

A Decomposition of the Downlink Utility Maximization Problem

Johannes Brehmer and Wolfgang Utschick
Associate Institute for Signal Processing
Munich University of Technology
{brehmer, utschick}@tum.de

Abstract—An alternative decomposition method for sum-utility maximization in the multi-user downlink is proposed. The proposed method is based on repeated local approximations of the Pareto efficient boundary of the rate region. Similar to known “Layering as Optimization” approaches, the utility maximization problem is vertically decomposed into a set of coupled sub-problems. Mathematically, however, the proposed method is not based on primal or dual decompositions, but on an optimization on manifolds.

I. INTRODUCTION

In *network utility maximization* (NUM), routing and resource allocation are optimized to maximize a sum utility over all nodes. To this end, an optimization problem is formulated and solved. Recently, techniques to decompose the NUM optimization problem into a set of coupled subproblems have received wide attention [1], [2]. In these works, each subproblem is associated with a layer, giving rise to the term “Layering as Optimization Decomposition” (LOD), see [2] and references therein. The LOD framework provides two main classes of decompositions: horizontal decomposition and vertical decomposition. In this work, utility maximization in the wireless multi-user downlink is considered, in the following denoted as *downlink utility maximization* (DUM) problem. A central transmitter is assumed, therefore the focus is on a vertical decomposition into functional modules. As the downlink represents just a very simple case of a network, the techniques developed in the LOD framework are readily applicable [2]. Within the LOD framework, from an optimization viewpoint the main techniques employed are primal and dual decompositions. In this work, a different approach to achieve a vertical decomposition is proposed. The key idea is to view the DUM problem as an optimization over a manifold. This allows to directly apply results from the optimization literature to the DUM problem. As in the LOD framework, a decomposition into a set

of coupled subproblems is achieved. The coupling as well as the subproblems, however, are of a very different type. In particular, it is sufficient to compute a single update in the APP layer subproblem for each inter-problem iteration, while a dual decomposition requires a completely solved APP layer problem at each iteration. Moreover, the coupling between layers is not based on a subgradient iteration, but on a first order approximation of the manifold, which may lead to faster convergence. These observations hint at the competitiveness of the proposed approach compared to a duality-based decomposition.

II. DUM PROBLEM

Let $\mathcal{R} \subset \mathbb{R}_{0,+}^K$ denote the achievable rate region of a K user downlink (resp. broadcast channel). In general, if information rates are used as performance metric, \mathcal{R} completely characterizes the PHY layer. Which rate vectors $\mathbf{r} \in \mathbb{R}_{0,+}^K$ are achievable depends on the specific choice of the PHY layer system model. In this work, it is assumed that the PHY layer provides a time-sharing mode, i.e., it is assumed that the rate region \mathcal{R} is convex.

Let $u_k(r_k)$ express the utility perceived by user k given a rate r_k is allocated to user k . It is assumed that u_k is differentiable, monotonically increasing, and concave. Define the sum-utility $u : \mathbb{R}_{0,+}^K \rightarrow \mathbb{R}$ by

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

Downlink utility maximization seeks to maximize sum-utility over the set of achievable rates:

$$\max_{\mathbf{r} \in \mathcal{R}} u(\mathbf{r}). \quad (1)$$

In the form of (1), DUM is a simple version of the general NUM problem in [2]. Recently, similar setups were considered in [3], [4].

III. DUAL DECOMPOSITION

As already stated in the introduction, the techniques from the LOD framework are readily applicable to the DUM problem. In this section, the application of a dual decomposition to achieve a vertical decomposition into functional modules is briefly summarized.

