[15] X. Kang, Y.-C. Liang, H. K. Garg, and L. Zhang, "Sensing-based spectrum sharing in cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 4649–4654, Oct. 2009.

[16] Z. Quan, S. Cui, A. H. Sayed, and H. V. Poor, "Optimal multiband joint detection for spectrum sensing in cognitive radio networks," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 1128–1140, Mar. 2009.

[17] K. B. Letaief and W. Zhang, "Cooperative communications for cognitive radio networks," *Proc. IEEE*, vol. 97, no. 5, pp. 878–893, May 2009.

[18] X. Gong, S. A. Vorobyov, and C. Tellambura, "Joint bandwidth and power allocation with admission control in wireless multi-user networks with and without relaying," *IEEE Trans. Signal Process.*, accepted with minor revision.

[19] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[20] C. Floudas and P. M. Pardalos, *Encyclopedia of Optimization*, 2nd ed. New York: Springer-Verlag, 2009.

A Layered Decomposition Framework for Resource Allocation in Multiuser Communications

Cho Yiu Ng, *Student Member, IEEE*,
Kenneth W. Shum, *Member, IEEE*, Chi Wan Sung, *Member, IEEE*,
and Tat Ming Lok, *Senior Member, IEEE*

***Abstract*—The resource allocation problem for multiuser channels is decomposed into two layers. The lower layer is the weighted sum rate maximization, which is widely considered for many different multiuser channels. The weighted sum rate maximization is employed as a subroutine and called the upper layer. The upper layer is the optimization of dual variables for maximizing a joint utility function, which is very general and includes proportional fairness and max-min fairness as special cases. This layered approach decouples the physical-layer technologies from system-layer consideration. For example, if we want to evaluate a new objective in resource allocation, changes are required only in the outer loop, whereas the inner loop remains the same. This induces flexibility in the software structure. To numerically obtain the solution, a Gauss–Seidel-type algorithm is proposed, and its effectiveness in achieving proportional fairness in the parallel Gaussian broadcast channel is demonstrated by computer simulation.**

***Index Terms*—Dual decomposition, network utility maximization, orthogonal frequency-division multiple-access (OFDMA), proportional fairness.**

I. Introduction

For multiuser communications, a common performance measure is the sum rate of all users. For downlink cellular systems, as the distances between the users and the base station may spread over a very wide range, maximizing the sum rate will penalize users who are far away from the base station. To alleviate this near–far unfairness, one solution is to maximize the weighted sum rate. Weighted sum rate maximization for a parallel Gaussian broadcast channel (BC) is considered in [1] and [2], and that for orthogonal frequency-division multiple-access (OFDMA) systems is considered in [3]. On the other hand, it is unclear how to pick a set of good weighting factors. Another alternative is to assign data rates according to some notion of fairness, e.g., proportional fairness [4], max-min fairness, etc. While the proportional fairness problem for OFDMA systems has been considered in [5], we consider the same problem for the parallel Gaussian BC but under a more general framework based on the concept of joint utility function.

In view of the fact that much research effort has been put into maximizing the achievable rate region of a given multiuser communication channel, our framework employs the weighted sum rate maximization algorithm, which is a common way to characterize the rate region, as a subroutine. This isolates the physical-layer issue from other higher layer consideration. It can be applied to many multiuser channel models such as the parallel Gaussian BC, fading multiple access channel [6], and the cooperative transmission scheme in [7]. Whenever there is a need to adopt a new resource-allocation objective or an advance in physical-layer technologies, we only need to change some of the modules rather than redevise the whole optimization algorithm. Such flexible software structure can reduce the development cost and maintenance cost of the resource-allocation software. In addition, it also allows researchers to decouple a complex problem and focus on one of the aspects.

