optimization on manifolds [23], [27].

According to Eq. (10), the iterates $\boldsymbol{r}^{(n)}$ generate an increasing sequence $u(\boldsymbol{r}^{(n)})$. The iteration stops if

$$\nabla f_{\boldsymbol{r}}(\mathbf{0}) = \mathbf{0}. \tag{14}$$

In this work, points $\boldsymbol{r} \in \mathcal{E}$ for which (14) holds are denoted as *stationary points*. The tangent space of $\tilde{\mathcal{E}}$ at $\boldsymbol{r}$ is defined as

$$\mathcal{T}_{\boldsymbol{r}} = \operatorname{span}\left(\nabla \boldsymbol{\phi}_{\boldsymbol{r}}(\mathbf{0})^{\mathrm{T}}\right).$$

Thus, geometrically, stationary points correspond to points on the efficient boundary where the gradient of the utility function is orthogonal to the tangent space, cf. Eq. (8). In the context of minimizing a differentiable function over a differentiable manifold, Eq. (14) represents a necessary first-order optimality condition [22].

The step size t is determined with an inexact line search. As evaluations of $f_{\boldsymbol{r}}$ are usually computationally expensive, the step size t is chosen such that an increase in the utility value results, while keeping the number of evaluations of $f_{\boldsymbol{r}}$ as small as possible. Define

$$\theta(t) = f_{\boldsymbol{r}}(t \nabla f_{\boldsymbol{r}}(\mathbf{0})) = u(\boldsymbol{\phi}_{\boldsymbol{r}^{(n)}}(t \nabla f_{\boldsymbol{r}^{(n)}}(\mathbf{0}))).$$

Starting with an initial step size $t = t_0$ that satisfies Eq. (9), the step size t is halved until

$$\theta(t) \geq \theta(0) + \alpha \nabla \theta(0) t, \tag{15}$$

for fixed α, $0 < \alpha < 1$. Note that (15) corresponds to Armijo's rule [28] for accepting a step size as not too large. In contrast to Armijo's rule, however, there is no test whether the step size is too small, i.e., t_0 is always considered large enough.

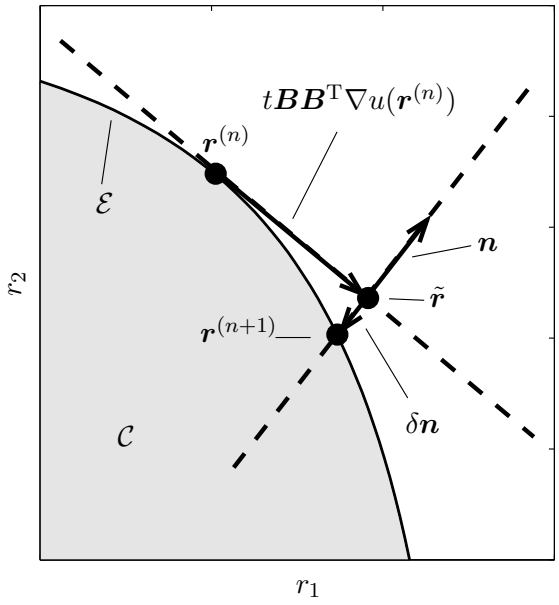

Figure 1. One iteration of the IEA method.

There exists a choice for the parameterizations $\boldsymbol{\phi}_{\boldsymbol{r}}$ for which $\nabla \boldsymbol{\phi}_{\boldsymbol{r}}(\mathbf{0})$, and thus $\nabla f_{\boldsymbol{r}}(\mathbf{0})$, is particularly simple to compute. Let $\boldsymbol{B} \in \mathbb{R}^{K \times K-1}$ denote an orthonormal basis of the tangent space $\mathcal{T}_{\boldsymbol{r}}$. Choose $\boldsymbol{n}$ such that the columns of $\begin{bmatrix} \boldsymbol{B} & \boldsymbol{n} \end{bmatrix}$ constitute an orthonormal basis of $\mathbb{R}^K$. Choose the parameterization $\boldsymbol{\phi}_{\boldsymbol{r}}$ as follows:

$$\boldsymbol{\phi}_{\boldsymbol{r}}(\boldsymbol{\mu}) = \boldsymbol{r} + \boldsymbol{B}\boldsymbol{\mu} + \boldsymbol{n}\delta(\boldsymbol{\mu}), \tag{16}$$

where $\delta(\boldsymbol{\mu})$ is chosen such that $\boldsymbol{\phi}_{\boldsymbol{r}}(\boldsymbol{\mu}) \in \tilde{\mathcal{E}}$ (correction step). Then

$$\nabla \boldsymbol{\phi}_{\boldsymbol{r}}(\mathbf{0}) = \boldsymbol{B}^{\mathrm{T}}. \tag{17}$$

As shown in Subsection V-B, it is straightforward to find a basis $\boldsymbol{B}$. Combining Eqs. (12),(13),(8),(16) and (17) yields

$$\boldsymbol{r}^{(n+1)} = \boldsymbol{r}^{(n)} + t\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} \nabla u(\boldsymbol{r}^{(n)}) + \boldsymbol{n}\delta(t), \tag{18}$$

with $\delta(t) = \delta(t\boldsymbol{B}^{\mathrm{T}} \nabla u(\boldsymbol{r}^{(n)})$. Accordingly, the update in rate space is given by

$$\boldsymbol{r}^{(n+1)} - \boldsymbol{r}^{(n)} = t\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} \nabla u(\boldsymbol{r}^{(n)}) + \boldsymbol{n}\delta(t). \tag{19}$$

The first summand in (19) is the orthogonal projection of $\nabla u(\boldsymbol{r}^{(n)})$ on the tangent space. Based on this observation, the proposed method can be interpreted as follows: First, approximate the efficient set by its tangent space at $\boldsymbol{r}^{(n)}$. Next, compute a gradient step,

using this approximation. Finally, make a correction step from the approximation back to the efficient set, yielding $\boldsymbol{r}^{(n+1)}$. Based on the observation that at each iteration, an approximation of the efficient set is computed, the proposed method is denoted as *iterative efficient set approximation* (IEA). For the case of $K = 2$ users, one iteration of the IEA method is illustrated in Figure 1.

Eq. (9) defines an upper bound on the step size t, which ensures $\boldsymbol{\mu}^{(n+1)}$ stays within the domain of the parameterization $\phi_{\boldsymbol{r}^{(n)}}$. The domain of the parameterization defined in (16) is defined implicitly by the requirement that all entries of the resulting rate vector have to be positive, i.e.,

$$\mathcal{U}_{\boldsymbol{r}} = \{\boldsymbol{\mu} : \boldsymbol{\phi}_{\boldsymbol{r}}(\boldsymbol{\mu}) \gg \mathbf{0}\}. \tag{20}$$

In fact, the image and domain of the parameterization defined in (16) can be extended to also include rate vectors with zero entries. From Eqs. (20) and (18), an upper bound on the step size t can then be derived by interpreting $\boldsymbol{r}^{(n+1)}$ as a function of t. An upper bound on t is given by the value of t where the smallest entry in $\boldsymbol{r}^{(n+1)}(t)$ is exactly zero:

$$\bar{t} : \min_k \boldsymbol{r}_k^{(n+1)}(\bar{t}) = 0.$$

Note that by Eq. (18), the upper bound $\bar{t}$ depends on $\boldsymbol{r}^{(n)}$ – thus the validity range $0 < t < \bar{t}$ changes over $\tilde{\mathcal{E}}$, and it may get small close to the boundary of $\mathcal{E}$.