Following [2], the DUM problem is first modified by introducing additional variables:

$$\max_{\mathbf{r}, \mathbf{s}} u(\mathbf{s}) \quad \text{s.t.} \quad \mathbf{s} \leq \mathbf{r}, \mathbf{r} \in \mathcal{R}. \quad (2)$$

After introducing the Lagrangian

$$L(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}) = u(\mathbf{s}) + \boldsymbol{\lambda}^T (\mathbf{r} - \mathbf{s})$$

the dual function is given by

$$g(\boldsymbol{\lambda}) = g_A(\boldsymbol{\lambda}) + g_P(\boldsymbol{\lambda}),$$

with $\boldsymbol{\lambda} > \mathbf{0}_K$ (the cases $\lambda_k = 0$ can be excluded) and

$$g_A(\boldsymbol{\lambda}) = \max_{\mathbf{s}} u(\mathbf{s}) - \boldsymbol{\lambda}^T \mathbf{s}, \quad \text{and} \quad (3)$$

$$g_P(\boldsymbol{\lambda}) = \max_{\mathbf{r} \in \mathcal{R}} \boldsymbol{\lambda}^T \mathbf{r}. \quad (4)$$

For a fixed $\boldsymbol{\lambda}$, the optimization is decomposed into two independent subproblems (3) and (4). Subproblem (3) can be solved at the APP layer, subproblem (4) is solved at the PHY layer. The optimum $\boldsymbol{\lambda}$ is found via a subgradient method. For each subgradient iteration, problems (3) and (4) have to be solved for the current $\boldsymbol{\lambda}$. Subproblem (4) is a *weighted sum-rate maximization* (WSRmax). WSRmax is a well-researched problem and efficient solutions exist for a wide range of PHY layer setups. To summarize, two fundamental features are provided by the dual decomposition:

- 1) vertical decomposition of the original problem into two ‘‘inner’’ subproblems, which can be solved independently at each iteration, and an ‘‘outer’’ subgradient-based optimization,
- 2) re-use of existing algorithms to solve PHY subproblem.

In Section IV, a decomposition is proposed that provides similar features, while being based on significantly different optimization methods.

IV. AN ITERATIVE EFFICIENT SET APPROACH

Due to the monotonicity of the utility functions u_k , the sum utility is monotone in \mathbf{r} :

$$\mathbf{r} \leq \mathbf{r}' \Rightarrow u(\mathbf{r}) \leq u(\mathbf{r}'). \quad (5)$$

Define the Pareto efficient set as follows:

$$\mathcal{E} = \{ \mathbf{r} \in \mathcal{R} : \nexists \mathbf{r}' \in \mathcal{R} \quad \text{with} \quad \mathbf{r}' > \mathbf{r} \}. \quad (6)$$

Verbally, \mathcal{E} contains the largest rate vectors (under the partial Pareto order). Due to (5), maximizing sum-utility over the entire rate region is equivalent to maximizing over the Pareto efficient boundary of \mathcal{R} , i.e.,

$$\max_{\mathbf{r} \in \mathcal{R}} u(\mathbf{r}) = \max_{\mathbf{r} \in \mathcal{E}} u(\mathbf{r}). \quad (7)$$

A closed-form expression for \mathcal{E} is in general not available. In most cases, however, algorithms for computing points $\mathbf{r} \in \mathcal{E}$ are known, such as, e.g., algorithms for maximizing the weighted sum-rate. Now, the following additional assumptions are made:

- 1) At all points $\mathbf{r} \in \mathcal{E}$ with $r_k > 0, \forall k$ exists a tangent space $\mathcal{T}_{\mathbf{r}}$ of \mathcal{E} .
- 2) At all such points \mathbf{r} , an orthonormal basis $\mathbf{Q} \in \mathbb{R}^{K \times K-1}$ of $\mathcal{T}_{\mathbf{r}}$ is available.