The optimization framework discussed in this paper is based on a layered decomposition approach similar to those in [8] and [9]. However, a more general joint utility function is adopted in this paper, whereas, in [8] and [9], the utility function to be optimized is assumed to be separable, i.e., it is the sum of the utility functions pertaining to the users. The approach in this paper is thus more flexible. Our iterative algorithm, which updates the dual variables in a Gauss–Seidel manner, is more efficient than the subgradient approach suggested in [8] and [9]. To illustrate the difference in computation time, we compare our algorithm with an algorithm proposed in [10], which is based on the subgradient approach and designed for the parallel Gaussian BC. Simulation results show that our algorithm has faster convergence. On the other hand, we remark that the subgradient approach is more appropriate for distributed implementation. Our proposed algorithm in this paper can be considered as a tradeoff between performance and the how distributed the algorithm is.

In Section II, we present a joint utility optimization for parallel Gaussian BC as a motivating example. The decomposition framework and convergence results are detailed in Section III. Some numerical examples are given in Section IV.

II. Motivating Example and Problem Formulation

Consider a family of K parallel Gaussian BCs with N receivers. For $n = 1, \dots, N$ and $j = 1, \dots, K$, let $g_{n,j}$ be the power gain for $j = 1, 2, \dots K$ user n in channel j, and let π_j be a permutation over the set $\{1, 2, \dots, N\}$ such that $g_{\pi_j(i_1),j} > g_{\pi_j(i_2),j}$ if $i_1 < i_2$. According to [1], a rate vector $\mathbf{R} \triangleq (R_1, \dots, R_N)$ is achievable in the parallel Gaussian BC if we can write

$$P = \sum_{n=1}^{N}\sum_{j=1}^{K} p_{n,j} \quad \text{and} \quad R_n = \sum_{j=1}^{K} R_{n,j}$$

Manuscript received October 20, 2009; revised March 24, 2010, July 19, 2010, and September 27, 2010; accepted November 8, 2010. Date of publication December 3, 2010; date of current version February 18, 2011. This work was supported in part by the grants from the Research Grant Council of the Hong Kong Special Administrative Region, China under Project CityU 120107 and Project CUHK 415608. The review of this paper was coordinated by Prof. Y. Su.

C. Y. Ng and T. M. Lok are with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: michaelng@ieee.org; tmlok@ie.cuhk.edu.hk).

K. W. Shum is with the Institute of Network Coding, The Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: wkshum@inc.cuhk.edu.hk).

C. W. Sung is with the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong (e-mail: albert.sung@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2010.2096439

for some nonnegative $R_{n,j}$ and $p_{n,j}$ such that

$$R_{\pi_j(n),j} \leq \log_2 \left(1 + \frac{g_{\pi_j(n),j} \cdot p_{\pi_j(n),j}}{\sigma^2 + g_{\pi_j(n),j} \sum_{k=1}^{n-1} p_{\pi_j(k),j}} \right) \qquad (1)$$

for $n = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, K$. Let $\mathcal{C}$ be the set of all achievable rate vectors.

Our objective is to maximize a given joint utility function $U(\mathbf{R})$ over the capacity region $\mathcal{C}$. In the following, for two vectors $\mathbf{x}$ and $\mathbf{y}$, we will write $\mathbf{x} \preceq \mathbf{y}$ if each component of $\mathbf{x}$ is smaller than or equal to the corresponding component in $\mathbf{y}$. We assume that $U(\mathbf{R})$ is a function on $\mathbb{R}_+^N$ that is 1) strictly concave, 2) twice continuously differentiable, and 3) $U(\mathbf{R}) \leq U(\mathbf{R}')$ whenever $\mathbf{R} \preceq \mathbf{R}'$. A large set of commonly considered resource allocation problems can be modeled by a joint utility function satisfying the aforementioned assumptions. For example, the so-called α-fairness, where $\alpha > 0$, can be modeled by [11]

$$U(\mathbf{R}) = \begin{cases} \sum_{n=1}^N \ln(R_n), & \text{if } \alpha = 1 \\ \sum_{n=1}^N (1-\alpha) R_n^{1-\alpha}, & \text{otherwise.} \end{cases} \qquad (2)$$

The notion of α-fairness generalizes many widely considered fairness objectives. It reduces to proportional fairness and harmonic-mean fairness when $\alpha = 1, 2$, respectively. It also approximates max-min fairness when $\alpha \to \infty$.