B. Correction Step

The most involved step is the computation of $\delta(\boldsymbol{\mu}^{(n+1)})$. Write $\boldsymbol{r}^{(n+1)}$ as

$$\boldsymbol{r}^{(n+1)} = \tilde{\boldsymbol{r}} + \delta \boldsymbol{n}, \tag{21}$$

with $\tilde{\boldsymbol{r}} = \boldsymbol{r}^{(n)} + \boldsymbol{B}\boldsymbol{\mu}^{(n+1)}$. Based on (21), the correction step can be interpreted as the projection of $\tilde{\boldsymbol{r}}$ on $\mathcal{E}$ by computing the intersection between $\mathcal{E}$ and the line $\{\boldsymbol{r} = \tilde{\boldsymbol{r}} + x\boldsymbol{n}, x \in \mathbb{R}\}$, cf. Fig. 1. Assume that $\boldsymbol{n} \geq \mathbf{0}$ (the validity of this assumption is verified at the end of this subsection). Then δ can be found by solving the following optimization problem:

$$\delta = \max_{x,\boldsymbol{r}} \; x \quad \text{s.t.} \quad \tilde{\boldsymbol{r}} + x\boldsymbol{n} \leq \boldsymbol{r}, \; \boldsymbol{r} \in \mathcal{C}. \tag{22}$$

Note that (22) is a convex problem. In particular, it is independent of the utility function u, i.e., it is convex regardless whether u is concave or not. Moreover, Slater's condition is satisfied, i.e., strong duality holds. Accordingly, (22) can be solved via Lagrange duality.

The Lagrangian of problem (22) is given by

$$L(x, \boldsymbol{r}, \boldsymbol{\lambda}) = x + \boldsymbol{\lambda}^T(\boldsymbol{r} - \tilde{\boldsymbol{r}} - x\boldsymbol{n}).$$

The dual function follows as

$$\begin{aligned} g(\boldsymbol{\lambda}) &= \sup_{\substack{x\in\mathbb{R} \\ \boldsymbol{r}\in\mathcal{C}}} \left(x(1 - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{n}) + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{r} - \tilde{\boldsymbol{r}})\right) \\ &= \begin{cases} +\infty, \; \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{n} \neq 1, \\ \max_{\boldsymbol{r}\in\mathcal{C}} \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{r} - \tilde{\boldsymbol{r}}), \; \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{n} = 1. \end{cases} \end{aligned} \tag{23}$$

Note that for $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{n} = 1$, again a weighted sum-rate maximization problem is to be solved. Recall from Section IV that WsrMax can be efficiently solved as a convex problem in the dual MAC.

Let $\boldsymbol{r}^*(\boldsymbol{\lambda})$ denote a maximizer of the weighted sum-rate maximization in (23) for a given $\boldsymbol{\lambda} \in \mathbb{R}_+^K$. The optimal dual variable $\boldsymbol{\lambda}$ is found by solving

$$\min_{\boldsymbol{\lambda}\geq\mathbf{0}} \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{r}^*(\boldsymbol{\lambda}) - \tilde{\boldsymbol{r}}) \quad \text{s.t.} \quad \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{n} = 1. \tag{24}$$

According to Danskin's Theorem [16], a subgradient (at $\boldsymbol{\lambda}$) of the cost function of problem (24) is given by $(\boldsymbol{r}^*(\boldsymbol{\lambda}) - \tilde{\boldsymbol{r}})$. If $\boldsymbol{\lambda}$ has equal entries, $\boldsymbol{r}^*(\boldsymbol{\lambda})$ is not unique [4]. Thus, the subgradient is not unique and the cost function is nondifferentiable. Accordingly, the minimization in (24) has to be carried out using any of the methods for nondifferentiable convex optimization, such as subgradient methods, cutting plane methods, or the ellipsoid method [29]. All these methods have in common that they generate iterates $\boldsymbol{\lambda}^{(i)}$ (which converge to the optimal dual variable $\boldsymbol{\lambda}^*$), and at each iteration i, they require the computation of a subgradient at $\boldsymbol{\lambda}^{(i)}$ – which basically corresponds to solving a WsrMax problem with weight $\boldsymbol{\lambda}^{(i)}$. In this work, an outer-linearization cutting plane method [16] is used to solve problem (24).

As strong duality holds, $\delta = g(\boldsymbol{\lambda}^*)$, and

$$\boldsymbol{r}^{(n+1)} = \tilde{\boldsymbol{r}} + g(\boldsymbol{\lambda}^*)\boldsymbol{n}.$$

From the optimal dual variable $\boldsymbol{\lambda}^*$ also follows the tangent space at $\boldsymbol{r}^{(n+1)}$. Due to strong duality, $\boldsymbol{r}^{(n+1)}$ maximizes $L(x^*, \boldsymbol{r}, \boldsymbol{\lambda}^*)$ over $\mathcal{C}$ [16]. Accordingly, $\boldsymbol{r}^{(n+1)}$ is a maximizer of a WsrMax problem with weight $\boldsymbol{\lambda}^*$. Recall that for WsrMax, $u(\boldsymbol{r}) = \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{r}$, with $\nabla u(\boldsymbol{r}) = \boldsymbol{\lambda}$. The corresponding composite function $f_{\boldsymbol{r}}$ is given by $f_{\boldsymbol{r}}(\boldsymbol{\mu}) = \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\phi}_{\boldsymbol{r}}(\boldsymbol{\mu})$. As $\boldsymbol{r}^{(n+1)}$ is a maximizer of the WsrMax problem, it has to be a stationary point (for this particular composite function, with $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$). From Eq. (14) follows:

$$\nabla\left((\boldsymbol{\lambda}^*)^{\mathrm{T}}\boldsymbol{\phi}_{\boldsymbol{r}^{(n+1)}}\right)(\mathbf{0}) = \nabla\boldsymbol{\phi}_{\boldsymbol{r}^{(n+1)}}(\mathbf{0})\boldsymbol{\lambda}^* = \mathbf{0},$$

thus

$$\mathcal{T}_{\boldsymbol{r}^{(n+1)}} = \operatorname{null}\left((\boldsymbol{\lambda}^*)^{\mathrm{T}}\right). \tag{25}$$

In other words, the basis $\boldsymbol{B}$ needed in the *next* iteration can be obtained by computing an orthonormal basis of the null space of $(\boldsymbol{\lambda}^*)^{\mathrm{T}}$, where $\boldsymbol{\lambda}^*$ is the optimal dual variable of the current iteration. In addition, in the next iteration a unit vector $\boldsymbol{n} \geq \mathbf{0}$ orthogonal to $\boldsymbol{B}$ is needed. From Eq. (25) it follows that $\boldsymbol{n}$ (in the next iteration) is simply

$$\boldsymbol{n} = \frac{\boldsymbol{\lambda}^*}{\|\boldsymbol{\lambda}^*\|_2}.$$

C. Time-Sharing Solutions

The algorithm described in Subsections V-A and V-B yields a stationary point $\boldsymbol{r}^*$ of problem (4). The final step is the recovery of an optimal parameter setup $\boldsymbol{x}_{\mathrm{p}}^*$ from $\boldsymbol{r}^*$. The complexity of the recovery step depends on the location of $\boldsymbol{r}^*$: If $\boldsymbol{r}^* \notin \mathcal{R}$, then $\boldsymbol{r}^*$ lies in a time-sharing region. Throughout this work, the term *time-sharing region* denotes a subset of $\mathcal{E}$ whose elements are only achievable by time-sharing. In case of time-sharing optimality, the optimal parameter setup has to be found by identifying a set of points in $\mathcal{E} \cap \mathcal{R}$ whose convex combination yields $\boldsymbol{r}^*$.