In other words,

$$\hat{\mathcal{E}}_{\mathbf{r}} = \{ \mathbf{r} + \mathbf{Q}\boldsymbol{\mu}, \boldsymbol{\mu} \in \mathbb{R}^{K-1} \},$$

represents a first-order approximation of \mathcal{E} around \mathbf{r} . Let $\mathbf{r}^* \in \mathcal{E}$ denote a rate vector that maximizes sum-utility, and assume that \mathbf{r}^* is unique. Obviously, starting at an arbitrary $\mathbf{r} \in \mathcal{E}$, the goal is to move toward \mathbf{r}^* . To do so, the first order approximation $\hat{\mathcal{E}}_{\mathbf{r}}$ is provided to the APP layer. The APP layer determines an update $\mathbf{Q}\tilde{\boldsymbol{\mu}}$, resulting in

$$\tilde{\mathbf{r}} = \mathbf{r} + \mathbf{Q}\tilde{\boldsymbol{\mu}}.$$

In this work, a gradient-based update is considered, i.e.,

$$\tilde{\boldsymbol{\mu}} = t\mathbf{Q}^T \nabla u(\mathbf{r}), \quad (8)$$

with a stepsize $t \geq 0$. Note that for orthogonal \mathbf{Q} ,

$$\Delta \mathbf{r} = \tilde{\mathbf{r}} - \mathbf{r} = t\mathbf{P}_{\mathbf{r}} \nabla u(\mathbf{r}),$$

where $\mathbf{P}_{\mathbf{r}}$ is the orthogonal projector on $\mathcal{T}_{\mathbf{r}}$.

In general, $\tilde{\mathbf{r}} \notin \mathcal{E}$, therefore the PHY layer has to project $\tilde{\mathbf{r}}$ back onto \mathcal{E} , resulting in a new $\mathbf{r}' \in \mathcal{E}$. Provided that a sufficient increase in u was achieved (by proper adjustment of the stepsize t), a new approximation is computed at \mathbf{r}' and the whole process is repeated until $\mathbf{r}' = \mathbf{r}^*$.

There exist different possibilities to project $\tilde{\mathbf{r}}$ on \mathcal{E} . Due to the nature of \mathcal{E} , a Euclidean projection on \mathcal{E} seems prohibitive. Instead, a projection orthogonal to the tangent space $\mathcal{T}_{\mathbf{r}}$ is employed. Let \mathbf{n} denote the unit-norm vector that is orthogonal to $\mathcal{T}_{\mathbf{r}}$ and points away from \mathcal{R} . To project $\tilde{\mathbf{r}}$ on \mathcal{E} , the following problem is solved:

$$\max_{x, \mathbf{r}} x \quad \text{s.t.} \quad \tilde{\mathbf{r}} + x\mathbf{n} \leq \mathbf{r}, \mathbf{r} \in \mathcal{R}.$$

The Lagrangian is given by

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

The dual function follows as

$$\begin{aligned} g(\boldsymbol{\lambda}) &= \sup_{\substack{x \in \mathbb{R} \\ \mathbf{r} \in \mathcal{R}}} (x(1 - \boldsymbol{\lambda}^T \mathbf{n}) + \boldsymbol{\lambda}^T (\mathbf{r} - \tilde{\mathbf{r}})) \\ &= \begin{cases} +\infty, & \boldsymbol{\lambda}^T \mathbf{n} \neq 1, \\ \max_{\mathbf{r} \in \mathcal{R}} \boldsymbol{\lambda}^T (\mathbf{r} - \tilde{\mathbf{r}}), & \boldsymbol{\lambda}^T \mathbf{n} = 1. \end{cases} \end{aligned} \quad (9)$$

Note that for $\boldsymbol{\lambda}^T \mathbf{n} = 1$, again a weighted sum-rate maximization problem is to be solved. Let $\mathbf{r}^*(\boldsymbol{\lambda})$ denote an optimizer of the weighted sum-rate maximization in (9). The optimum dual variable $\boldsymbol{\lambda}$ is found by solving

$$\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \boldsymbol{\lambda}^T (\mathbf{r}^*(\boldsymbol{\lambda}) - \tilde{\mathbf{r}}) \quad \text{s.t.} \quad \boldsymbol{\lambda}^T \mathbf{n} = 1 \quad (10)$$

via a subgradient method.