We are going to present an efficient method, which can be applied not only to the parallel Gaussian BC but also to a general class of models, provided that the underlying rate region $\mathcal{C} \subseteq \mathbb{R}_+^N$ possesses four properties:

1) It is compact.
2) It is convex.
3) It is comprehensive.
4) Every Pareto-optimal rate vector is an extreme point.

We recall that a set in an Euclidean space is *compact* if it is bounded and closed and a set $\mathcal{C}$ in $\mathbb{R}_+^N$ is *comprehensive* if $\mathbf{x} \in \mathcal{C}$ implies that $\mathbf{y} \in \mathcal{C}$ for all $\mathbf{y} \preceq \mathbf{x}$. A point $\mathbf{x}$ is *Pareto-optimal* in $\mathcal{C}$ if $\mathbf{x} \preceq \mathbf{y} \in \mathcal{C}$ implies $\mathbf{x} = \mathbf{y}$. Furthermore, a point $\mathbf{x} \in \mathcal{C}$ is an *extreme point* of $\mathcal{C}$ if there is no way to express $\mathbf{x}$ as $\alpha \mathbf{y} + (1-\alpha)\mathbf{z}$ such that $\mathbf{y}, \mathbf{z} \in \mathcal{C}$, $\mathbf{y} \neq \mathbf{x} \neq \mathbf{z}$, and $0 < \alpha < 1$. It is straightforward to check that the capacity region of the parallel BC with distinct link gains satisfies all these properties.

III. Layered Optimization Framework

A. Dual Decomposition

The problem defined in the previous section is convex. We assume that strong duality holds, which is usually the case. (See [9] for more discussions on this assumption.) We are going to show how the dual problem can be decomposed into two layers. To do that, we first introduce an auxiliary rate vector $\tilde{\mathbf{R}} \triangleq (\tilde{R}_1, \tilde{R}_2, \ldots, \tilde{R}_N)$ and reformulate the problem as

$$\max \left\{ U(\tilde{\mathbf{R}}) : \tilde{\mathbf{R}} \preceq \mathbf{R}, \mathbf{R} \in \mathcal{C}, \tilde{\mathbf{R}} \in \mathcal{B} \right\}$$

where $\mathcal{B}$ is a rectangular box $\prod_{n=1}^N [0, b_n]$ containing $\mathcal{C}$ as a subset. Note that the bounding box $\mathcal{B}$ always exists since $\mathcal{C}$ is bounded. For example, for the parallel Gaussian BC, we may choose $b_n = \log_2(1 + \max_{1 \leq j \leq K} g_{n,j} P/\sigma^2)$, which is the upper bound of user n's rate.

The reformulated problem is indeed equivalent to the original problem. Note that a feasible $\tilde{\mathbf{R}}$ is within $\mathcal{C}$, because $\tilde{\mathbf{R}} \preceq \mathbf{R} \in \mathcal{C}$, and $\mathcal{C}$ is comprehensive. Since $\mathbf{R}$ can be any point in $\mathcal{C}$, so can $\tilde{\mathbf{R}}$. In other words, the feasible region of $\tilde{\mathbf{R}}$ is exactly equal to $\mathcal{C}$. Hence, a solution to the reformulated problem also solves the original problem.

Next, we relax the constraint $\mathbf{R} \succeq \tilde{\mathbf{R}}$ to form the *partial Lagrangian*, i.e.,

$$L(\mathbf{R}, \tilde{\mathbf{R}}, \boldsymbol{\mu}) \triangleq U(\tilde{\mathbf{R}}) + \sum_{n=1}^N \mu_n (R_n - \tilde{R}_n) \qquad (3)$$