The recovery is based on the optimal dual variable of the last correction step: If at least two entries in $\boldsymbol{\lambda}^*$ are equal, time-sharing may be required. In the case of equal entries in $\boldsymbol{\lambda}^*$, there exist multiple rate vectors $\boldsymbol{r} \in \mathcal{R}$ that are maximizers of a WsrMax problem with weight $\boldsymbol{\lambda}^*$ [4], and $\boldsymbol{r}^*$ is a convex combination of these points. In the case that all entries in $\boldsymbol{\lambda}^*$ are equal, all permutations π are optimal, resulting in $K!$ points $\boldsymbol{r}_{\mathrm{wsr}}(\boldsymbol{\lambda}^*, \pi)$. As a consequence, enumerating all $K!$ points first and then selecting the (at most) K points that are actually required to implement $\boldsymbol{r}^*$ is only feasible for small K. For larger K, an efficient method for identifying a set of relevant points is provided in [30].

If no two entries in $\boldsymbol{\lambda}^*$ are equal, the optimum encoding order π^* is uniquely defined, $\boldsymbol{r}^* = \boldsymbol{r}_{\mathrm{wsr}}(\boldsymbol{\lambda}^*, \pi^*)$, and $\boldsymbol{Q}^*$ maximizes $(\boldsymbol{\lambda}^*)^{\mathrm{T}} \boldsymbol{r}(\boldsymbol{Q}, \pi^*)$, cf. Eq. (5).

From an implementation viewpoint, entries in $\boldsymbol{\lambda}^*$ will usually not be exactly equal, even if the theoretical solution lies in a time-sharing region. As a result, time-sharing between users is declared if the difference between weights is below a certain threshold.

D. Coarse Projection

The proposed algorithm consists of two nested loops: a gradient-based outer loop and an inner loop for the correction step at each outer iteration. A significant reduction in computational complexity can be achieved if the required precision of the inner loop is adapted to the outer loop. In fact, the convergence of the outer loop is ensured by an increase in the cost function at each step, based on condition (10). The inner iteration generates rate vectors $\boldsymbol{r}(\boldsymbol{\lambda}^{(i)})$ during convergence to $\boldsymbol{\lambda}^*$. If $\boldsymbol{r}(\boldsymbol{\lambda}^{(i)})$ fulfills condition (10) and $\boldsymbol{r}(\boldsymbol{\lambda}^{(i)}) \in \tilde{\mathcal{E}}$, the projection of $\tilde{\boldsymbol{r}}$ on $\mathcal{C}$ is sufficiently good to yield an ascent step on $\tilde{\mathcal{E}}$. In this case, the projection is aborted and the outer loop continues with

$$\boldsymbol{r}^{(n+1)} = \boldsymbol{r}(\boldsymbol{\lambda}^{(i)}).$$

The resulting reduction in the number of inner iterations comes at the price of an evaluation of the function u at each inner iteration. As a result, the overall gain in terms of complexity clearly depends on the cost associated with evaluating u.

E. Boundary Points

So far, it has been assumed that at the optimal solution $\boldsymbol{r}^*$, all users have non-zero rate (i.e., $\boldsymbol{r}^* \in \tilde{\mathcal{E}}$). If this assumption does not hold, the sequence $\left\{\boldsymbol{r}^{(n)}\right\}$ converges to a point on the boundary of $\mathcal{E}$, cf. Section V-F. Define

$$\mathcal{I}(\boldsymbol{r}) = \{k : r_k = 0\}. \tag{26}$$

The boundary of $\mathcal{E}$ is given by

$$\partial\mathcal{E} = \mathcal{E} \setminus \tilde{\mathcal{E}} = \{\boldsymbol{r} \in \mathcal{E} : \mathcal{I}(\boldsymbol{r}) \neq \varnothing\}.$$

Observe that the boundary can be written as the union of K sets $\partial\mathcal{E}_{\{k\}}$, with

$$\partial\mathcal{E}_{\{k\}} = \{\boldsymbol{r} \in \mathcal{E} : \{k\} \subset \mathcal{I}(\boldsymbol{r})\}.$$

Finally, define a set $\mathcal{E}_{\{k\}}$ by removing the k-th entry (which is zero) from all elements in $\partial\mathcal{E}_{\{k\}}$:

$$\mathcal{E}_{\{k\}} = \left\{\boldsymbol{x} \in \mathbb{R}^{K-1} : x_\ell = r_\ell, \ell \notin \{k\}, \boldsymbol{r} \in \partial\mathcal{E}_{\{k\}}\right\}.$$

Note that the resulting set $\mathcal{E}_{\{k\}}$ is the efficient boundary of a capacity region of a $K-1$ user MIMO BC, with user k removed. It follows immediately that the interior $\mathcal{E}_{\{k\}}$ is again a differentiable manifold, now of dimension $K-1$. The boundary of $\mathcal{E}_{\{k\}}$ can be decomposed in the same manner, resulting in a set of $K-2$ dimensional manifolds, and so on. Accordingly, the set $\mathcal{E}_{\mathcal{D}}$, with $\mathcal{D} \subseteq \{1, \ldots, K\}$ corresponds to the efficient boundary of a capacity region of a $K - |\mathcal{D}|$ user MIMO BC, with users in $\mathcal{D}$ removed.

Accordingly, the general case is incorporated as follows: Denote by $\mathcal{A} = \{1, \ldots, K\} \setminus \mathcal{D}$ the set of active users. Only active users are considered in the optimization, i.e., replace K by $|\mathcal{A}|$ and let k be the index of the k-th active user in all steps of the algorithm. If the sequence $\left\{\boldsymbol{r}^{(n)}\right\}$ converges to a point on the boundary

of $\mathcal{E}_\mathcal{D}$, the users with zero entries in the rate vector are removed from $\mathcal{A}$ and assigned to $\mathcal{D}$. Initialize with $\mathcal{A} = \{1, \ldots, K\}$, $\mathcal{D} = \varnothing$ and $\boldsymbol{r}^{(0)} \in \tilde{\mathcal{E}}$. With these modifications, the algorithm always operates on differentiable manifolds $\tilde{\mathcal{E}}_\mathcal{D} \subset \mathbb{R}^{|\mathcal{A}|}$, with $\boldsymbol{r} \gg \mathbf{0}, \forall \boldsymbol{r} \in \tilde{\mathcal{E}}_\mathcal{D}$.

In practice, convergence to the boundary is detected as follows: If the rate $r_k^{(n)}$ of an active user falls below a threshold, and the projected utility gradient results in $r_k^{(n+1)} < r_k^{(n)}$, the user is deactivated. The decision to deactivate a user is based on the iterates and not on the limit point, thus the modified algorithm may lead to suboptimal results if a user is deactivated that actually has non-zero rate in the limit.

F. Convergence of the IEA Method

Concerning the convergence of the IEA method, two cases can be distinguished: In the first case, the sequence $\left\{\boldsymbol{r}^{(n)}\right\}$ converges to a point in $\tilde{\mathcal{E}}$. In the second case, the sequence $\left\{\boldsymbol{r}^{(n)}\right\}$ converges to a point on the boundary of $\mathcal{E}$. According to Section V-E, after removing the users with zero rate, the boundary itself is a $K-1$ dimensional manifold with boundary, and the algorithm converges in the interior or on the boundary of this manifold. The argument continues until the dimension of the manifold under consideration is 0. Thus, it suffices to consider the convergence behaviour in the interior of $\mathcal{E}_\mathcal{D}$, which, from the perspective of the algorithm is equivalent to $\tilde{\mathcal{E}}$ – an open set equipped with a differentiable manifold structure.