Similar to the dual decomposition, a dual problem is solved via a subgradient method. However, in the efficient set approach, there are now two iteration levels: outer iterations based on a gradient update and inner iterations within the PHY layer to determine the projection. In contrast to the dual decomposition, the inner subgradient iterations performed for projection on \mathcal{E} do not involve the application layer. While in the dual decomposition the APP layer has to fully solve an optimization problem (3) for each subgradient iteration, in the proposed efficient set method, the APP only has to compute a gradient at each outer iteration.

A fundamental assumption is the availability of a tangent space basis \mathbf{Q} . If a weighted-sum rate maximization is employed in the projection step, the computation of \mathbf{Q} is trivial. Let $\boldsymbol{\lambda}'$ denote the optimum dual variable of (10) and \mathbf{r}' a rate vector that maximizes the weighted-sum rate for this particular weighting $\boldsymbol{\lambda}'$, i.e.,

$$(\boldsymbol{\lambda}')^T \mathbf{r}' = \max_{\mathbf{r} \in \mathcal{R}} (\boldsymbol{\lambda}')^T \mathbf{r}.$$

It is known that $\boldsymbol{\lambda}'$ is orthogonal to the tangent space $\mathcal{T}_{\mathbf{r}'}$ at \mathbf{r}' (see [5] for an intuitive treatment), thus

$$\mathcal{T}_{\mathbf{r}'} = \text{null}((\boldsymbol{\lambda}')^T).$$

In other words, a basis \mathbf{Q} of $\mathcal{T}_{\mathbf{r}'}$ is found by computing an orthonormal basis of $\text{null}((\boldsymbol{\lambda}')^T)$ — thus, the proposed projection algorithm provides a basis of the tangent space needed for the next iteration “almost for free”.

From the perspective of layering as optimization decomposition, again a decomposition into two functional modules takes place: At each outer iteration,

- 1) PHY layer: compute basis \mathbf{Q} and project $\tilde{\mathbf{r}}$ on \mathcal{E} ,
- 2) APP layer: compute update $\mathbf{Q}\tilde{\boldsymbol{\mu}}$.

Communication between layers is in terms of (\mathbf{r}, \mathbf{Q}) (PHY \rightarrow APP) resp. $\tilde{\mathbf{r}}$ (APP \rightarrow PHY).

V. OPTIMIZATION ON MANIFOLDS

In this section, the main results concerning the mathematical properties of the proposed decomposition approach are briefly discussed. Denote by $\mathcal{C} \subset \mathcal{E}$ the set of efficient rate vectors where at least one user has rate zero. Define the open set \mathcal{M} as follows:

$$\mathcal{M} = \mathcal{E} \setminus \mathcal{C}. \quad (11)$$

The assumption that a tangent space $\mathcal{T}_{\mathbf{r}}$ exists at each point in \mathcal{M} corresponds to the assumption that \mathcal{M} is a differentiable manifold. Thus

$$\max_{\mathbf{r} \in \mathcal{M}} u(\mathbf{r}) \quad (12)$$

is an optimization over a differentiable manifold. Two early works on optimization on manifolds are [6], [7]. In [7], steepest descent and Newton methods are generalized to optimization on manifolds by moving along geodesics. Recently, optimization on manifolds has also received attention in the signal processing community (see [8] for an overview).

Define a local parameterization $\boldsymbol{\theta}_{\mathbf{r}}$ of \mathcal{M} at \mathbf{r} as follows:

$$\boldsymbol{\theta}_{\mathbf{r}}(\boldsymbol{\mu}) = \mathbf{r} + \mathbf{Q}\boldsymbol{\mu} + n_x(\boldsymbol{\mu}) \in \mathcal{M}.$$