where $\boldsymbol{\mu} \triangleq (\mu_1, \mu_2, \ldots, \mu_N) \succeq \mathbf{0}$ is the vector of Lagrange multipliers. Let the maximum of $L(\mathbf{R}, \tilde{\mathbf{R}}, \boldsymbol{\mu})$ over $\mathbf{R} \in \mathcal{C}$ and $\tilde{\mathbf{R}} \in \mathcal{B}$ be denoted by $q(\boldsymbol{\mu})$, which is called the *partial dual function*. The dual problem is to minimize $q(\boldsymbol{\mu})$ over all $\boldsymbol{\mu} \succeq \mathbf{0}$. Since $\tilde{\mathbf{R}}$ and $\mathbf{R}$ belong to two different sets, the calculation of $q(\boldsymbol{\mu})$ can be decoupled into two separate subproblems, i.e.,

$$q(\boldsymbol{\mu}) = \max_{\tilde{\mathbf{R}} \in \mathcal{B}} \left\{ U(\tilde{\mathbf{R}}) - \boldsymbol{\mu} \cdot \tilde{\mathbf{R}} \right\} + \max_{\mathbf{R} \in \mathcal{C}} \boldsymbol{\mu} \cdot \mathbf{R}. \qquad (4)$$

The first subproblem is the maximization of

$$U(\tilde{\mathbf{R}}) - \boldsymbol{\mu} \cdot \tilde{\mathbf{R}} \qquad (5)$$

over all $\tilde{\mathbf{R}} \in \mathcal{B}$. Since U is strictly concave and $\mathcal{B}$ is compact, a unique maximum exists. We denote the optimal solution to (5) for a given $\boldsymbol{\mu}$ by $\tilde{\mathbf{R}}^*(\boldsymbol{\mu}) \triangleq (\tilde{R}_1^*(\boldsymbol{\mu}), \ldots, \tilde{R}_N^*(\boldsymbol{\mu}))$. The variables $\tilde{R}_n$ are iteratively optimized in a round-robin manner. Let $\tilde{R}_n(\boldsymbol{\mu}, t)$ be the value of $\tilde{R}_n$ at iteration t. $\tilde{R}_n(\boldsymbol{\mu}, t)$ is obtained by

$$\tilde{R}_n(\boldsymbol{\mu}, t) = \arg \max_{0 \leq \tilde{r}_n \leq b_n} \Big[U\big(\tilde{R}_1(\boldsymbol{\mu}, t), \ldots, \tilde{R}_{n-1}(\boldsymbol{\mu}, t), \tilde{r}_n \\ \tilde{R}_{n+1}(\boldsymbol{\mu}, t-1), \ldots, \tilde{R}_N(\boldsymbol{\mu}, t-1)\big) - \mu_n \tilde{r}_n \Big]. \qquad (6)$$

By taking the partial derivative of (5), each Gauss–Seidel update of $\tilde{R}_n$ is done by solving the following equation:

$$\frac{\partial U(\tilde{\mathbf{R}})}{\partial \tilde{R}_n} = \mu_n. \qquad (7)$$

Since U is twice continuously differentiable and strictly concave, the left-hand side of (7) is a nonincreasing and continuous function of $\tilde{R}_n$. Hence, it can be efficiently solved by the bisection method for instance.

Proposition 1: For each fixed $\boldsymbol{\mu}$, the Gauss–Seidel-type algorithm in (6) converges to the optimal $\tilde{\mathbf{R}}^*(\boldsymbol{\mu})$.

Proof: We apply a convergence result in [12]. (See Theorem 6 in the Appendix.) The domain $\mathcal{B}$ is the Cartesian product of the compact sets $[0, b_1], [0, b_2], \ldots, [0, b_N]$. The objective function (5) is differentiable and strictly convex. Hence, it is pseudoconvex.[1] In addition, within the domain of $\tilde{\mathbf{R}}$, the level sets[2] of the objective function are compact. By Theorem 6, the sequence generated by the algorithm in (6) has limit points, and each limit point maximizes (5) over all $\tilde{\mathbf{R}} \in \mathcal{B}$. Since it is assumed that U is strictly concave, there is one and only one such limit point. Moreover, $U(\tilde{\mathbf{R}}(\boldsymbol{\mu}, t)) - \boldsymbol{\mu} \cdot \tilde{\mathbf{R}}(\boldsymbol{\mu}, t)$ is monotonically increasing as t increases. We see that the values converge to the maximal value. The convergence of the vectors $\tilde{\mathbf{R}}(\boldsymbol{\mu}, t)$ to the unique limit point is taken care of by the next lemma. ■