Accordingly, the IEA method is globally convergent if convergence to a point $\boldsymbol{r}^* \in \tilde{\mathcal{E}}$ implies that $\boldsymbol{r}^* \in \tilde{\mathcal{E}}$ is a stationary point. Convergence can be proved using Zangwill's Global Convergence Theorem [26]. Not all parameterizations, however, yield a convergent method. For the parameterization defined in Eq. (16), global convergence (in the sense of the Global Convergence Theorem) is proved in [31].

A more intuitive (and less rigorous) discussion of the convergence behaviour follows from considering the updates $\boldsymbol{\mu}^{(n+1)}$. Convergence to a point $\boldsymbol{r}^*$ implies

$$\boldsymbol{\mu}^{(n+1)} = t^{(n)} \nabla f_{\boldsymbol{r}^{(n)}}(\mathbf{0}) \to 0. \tag{27}$$

Now assume that $\boldsymbol{r}^*$ is not a stationary point. This implies $\nabla f_{\boldsymbol{r}^{(n)}}(\mathbf{0}) \neq \mathbf{0}, \forall n$, which, by Eq. (27) implies $t^{(n)} \to 0$. For the parameterization defined in Eq. (16), such a sequence of step sizes results if the sequence of upper bounds $\bar{t}(\boldsymbol{r}^{(n)})$ converges to zero. This behaviour, however, only occurs if the sequence $\left\{\boldsymbol{r}^{(n)}\right\}$ converges to a point on the boundary of $\mathcal{E}$, which contradicts the assumption that $\boldsymbol{r}^* \in \tilde{\mathcal{E}}$.

The theoretical convergence results based on Zangwill's Global Convergence Theorem assume infinite precision. Theoretically, if $\nabla f_{\boldsymbol{r}^{(n)}}(\mathbf{0}) \neq \mathbf{0}$, it is always possible to find a step size $t > 0$ such that (10) holds. In a practical implementation of the IEA method, the parameterization is evaluated numerically, in particular the correction step is a numerical solution of a convex optimization problem. Due to the convexity of the correction problem, a high numerical precision can be achieved. Still, the inherent finite precision of the correction step sets a limit to the precision of the overall algorithm. This property underlines the importance of the coarse projection described in V-D: The inner loop needs a tight convergence criterion in order to yield a high precision in cases where it is difficult to find an ascent step. In cases where an ascent step is easily found, however, it is not necessary to solve the problem to high precision. The latter case is detected by the coarse projection. Also note that the coarse projection does not impact the convergence behaviour in a negative way: The global convergence ensures that (theoretically) the algorithm does not get stuck at a non-stationary point. The coarse projection only comes into play if it is possible to move away from the current point.

It is clearly not guaranteed that a stationary point $\boldsymbol{r}^*$ maximizes utility. Due to the fact that the proposed algorithm is an ascent method, however, $\boldsymbol{r}^*$ is a good solution in the sense that given an initial value $\boldsymbol{r}^{(0)}$, utility is either improved, or the algorithm converges at the first iteration and stays at $\boldsymbol{r}^{(0)}$, in this case requiring no extra computations. That is, any investment in terms of computational effort is rewarded with a gain in utility.

VI. Monotonic Optimization

The gradient-based approach presented in Section V converges to a stationary point of the optimization problem, and cannot guarantee convergence to global optimality, as it relies on local information only.

The rate-space formulation (4) of the utility maximization problem corresponds to the maximization of a monotonic function (the utility function u) over a compact set in $\mathbb{R}_+^K$ (the capacity region $\mathcal{C}$), and hence is a monotonic optimization problem [15], which can be solved to global optimality.

A basic problem of monotonic optimization is the maximization of a monotonic function over a compact normal set [15]. A subset $\mathcal{S}$ of $\mathbb{R}_+^K$ is said to be *normal* in $\mathbb{R}_+^K$ (or briefly, normal), if $\boldsymbol{x} \in \mathcal{S}, \mathbf{0} \leq \boldsymbol{y} \leq \boldsymbol{x} \Rightarrow \boldsymbol{y} \in \mathcal{S}$. The capacity region $\mathcal{C}$ is normal: any rate vector $\boldsymbol{r}'$ that is smaller than an achievable rate vector $\boldsymbol{r}$ is also

achievable. Thus, $\mathcal{C}$ is a compact normal set and the rate-space problem (4) is a basic problem of monotonic optimization.

A. Polyblock Algorithm

The basic algorithm for solving monotonic optimization problems is the so-called *polyblock algorithm*. A polyblock is simply the union of a finite number of hyper-rectangles in $\mathbb{R}_+^K$: Given a discrete set $\mathcal{V} \subset \mathbb{R}_+^K$, a polyblock $\mathcal{P}(\mathcal{V})$ is defined as

$$\mathcal{P}(\mathcal{V}) = \bigcup_{\boldsymbol{v}\in\mathcal{V}} \{\boldsymbol{r} \in \mathbb{R}_+^K, \boldsymbol{r} \leq \boldsymbol{v}\}.$$

The set $\mathcal{V}$ contains the vertices of the polyblock $\mathcal{P}(\mathcal{V})$.

Due to the fact that $\mathcal{C}$ is a compact normal subset of $\mathbb{R}_+^K$, there exists a set $\mathcal{V}^{(0)}$ such that $\mathcal{C} \subseteq \mathcal{P}(\mathcal{V}^{(0)})$. Moreover, starting with $n = 0$, either $\mathcal{C} = \mathcal{P}(\mathcal{V}^{(n)})$ or there exists a discrete set $\mathcal{V}^{(n+1)} \subset \mathbb{R}_+^K$ such that

$$\mathcal{C} \subseteq \mathcal{P}(\mathcal{V}^{(n+1)}) \subset \mathcal{P}(\mathcal{V}^{(n)}). \tag{28}$$

In other words, the polyblocks $\mathcal{P}(\mathcal{V}^{(n)})$ represent an iteratively refined outer approximation of the capacity region.

Consider the problem of maximizing utility over the polyblock $\mathcal{P}(\mathcal{V}^{(n)})$:

$$\max_{\boldsymbol{r}\in\mathcal{P}(\mathcal{V}^{(n)})} u(\boldsymbol{r}). \tag{29}$$

Let $\check{\boldsymbol{r}}^{(n)}$ denote a maximizer of problem (29). Due to the monotonicity of u, $\check{\boldsymbol{r}}^{(n)} \in \mathcal{V}^{(n)}$, i.e., the maximum of a monotonic function over a polyblock is attained on one of the vertices [15]. Due to the fact that the vertex set of a polyblock is discrete, problem (29) can be solved to global optimality by searching over all $\boldsymbol{v} \in \mathcal{V}^{(n)}$.

If $\check{\boldsymbol{r}}^{(n)} \in \mathcal{E}$, the globally optimal rate vector is found. In general, however, $\check{\boldsymbol{r}}^{(n)}$ will lie outside the capacity region, due to the fact that the polyblock represents an outer approximation. Denote by $\boldsymbol{y}^{(n)} \in \mathcal{E}$ the intersection between $\mathcal{E}$ and the line segment connecting the origin with $\check{\boldsymbol{r}}^{(n)}$. Let $\hat{\boldsymbol{r}}^{(n)}$ denote the best intersection point computed so far, i.e.,

$$\hat{\boldsymbol{r}}^{(n)} = \boldsymbol{y}^{(\ell^*)}, \ell^* = \operatorname*{argmax}_{\ell\in\{1,\ldots,n\}} u(\boldsymbol{y}^{(\ell)}).$$

Moreover, let u^* denote the global maximum of (4). It follows that

$$u(\hat{\boldsymbol{r}}^{(n)}) \leq u^* \leq u(\check{\boldsymbol{r}}^{(n)}). \tag{30}$$

Intuitively, as the outer approximation of $\mathcal{C}$ by a polyblock is refined at each step, $u(\check{\boldsymbol{r}}^{(n)})$ eventually converges to u^*. Due to the continuity of u, this convergence also holds for $\hat{\boldsymbol{r}}^{(n)}$, i.e., $\hat{\boldsymbol{r}}^{(n)}$ converges to a global maximizer of u. See [15] for a rigorous proof. According to Eq. (30), an ϵ-optimal solution is found if $u(\hat{\boldsymbol{r}}^{(n)}) \geq u(\check{\boldsymbol{r}}^{(n)}) - \epsilon$.