According to the previous section, $\mathbf{r} + \mathbf{Q}\boldsymbol{\mu} \in \hat{\mathcal{E}}_{\mathbf{r}}$ and $n_x(\boldsymbol{\mu})$ corresponds to the projection step. Moreover,

$$\begin{aligned} \boldsymbol{\theta}_{\mathbf{r}}(\mathbf{0}) &= \mathbf{r}, \quad \text{and} \\ \nabla \boldsymbol{\theta}_{\mathbf{r}}(\mathbf{0}) &= \mathbf{Q}^T, \end{aligned}$$

where $\nabla \boldsymbol{\theta}_{\mathbf{r}}(\boldsymbol{\mu})$ denotes the transpose of the Jacobian matrix of $\boldsymbol{\theta}_{\mathbf{r}}$ at $\boldsymbol{\mu}$. Locally, maximizing $u \circ \boldsymbol{\theta}_{\mathbf{r}}$ corresponds to an unconstrained optimization problem. In particular, the gradient at $\boldsymbol{\mu} = \mathbf{0}$ is given by

$$\nabla(u \circ \boldsymbol{\theta}_{\mathbf{r}})(\mathbf{0}) = \mathbf{Q}^T \nabla u(\mathbf{r}),$$

which is also the gradient used in Eq. (8). In other words, the APP layer can compute the “correct” gradient update based on $\hat{\mathcal{E}}$. This property clearly depends on the fact that the projection step fullfills

$$\nabla x(\mathbf{0}) = \mathbf{0}.$$

Finally, note that the parameterization $\boldsymbol{\theta}_{\mathbf{r}}$ corresponds to the so-called “tangent restoration approach” in [7].

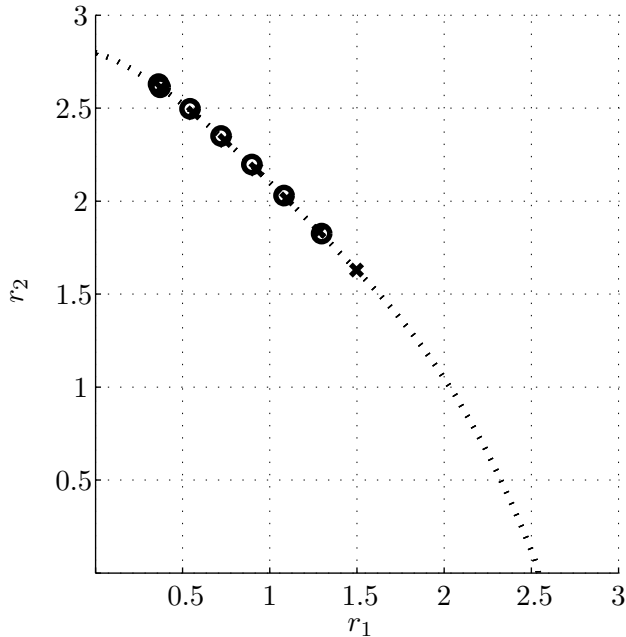


Fig. 1. Iterative Efficient Set Method

VI. NUMERICAL RESULTS

In this section, some numerical results are presented to outline the basic behaviour of the discussed decomposition approaches. At the PHY layer, an OFDMA mode with 512 subcarriers is considered. For this PHY setup, an efficient algorithm to compute a (closely) optimum subcarrier and power allocation under a weighted sum-rate criterion is provided in [9]. The algorithm from [9] is employed without an additional time-sharing mode. As a result, the set of achievable rate points is non-convex and the efficient set is non-smooth. Simulations show that the algorithms discussed in this paper still provide good performance in such a setting, see also [4]. A detailed theoretical analysis of such non-convexity issues for a finite number of subcarriers remains an open problem.

Results are limited to the case of $K = 2$ users. In this case, the tangent space $\mathcal{T}_{\tilde{r}}$ is simply a line. Fig. 1 shows an exemplary run of the efficient set-based decomposition. The dotted line corresponds to the Pareto efficient boundary \mathcal{E} of the rate region \mathcal{R} . For the scenario under consideration, the rate vector that maximizes sum-utility is (approximately) $\mathbf{r}^* = (0.4, 2.6)$. The algorithm is initialized with a sum-rate maximizing rate vector at $(1.5, 1.6)$. In Fig. 1, crosses correspond to rate vectors obtained by projecting on \mathcal{E} (except for the initial value at $(1.5, 1.6)$), and circles to rate vectors $\tilde{\mathbf{r}}$ requested by the APP layer. In order to save computational complex-