Lemma 2: Let $U(\mathbf{x})$ be a continuous function of $\mathbf{x}$ over a compact domain $\mathcal{X}$. Suppose that there is a unique optimal point $\mathbf{x}^*$, which maximizes $U(\mathbf{x})$ over all $\mathbf{x} \in \mathcal{X}$. If $\mathbf{x}_t$, $t = 1, 2, 3, \ldots$ is a sequence of points in $\mathcal{X}$ such that $\lim_{t \to \infty} U(\mathbf{x}_t) = U(\mathbf{x}^*)$, then $\mathbf{x}_t \to \mathbf{x}^*$ as $t \to \infty$.

[1] A differentiable function f is pseudoconvex if it satisfies the property that $\nabla f(x)(y - x) \geq 0$ implies $f(y) \geq f(x)$. Note that, if a function g is differentiable and convex, it is pseudoconvex.

[2] If $f(x)$ is a real function and γ is a scalar, the set $\{x | f(x) \leq \gamma\}$ is called a *level set of* f.

Proof: Let $\mathcal{G}$ be an arbitrarily open set in $\mathcal{X}$, which contains $\mathbf{x}^*$. For instance, $\mathcal{G}$ may be an open ball centered at $\mathbf{x}^*$. We want to show that $\mathbf{x}_t \in \mathcal{G}$ for all sufficiently large t. Consider the complement of $\mathcal{G}$ in $\mathcal{X}$ denoted by $\mathcal{G}^c$, which is a compact set, because it is a closed subset of the compact set $\mathcal{X}$. Hence, the function $U(\mathbf{x})$ attains a maximum over $\mathbf{x} \in \mathcal{G}^c$. Let the maximum be denoted by $M = \max_{\mathbf{x}\in\mathcal{G}^c} U(\mathbf{x})$. The value of M is strictly less than $U(\mathbf{x}^*)$, because $\mathbf{x}^* \notin \mathcal{G}^c$. Because of the assumption that $\lim_{t\to\infty} U(\mathbf{x}_t) = U(\mathbf{x}^*)$ for all sufficiently large t, $U(\mathbf{x}_t)$ is strictly larger than M, and hence, $\mathbf{x}_t$ cannot be in $\mathcal{G}^c$. This proves that, for all sufficiently large t, $\mathbf{x}_t \in \mathcal{G}$. ■

The second subproblem is the maximization of $\boldsymbol{\mu}\cdot\mathbf{R}$ over $\mathbf{R}\in\mathcal{C}$. This is the weighted sum rate maximization problem. Let $\mathbf{R}^*(\boldsymbol{\mu}) = (R_1^*(\boldsymbol{\mu}),\ldots,R_N^*(\boldsymbol{\mu}))$ be the unique optimal solution for a given $\boldsymbol{\mu}$. The uniqueness follows from assumption 4 on the region $\mathcal{C}$. It is assumed that an algorithm for the second subproblem is already available.

The idea of the overall algorithm is to search over the boundary of $\mathcal{C}$ by varying the weight vector $\boldsymbol{\mu}$. By the proper separation theorem in [13, Prop. 2.4.5], we can recover all points on the surface of the rate region $\mathcal{C}$ by running the weighted sum maximization algorithm with different weights, provided that every Pareto-optimal rate vector in $\mathcal{C}$ is an extreme point. The remaining question is how to search for the optimal $\boldsymbol{\mu}$, which is answered in the next section.