One possible method to construct a sequence of polyblocks $\mathcal{P}(\mathcal{V}^{(n)})$ that satisfies (28) is as follows [15]: Define

$$\mathcal{K}(\boldsymbol{r}) = \left\{\boldsymbol{x} \in \mathbb{R}_+^K : x_k > r_k, k \notin \mathcal{I}(\boldsymbol{r})\right\},$$

with $\mathcal{I}(\boldsymbol{r})$ as defined in (26). Clearly, $\hat{\boldsymbol{r}}^{(n)} \in \mathcal{E}$ implies $\mathcal{K}(\hat{\boldsymbol{r}}^{(n)}) \cap \mathcal{C} = \varnothing$. Thus, $\mathcal{K}(\hat{\boldsymbol{r}}^{(n)})$ can be removed from $\mathcal{P}(\mathcal{V}^{(n)})$ without removing any achievable rate vector. Moreover, if $\check{\boldsymbol{r}}^{(n)} \notin \mathcal{E}$,

$$\mathcal{K}(\hat{\boldsymbol{r}}^{(n)}) \cap \mathcal{P}(\mathcal{V}^{(n)}) \supset \left\{\check{\boldsymbol{r}}^{(n)}\right\} \supset \varnothing,$$

thus by removing $\mathcal{K}(\hat{\boldsymbol{r}}^{(n)})$, a tighter approximation results. Finally, $\mathcal{P}(\mathcal{V}^{(n)}) \setminus \mathcal{K}(\hat{\boldsymbol{r}}^{(n)})$ is again a polyblock [15]. To summarize, the desired rule for constructing a sequence of polyblocks that satisfies (28) is

$$\mathcal{P}(\mathcal{V}^{(n+1)}) = \mathcal{P}(\mathcal{V}^{(n)}) \setminus \mathcal{K}(\hat{\boldsymbol{r}}^{(n)}).$$

The rules for computing the corresponding vertex set $\mathcal{V}^{(n+1)}$ are provided in [15].

B. Intersection with $\mathcal{E}$

If the polyblock algorithm is applied to the rate-space problem (4), the only step in the algorithm in which the intricate properties of the capacity region $\mathcal{C}$ come into play is the computation of the intersection between $\mathcal{E}$ and the line connecting the origin with $\check{\boldsymbol{r}}^{(n)}$. Comparing the correction step of the IEA algorithm from Section V-B with the computation of the intersection point, it turns out that both operations are almost identical, only the line whose intersection with $\mathcal{E}$ is computed is different. As a result, the Lagrangian-based algorithm from Section V-B can also be used to compute the intersection point, by setting

$$\tilde{\boldsymbol{r}} = \check{\boldsymbol{r}}^{(n)}, \boldsymbol{n} = \check{\boldsymbol{r}}^{(n)}.$$

In Section V, it was stated that the most complex step in each iteration of the IEA method is the correction step. Similar results hold for the polyblock algorithm: At each iteration, the main complexity lies in the computation of the intersection point. Due to the similarity between IEA's correction step and the computation of the intersection point in the polyblock algorithm, the complexity of both approaches can be compared by comparing the number of gradient iterations with the number of polyblocks generated until a sufficiently tight outer approximation is found. The convergence properties of the polyblock algorithm are only asymptotic [15]

– thus, a relatively high complexity of the polyblock algorithm can be expected. This expectation is confirmed by simulation results, see Section VIII.

C. Implementation Issues

The presentation of the polyblock algorithm in Section VI-A closely follows [15]. In this basic version, simulations showed very slow convergence of the algorithm, due to the fact that close to regions on the boundary where at least on rate gets close to zero, a large number of iterations are needed until a significant refinement results. A similar behaviour is reported in [32]. Following [32], the convergence speed of the algorithm can be significantly improved by modifying the direction of the line whose intersection with $\mathcal{E}$ defines the next iterate $\boldsymbol{y}^{(n)}$. Computationally, this is achieved by setting $\boldsymbol{n} = \check{\boldsymbol{r}}^{(n)} + \boldsymbol{a}$, $\boldsymbol{a} \in \mathbb{R}_+^K$ in the algorithm from Section V-B.

An initial vertex set $\mathcal{V}^{(0)}$ can be determined as follows: Define a rate vector $\boldsymbol{v} \in \mathbb{R}_+^K$ whose k-th entry v_k corresponds to the maximum rate achievable for user k. Then $\mathcal{V}^{(0)} = \{\omega \boldsymbol{v}\}$ with $\omega \geq 1$ defines a polyblock that contains the capacity region.

VII. Dual Decomposition

For concave utilities, a dual approach to solve the utility maximization problem in the MIMO BC was recently proposed in [14]. The algorithm in [14] represents an application of the dual decomposition [10]. Similar to the gradient-based method developed in Section V, the solution is found in two steps: First, an optimal rate vector $\boldsymbol{r}^*$ is found by operating in the rate space, second, the optimal parameters are derived from $\boldsymbol{r}^*$.

In the first step, problem (4) is modified by introducing additional variables:

$$\max_{\boldsymbol{r},\boldsymbol{s}} u(\boldsymbol{s}) \quad \text{s.t.} \quad \boldsymbol{0} \leq \boldsymbol{s} \leq \boldsymbol{r}, \boldsymbol{r} \in \mathcal{C}. \tag{31}$$

The dual function is chosen as

$$g(\boldsymbol{\lambda}) = \underbrace{\max_{\boldsymbol{s} \geq \boldsymbol{0}} u(\boldsymbol{s}) - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{s}}_{g_{\mathrm{A}}(\boldsymbol{\lambda})} + \underbrace{\max_{\boldsymbol{r} \in \mathcal{C}} \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{r}}_{g_{\mathrm{P}}(\boldsymbol{\lambda})}.$$

Evaluating the dual function at $\boldsymbol{\lambda}$ results in two decoupled subproblems: Computing $g_{\mathrm{A}}(\boldsymbol{\lambda})$ and $g_{\mathrm{P}}(\boldsymbol{\lambda})$ by maximizing over the primal variables $\boldsymbol{s}$ and $\boldsymbol{r}$, respectively. Computing $g_{\mathrm{P}}(\boldsymbol{\lambda})$ is again a WsrMax problem.

The optimal dual variable is found by minimizing the dual function with respect to $\boldsymbol{\lambda}$. The dual function is always convex, regardless of the properties of the utility function u [16].

If the utility function u is concave, strong duality holds, and the optimal primal solution $\boldsymbol{r}^*$ can be recovered from the dual solution by employing standard methods for primal recovery, as in [14]. Also, for concave u, efficient methods exist to find a set of corner points that implement $\boldsymbol{r}^*$ in the case of time-sharing optimality [30].