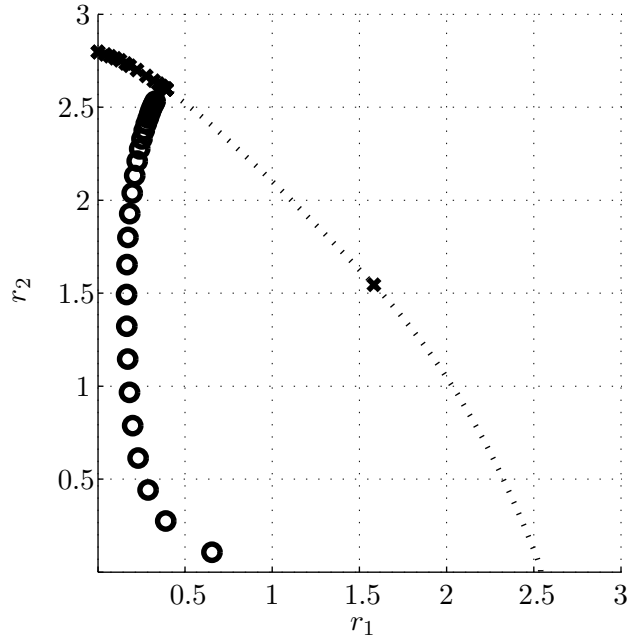


Fig. 2. Dual Decomposition

ity, the projection is terminated early if an efficient rate vector sufficiently close to $\tilde{\mathbf{r}}$ is found. Due to the small stepsize, the first order approximation is very good, and circles and crosses are hardly distinguishable. In this example, the optimum rate vector is found after 6 outer iterations. In comparison, Fig. 2 shows the convergence of a subgradient-based dual decomposition. In Fig. 2, crosses correspond to solutions of the PHY subproblems (i.e., rate vectors $\mathbf{r} \in \mathcal{R}$ that maximize the weighted sum-rate for a certain value of the Lagrangian multiplier λ), while circles correspond to variables \mathbf{s} . Due to the fact that the dual decomposition does not exploit the available local information about \mathcal{E} , more outer iterations are required to find the optimum “price” λ^* and the corresponding rate vector \mathbf{r}^* .

VII. CONCLUSIONS

A novel approach to decompose the downlink utility maximization problem is proposed. It is based on a local approximation of the set of efficient rate vectors. Based on an information exchange between layers, the APP layer guides the PHY layer towards the optimum rate allocation in an iterative manner. Mathematically, the proposed method can be understood as an optimization on a manifold. This allows for the direct application of results from the optimization literature.

REFERENCES

- [1] M. Chiang, S. H. Low, A. R. Calderbank, and J. Doyle, "Layering as optimization decomposition: current status and open issues," in *Proceedings of the Conference on Information Sciences and Systems (CISS)*, March 2006.
- [2] 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, 2007.
- [3] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Downlink scheduling and resource allocation for OFDM systems," in *Proceedings of the Conference on Information Sciences and Systems (CISS)*, March 2006.
- [4] T. C. Ng, W. Yu, J. Zhang, and A. Reid, "Joint optimization of relay strategies and resource allocations in cooperative cellular networks," in *Proceedings of the Conference on Information Sciences and Systems (CISS)*, March 2006.
- [5] I. Das and J. Dennis, "A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems," Rice University, Tech. Rep. 96-36, 1996.
- [6] D. G. Luenberger, "The gradient projection method along geodesics," *Management Science*, vol. 18, pp. 620-631, July 1972.
- [7] D. Gabay, "Minimizing a differentiable function over a differentiable manifold," *Journal of Optimization Theory and Applications*, vol. 37, pp. 177-219, June 1982.
- [8] J. H. Manton, "On the role of differential geometry in signal processing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, March 2005, pp. 1021-1024.
- [9] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *IEEE International Symposium on Information Theory (ISIT)*, 2006, pp. 1394-1398.