B. Iterative Optimization of Dual Variables

The proposed algorithm is a nonlinear Gauss–Seidel algorithm that maximizes the joint utility function by adjusting the dual variable $\boldsymbol{\mu}$ in a round-robin fashion. The algorithm computes a sequence of dual variables $\boldsymbol{\mu}(t)$, $t = 1, 2, \ldots$, which is recursively defined by

$$\mu_n(t+1) = \arg\min_{\xi\geq 0} q\,(\mu_1(t+1),\ldots \\ \mu_{n-1}(t+1),\xi,\mu_{n+1}(t),\ldots,\mu_N(t))\,. \quad (8)$$

By construction, the value of the partial dual function $q(\boldsymbol{\mu})$ decreases as we run the algorithm in (8). However, it does not imply that the value of the joint utility function is increasing. We recursively define another sequence of rate vector $\mathbf{R}_{\max}(t)$ by $\mathbf{R}_{\max}(1) = \mathbf{R}^*(\boldsymbol{\mu}(1))$, and for $t > 1$, $\mathbf{R}_{\max}(t)$

$$= \begin{cases} \mathbf{R}^*\left(\boldsymbol{\mu}(t)\right), & \text{if } U\left(\mathbf{R}^*\left(\boldsymbol{\mu}(t)\right)\right) > U\left(\mathbf{R}_{\max}(t-1)\right) \\ \mathbf{R}_{\max}(t-1), & \text{if } U\left(\mathbf{R}^*\left(\boldsymbol{\mu}(t)\right)\right) \leq U\left(\mathbf{R}_{\max}(t-1)\right). \end{cases}$$

$U(\mathbf{R}_{\max}(t))$ records the largest utility value up to iteration t.

The main result in this section is given here.

Theorem 3: The argmin in (8) exists and is finite. The sequence of rate vectors $\{\mathbf{R}_{\max}(t)\}$ converges, and the sequence of utility values $\{U(\mathbf{R}_{\max}(t))\}$ converges to the maximum value of $U(\mathbf{R})$ over all $\mathbf{R}\in\mathcal{C}$.

To prove Theorem 3, we need the following properties of the partial dual function;

Proposition 4: $q(\boldsymbol{\mu})$ is convex continuously differentiable with partial derivative given by

$$\frac{\partial q(\boldsymbol{\mu})}{\partial \mu_n} = -\tilde{R}_n^*(\boldsymbol{\mu}) + R_n^*(\boldsymbol{\mu}). \quad (9)$$

Proof: The fact that $q(\boldsymbol{\mu})$ is convex follows from the basic result that the pointwise maximum of affine functions is convex [14, Sec. III-B.3]. The property that $q(\boldsymbol{\mu})$ is differentiable and that the partial derivative is given as in (9) are consequences of Danskin's theorem. Since the partial Lagrangian L is differentiable, by Danskin's theorem, q is differentiable. Furthermore, since $\partial L/\partial\mu_n = -\tilde{R}_n + R_n$, by Danskin's theorem again, $\partial q/\partial\mu_n$ is given by (9). By the compactness of the domains of $\tilde{\mathbf{R}}$ and $\mathbf{R}$, i.e., $\mathcal{B}$ and $\mathcal{C}$, both $\tilde{R}_n^*(\boldsymbol{\mu})$ and $R_n^*(\boldsymbol{\mu})$ are continuous by Berge's maximal theorem. Hence, $-\tilde{R}_n^*(\boldsymbol{\mu}) + R_n^*(\boldsymbol{\mu})$ is continuous, implying that $q(\boldsymbol{\mu})$ is continuously differentiable. We refer to the Appendix for the statements of Danskin's and Berge's theorem. ■

Next, we want to investigate $\tilde{\mathbf{R}}^*(\boldsymbol{\mu})$ and $\mathbf{R}^*(\boldsymbol{\mu})$ as functions of a single component of $\boldsymbol{\mu}$, keeping the other components fixed. We introduce the notation $\boldsymbol{\mu}_{-n}$ to represent the vector $\boldsymbol{\mu}$ without the component μ_n and write $\tilde{\mathbf{R}}^*(\boldsymbol{\mu}) = \tilde{\mathbf{R}}^*(\mu_n, \boldsymbol{\mu}_{-n})$ and $\mathbf{R}^*(\boldsymbol{\mu}) = \mathbf{R}^*(\mu_n, \boldsymbol{\mu}_{-n})$ to emphasize that μ_n is the variable, whereas $\boldsymbol{\mu}_{-n}$ is fixed.