Being entirely based on Lagrange duality, a nonconcave utility poses significant problems to the dual decomposition. Most importantly, recovering an *optimal* primal solution $(\boldsymbol{r}^*, \boldsymbol{s}^*)$ from the dual solution is, in general, no longer possible. Moreover, the schemes for recovering all parameters $\boldsymbol{x}_{\mathrm{P}}$ of a time-sharing solution rely on strong duality to hold. [30]. For nonconcave u, however, strong duality cannot be assumed to hold. In fact, simulation results in Section VIII show a significant duality gap in the scenario under consideration.

As a result, for nonconcave u, the following *heuristic* is used to obtain a primal feasible solution $(\hat{\boldsymbol{r}}, \hat{\boldsymbol{s}})$: Given the optimal dual variable $\boldsymbol{\lambda}^*$, choose $\hat{\boldsymbol{r}} = \boldsymbol{r}_{\mathrm{wsr}}(\boldsymbol{\lambda}^*, \pi^*)$, where π^* is any optimal encoding order. Moreover, let $\hat{\boldsymbol{s}} = \hat{\boldsymbol{r}}$. An upper bound on the loss incurred by this approximation follows immediately from weak duality: Let u^* denote the (unknown) maximum utility value. By weak duality, $g(\boldsymbol{\lambda}^*) \geq u^*$, thus $u^* - u(\hat{\boldsymbol{r}}) \leq g(\boldsymbol{\lambda}^*) - u(\hat{\boldsymbol{r}})$. The tightness of this bound clearly depends on the duality gap, which is not known.

VIII. Simulation Results

Utility maximization in a $K = 3$ user Gaussian MIMO Broadcast channel with $N = 6$ transmit antennas and $M_k = 2$ receive antennas per user is simulated. The channels $\boldsymbol{H}_k$ are i.i.d. unit-variance complex Gaussian. Furthermore, the maximum transmit power is $P_{\mathrm{tr}} = 10$. To obtain rates in kbps, rates are multiplied by a bandwidth factor $W = 60$kHz.

In the simulations, the utility u is given by a weighted sum of the users' utilities u_k:

$$u(\boldsymbol{r}) = \sum_{k=1}^{K} w_k u_k(r_k).$$

The IEA method always uses a sum-rate maximizing rate vector as initial point $\boldsymbol{r}^{(0)}$. The results are averaged over 1000 channel realizations.

Two different models for the users' utilities u_k are considered: A concave logarithmic utility and a nonconcave sigmoidal utility.

A. Concave Utility

The logarithmic utility function is defined as

$$u_k(r_k) = b \ln(1 + c^{-1} r_k),$$

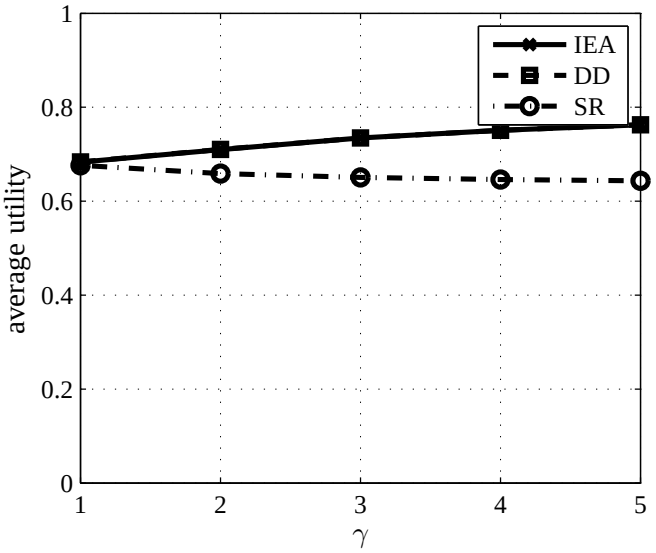

Figure 2. Average utility (concave utilities)

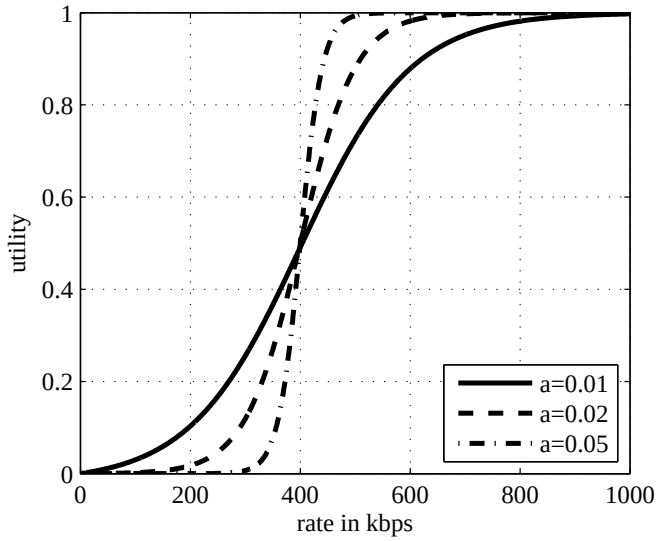

Figure 3. Sigmoid utility function, $b = 400$kbps

with constants b, c. In the simulations, $c = 40$kbps and b is chosen such that $u_k(1000\text{kbps}) = 1$. The weights w_k are chosen according to the following scheme:

$$\boldsymbol{\omega} = \begin{pmatrix} 1 & \gamma & \gamma^2 \end{pmatrix},$$
$$\boldsymbol{w} = \frac{\boldsymbol{\omega}}{\|\boldsymbol{\omega}\|_1}.$$

with $\gamma \in \{1, \ldots, 5\}$. Figure 2 shows the average utility for the case of logarithmic utility functions. Shown is the average utility for the gradient-based approach (IEA), for the dual decomposition (DD), and, as a reference, the average utility obtained by using for transmission the sum-rate maximizing rate vector that corresponds to encoding order $\pi = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$ (SR). Due to the fact that the utility maximization problem is convex, both IEA and DD achieve identical performance. Moreover, for identical weights w_k, cross-layer optimization does not provide a significant gain compared to the sum-rate maximizing strategy. The larger the difference between the users' weights, the larger the gain achieved by cross-layer optimization. This result is not surprising, as for asymmetric setups, it is more important to adapt the physical layer to the characteristics of the upper layers. Moreover, the decay of the logarithmic utility function is rather moderate around the optimal rate vector, therefore a maximizer of the weighted sum-rate is almost optimal for equal weights.

B. Nonconcave Utility

The nonconcave utility model is adopted from [8]. For each user k, the following sigmoidal utility function is used:

$$u_k(r_k) = c_k \left(\frac{1}{1 + \exp(-a_k(r_k - b_k))} + d_k \right),$$

where c_k and d_k are used to normalize u_k such that $u_k(0) = 0$ and $u_k(\infty) = 1$. The steepness of the transition between the minimum and the maximum value is controlled by the parameter a_k, whereas b_k determines the inflection point of the utility curve (cf. Figure 3). In the simulations, $a_k = a\,\text{kbps}^{-1}$, and a is varied in a range between 0.01 and 0.05, modelling different degrees of elasticity of the applications. For each channel realization, the constant b_k of each user is chosen randomly in the interval $[300\text{kbps}, 500\text{kbps}]$ according to a uniform distribution. Choosing the b_k randomly can be understood as a model for fluctuations in the data rate requirements of the users over time, e.g., transmission of a video source with varying scene activity. All users have equal weight $w_k = 1/K$.

Figure 4 shows the average utility for the case of sigmoidal utility functions. Shown is the average utility for the gradient-based approach (IEA), the polyblock algorithm (PB), the dual decomposition (DD), and the sum-rate maximizing rate vector (SR). In addition, the average minimum value of the dual function in the dual decomposition approach is shown (DUB). The PB algorithm finds the global maximum for each realization. As a result, the PB curve gives the maximum achievable average utility. In terms of average utility, the performance of the IEA method is close to optimal. It can be concluded that for the system setup under

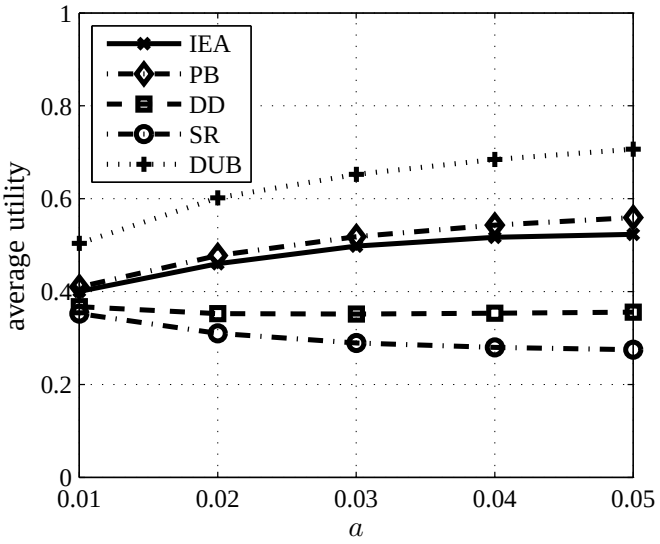

Figure 4. Average utility (sigmoidal utilities)

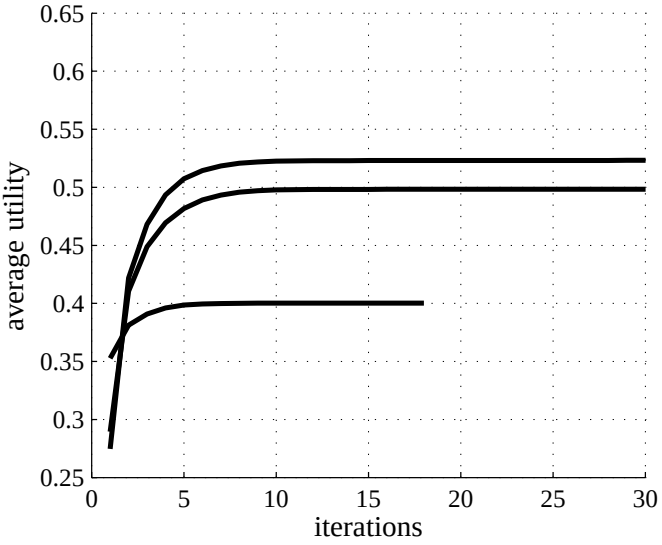

Figure 5. Average utility vs. number of iterations, IEA method

consideration, the IEA method succeeds in finding a stationary point which is identical or close to the global maximum for most realizations. In contrast, the dual decomposition-based method does not find a good rate vector in most cases. The poor performance of the computationally simple SR strategy emphasizes the need for cross-layer optimization. In particular, the performance gain achieved by both PB and IEA increases with a. This behaviour can be explained as follows: With increasing a, the interval in which the utility function makes a transition from small to large values becomes smaller. Therefore, it becomes more and more important to adapt the physical layer parameters to the utility characteristics.

The results in Figure 4 also show that the dual upper bound (DUB) obtained from the dual decomposition is rather loose. This implies that there is a significant duality gap in most cases.

C. Complexity Analysis

If average utility is the only figure of merit, the polyblock algorithm is obviously superior to all other approaches. From a practical viewpoint, a second metric of interest is the computational complexity of the different approaches. In the following, the utility-complexity trade-offs provided by the different approaches are investigated. All results are for the case of sigmoidal utility functions.

In Figure 5, average utility is plotted versus the number of iteration for the IEA method. The plot shows three graphs, corresponding to three different values of the steepness parameter a: $a \in \{0.01, 0.03, 0.05\}$. Note that the rightmost points of each graph corresponds to the average utility value in Figure 4. Only the gradient based outer iterations defined in Eqs. (11) - (13) are counted, the inner iterations in the correction step are neglected. Figure 5 shows that the IEA method needs on average between five and 10 iterations to get close to the maximum achieved utility value.

In Figure 6, average utility is plotted versus the number of iteration for the polyblock algorithm. The plot shows three pairs of graphs, with each pair corresponding to a different value of the steepness parameter a: $a \in \{0.01, 0.03, 0.05\}$. Each pair consists of two graph, one showing the average of the current best utility value $u(\hat{\boldsymbol{r}}^{(n)})$ (CBV, dash-dotted line), the other showing the average of the upper bound $u(\check{\boldsymbol{r}}^{(n)})$ (UB, solid line). Depending on the parameter a, between 50 to 75 iterations are needed until the current best value is close to the global maximum. Recall from Section VI-A that the convergence criterion for the PB algorithm is based on the difference between $u(\hat{\boldsymbol{r}}^{(n)})$ and $u(\check{\boldsymbol{r}}^{(n)})$. Figure 6 shows that a large number of iterations may be required until convergence is declared, due to the relatively slow convergence of the upper bound.

In both Figure 5 and Figure 6, the number of inner iterations required in the correction step and the computation of the intersection point, respectively, are not counted. In each inner iteration, a WsrMax problem is solved. Moreover, a WsrMax problem is also solved at each iteration of the dual decomposition. Accordingly, all three approaches can be compared based on the number

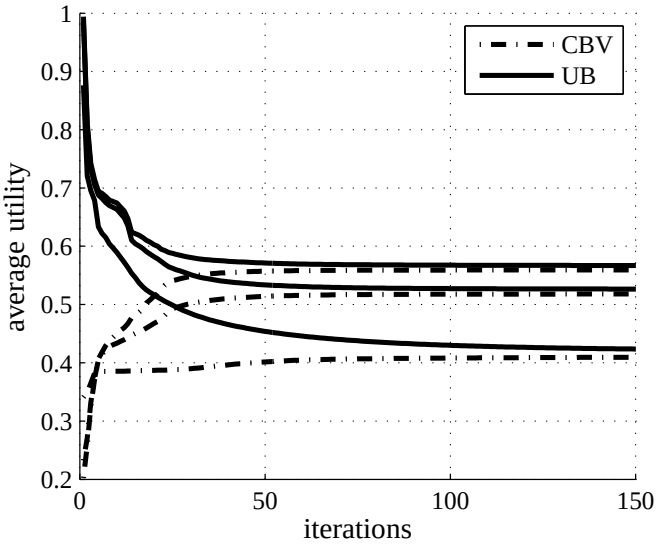

Figure 6. Average utility vs. number of iterations, PB algorithm

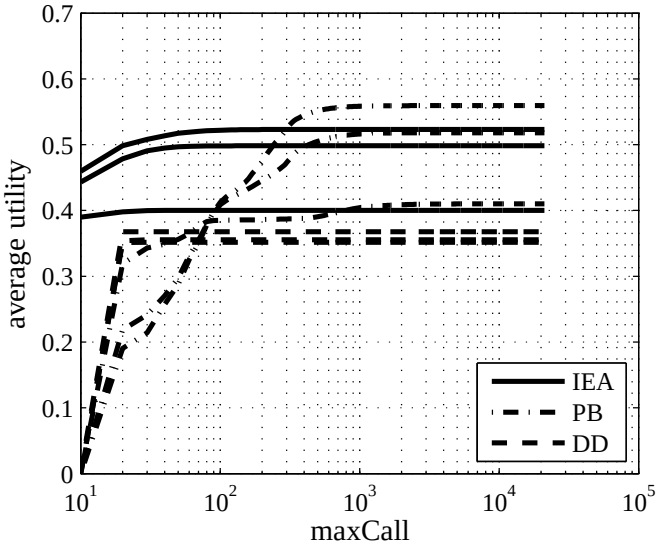

Figure 7. Average utility vs. maximum number of WsrMax calls

of calls to the WsrMax subroutine. Figure 7 shows the average utility that is achieved if the maximum number of calls to WsrMax is limited to a value maxCall, with maxCall increased in steps of 10 calls. Again, three groups of graphs are shown, each group corresponding to a value of a, with $a \in \{0.01, 0.03, 0.05\}$. As an example, the results show that the dual decomposition needs between 10 and 20 iterations until convergence (to a clearly suboptimal solution). Of particular interest are the cross-over points between IEA method and PB algorithm. For $a = 0.05$, the cross-over point is at maxCall $= 300$, i.e., only if more than a maximum of 300 calls to WsrMax are feasible does the PB algorithm outperform the IEA method. Moreover, for small values of maxCall, the IEA method provides significantly larger average utility.