Proposition 5: For any $n = 1, 2, \ldots, N$, given any fixed $\boldsymbol{\mu}_{-n}$, $R_n^*(\mu_n, \boldsymbol{\mu}_{-n})$ is an increasing function of μ_n, and $\tilde{R}_n^*(\mu_n, \boldsymbol{\mu}_{-n})$ is a decreasing function of μ_n. In particular, the partial derivative of $q(\mu_n, \boldsymbol{\mu}_{-n})$ with respect to μ_n is an increasing function of μ_n.

Proof: These monotonic properties readily follow from the definitions of $\mathbf{R}(\boldsymbol{\mu})$ and $\tilde{\mathbf{R}}(\boldsymbol{\mu})$. Details of the proof are omitted. ■

To compute the minimum in (8), we search for μ_n such that $\partial q/\partial\mu_n$ is equal to zero. As the partial derivative is monotonically increasing, this can be done by the bisection method for instance.

Proof of Theorem 3: For the convergence, we apply Theorem 6 in the Appendix, with $f(\boldsymbol{\mu})$ equal to $q(\boldsymbol{\mu})$. We can easily check that condition (a) in Theorem 6 is satisfied by Proposition 4. For condition (b), we need to show that, for any constant c, the level set $L_c := \{\boldsymbol{\mu} : q(\boldsymbol{\mu}) \leq c\}$ is compact. Since q is a continuous function of $\boldsymbol{\mu}$, L_c is a closed set. For boundedness, we will show that L_c is contained in a bounded set of the form $\{\boldsymbol{\mu}\in\mathbb{R}_+^N : \sum_n \mu_n \leq M\}$, where M is a sufficiently large constant.

Let $\mathcal{H}_1$ be the simplex in $\mathbb{R}_+^N$ consisting of all vectors with the sum of components equal to 1. Let $\min_{\boldsymbol{\mu}\in\mathcal{H}_1}\max_{\mathbf{R}\in\mathcal{C}} \boldsymbol{\mu}\cdot\mathbf{R}$ be denoted by m. The minimum is well defined, because $\max_{\mathbf{R}\in\mathcal{C}} \boldsymbol{\mu}\cdot\mathbf{R}$ is a continuous function of $\boldsymbol{\mu}$ (by Berge's maximum theorem) and $\mathcal{H}_1$ is compact. Note that m is a positive constant. It is zero only if all points in $\mathcal{C}$ have one or more components identically zero. This degenerate case is not of practical interest and can be excluded from consideration. Let $\mathbf{1}$ be the N-dimensional all-one vector. Pick an $\epsilon > 0$, so that ϵ is strictly less than m and $\epsilon\mathbf{1}\in\mathcal{C}$. Then, for any $\boldsymbol{\mu}\succeq\mathbf{0}$, we have $q(\boldsymbol{\mu}) \geq U(\epsilon\mathbf{1}) - \boldsymbol{\mu}\cdot(\epsilon\mathbf{1}) + \max_{\mathbf{R}\in\mathcal{C}}\boldsymbol{\mu}\cdot\mathbf{R}$. Here, we have used the fact that $\epsilon\mathbf{1}\in\mathcal{C}\subseteq\mathcal{B}$. Next, we use the property that $\max_{\mathbf{R}\in\mathcal{C}}\boldsymbol{\mu}\cdot\mathbf{R}$ is a homogeneous function of $\boldsymbol{\mu}$ of degree 1 and obtain