IX. Conclusions

Two methods for solving the nonconcave utility maximization problem in the MIMO broadcast channel are proposed: a gradient-based method that converges to a locally optimal solution, and an approach based on monotonic optimization that yields the global optimum. Due to the structure of the MIMO BC capacity region, a direct optimization in terms of the physical layer parameters transmit covariance matrices and encoding order is not feasible. Thus, as an intermediate step, both methods first determine an optimal rate vector. The optimal physical layer parameter setup, which may include a time-sharing solution, is then obtained from this rate vector. For both methods, the formulation of the utility maximization problem in the rate space represents a key step: The IEA method exploits the differentiable manifold structure of the efficient boundary of the capacity region, while the polyblock algorithm relies on the fact that maximizing utility over the set of achievable rate vectors represents a monotonic optimization problem.

The polyblock algorithm provides globally optimal performance, at the price of a relatively high computational complexity. From a practical viewpoint, the proposed IEA method provides an attractive trade-off between utility performance and computational complexity: In the simulation setup used in this work, the average utility achieved by the IEA method is close to optimal, at significantly lower complexity than the polyblock algorithm.

Throughout this work, it is assumed that users' rates are the only relevant performance metrics of the physical layer, implying that rate cannot be traded for delay and reliability. In a more general setup, more than one performance metric per user may be required to characterize the physical layer, corresponding to a utility function that is a function of all these metrics [7]. Concerning the results presented in this work, this would clearly impact the mapping from physical layer parameters to set of achievable performance vectors. The methods proposed in this work, however, would still be applicable, provided the structural assumptions of each method are still met (i.e., the utility function is monotone in all physical layer metrics, the set of achievable performance vectors is compact and, in case of the IEA method, can be

equipped with a differentiable manifold structure). While the capacity region is convex, it is not clear whether a generalized achievable region can still be assumed to be convex. This observation represents a further motivation for an optimization framework that does not rely on the assumption of convexity.

REFERENCES

[1] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[2] P. Viswanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.

[3] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 49, pp. 2658–2668, 2003.

[4] H. Viswanathan, S. Venkatesan, and H. Huang, "Downlink capacity evaluation of cellular networks with known-interference cancellation," *IEEE Journal on Selected Areas in Communications*, vol. 21, pp. 802–811, 2003.

[5] N. Jindal and A. Goldsmith, "Dirty-paper coding versus TDMA for MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 51, no. 5, pp. 1783–1794, May 2005.

[6] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on Information Theory*, vol. 52, pp. 3936–3964, 2006.

[7] S. Shenker, "Fundamental design issues for the future internet," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1176–1188, September 1995.

[8] J.-W. Lee, R. Mazumdar, and N. Shroff, "Non-convex optimization and rate control for multi-class services in the internet," *IEEE/ACM Transactions on Networking*, vol. 13, no. 4, pp. 827–840, Aug. 2005.

[9] M. Fazel and M. Chiang, "Network utility maximization with nonconcave utilities using sum-of-squares method," in *IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, December 2005, pp. 1867–1874.

[10] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: a mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, January 2007.

[11] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, pp. 33–37, January 1997.

[12] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, pp. 1452–1463, 2006.

[13] S. Stanczak, M. Wiczanowski, and H. Boche, *Resource Allocation in Wireless Networks - Theory and Algorithms*, ser. Lecture Notes in Computer Science (LNCS 4000). Springer-Verlag, 2006.

[14] J. Liu and Y. T. Hou, "Cross-layer optimization of MIMO-based mesh networks with Gaussian vector broadcast channels," April 2007, *Submitted to IEEE Transactions on Information Theory. Preprint available on arXiv.* [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:0704.0967

[15] H. Tuy, "Monotonic optimization: Problems and solution approaches," *SIAM Journal on Optimization*, vol. 11, no. 2, pp. 464–494, 2000.

[16] D. Bertsekas, A. Nedic, and A. Ozdaglar, *Convex analysis and optimization*. Athena Scientific, 2003.

[17] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable H.264/MPEG4-AVC extension," in *IEEE International Conference on Image Processing (ICIP)*, October 2006, pp. 161–164.

[18] D. Tse and S. Hanly, "Multiaccess fading channels. I. polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2796–2815, Nov 1998.

[19] 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 Communications*, vol. 24, pp. 1640–1646, 2006.

[20] R. Böhnke and K. Kammeyer, "Weighted sum rate maximization for the MIMO-downlink using a projected conjugate gradient algorithm," in *International Workshop on Cross Layer Design (IWCLD)*, Jinan, China, September 2007.

[21] D. G. Luenberger, "The gradient projection method along geodesics," *Management Science*, vol. 18, pp. 620–631, July 1972.

[22] D. Gabay, "Minimizing a differentiable function over a differentiable manifold," *Journal of Optimization Theory and Applications*, vol. 37, pp. 177–219, June 1982.

[23] J. H. Manton, "Optimization algorithms exploiting unitary constraints," *IEEE Transactions on Signal Processing*, vol. 50, pp. 635–650, 2002.

[24] J. B. Rosen, "The gradient projection method for nonlinear programming, II: Nonlinear constraints," *SIAM Journal on Applied Mathematics*, vol. 9, pp. 514–553, 1961.

[25] W. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, 1975.

[26] W. Zangwill, *Nonlinear Programming: A Unified Approach*. Prentice-Hall, 1969.

[27] J. H. Manton, "On the various generalisations of optimisation algorithms to manifolds," in *International Symposium on Mathematical Theory of Networks and Systems (MTNS)*, 2004.

[28] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed. Addison-Wesley, Inc., 1984.

[29] J.-L. Goffin and J.-P. Vial, "Convex nondifferentiable optimization: a survey focussed on the analytic center cutting plane method," Logilab, Geneva, Switzerland, Tech. Rep. 99.02, 1999.

[30] J. Brehmer, Q. Bai, and W. Utschick, "Time-sharing solutions in MIMO broadcast channel utility maximization," in *International Workshop on Smart Antennas (WSA)*, Darmstadt, Germany, February 2008.

[31] M. Riemensberger and W. Utschick, "Gradient descent on differentiable manifolds," Technische Universität München, Munich, Germany, Tech. Rep. MSV-2008-1, 2008.

[32] H. Tuy, F. Al-Khayyal, and P. Thach, "Monotonic optimization: Branch and cut methods," in *Essays and Surveys in Global Optimization*, C. Audet, P. Hansen, and G. Savard, Eds. Springer US, 2005, pp. 39–78.