$$q(\boldsymbol{\mu}) \geq U(\epsilon\mathbf{1}) - \boldsymbol{\mu}\cdot(\epsilon\mathbf{1}) + \left(\sum_n \mu_n\right)\max_{\mathbf{R}\in\mathcal{C}}\frac{\boldsymbol{\mu}}{\sum_n\mu_n}\cdot\mathbf{R}$$
$$\geq U(\epsilon\mathbf{1}) + [-\epsilon + m]\sum_n \mu_n. \quad (10)$$

for nonzero $\boldsymbol{\mu}$. The value within the square bracket in (10) is positive by our choice of ϵ. Given an arbitrary constant c, there is a sufficiently large M such that $q(\boldsymbol{\mu})$ is strictly larger than c whenever $\sum_n \mu_n$ is larger than M. This proves that $q(\boldsymbol{\mu}) \leq c$ implies $\sum_n \mu_n \leq M$ for some sufficiently large M. Hence, L_c is bounded.

As a result, the minimum in (8) can be taken over a compact set, instead of over $\xi \geq 0$. This proves that the arg min in (8) exists and is finite.

By Theorem 6, $\inf_t q(\boldsymbol{\mu}(t))$ is equal to the minimum value, e.g., v^*, of $q(\boldsymbol{\mu})$ over all $\boldsymbol{\mu}\succeq\mathbf{0}$. By the standard duality theory in convex analysis, $\sup_t U(\mathbf{R}^*(\boldsymbol{\mu}(t)))$ is also equal to v^* and is the maximum value of the primal problem. Since U is strictly concave, v^* is achieved by a unique point, e.g., $\bar{\mathbf{R}}$. The sequence $U(\mathbf{R}_{\max}(t))$ is monotonically increasing and converges to v^*. By Lemma 2, the sequence $\mathbf{R}_{\max}(t)$ converges to $\bar{\mathbf{R}}$ as t tends to infinity. ■

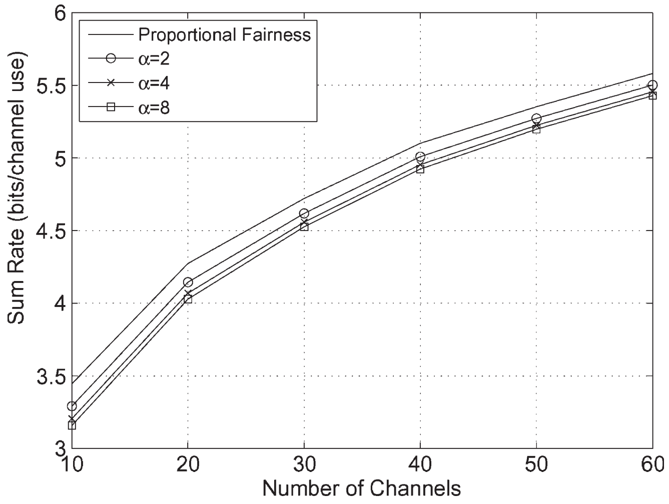

Fig. 1. Sum rate of ten users (Total power $= 1$).

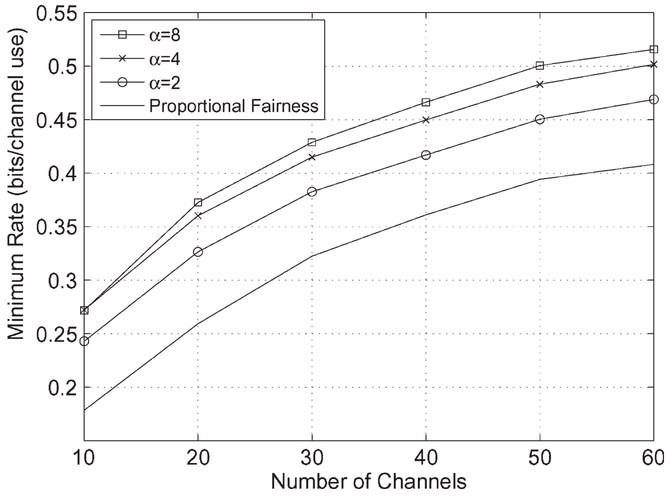

Fig. 2. Minimum rate of ten users (Total power $= 1$).

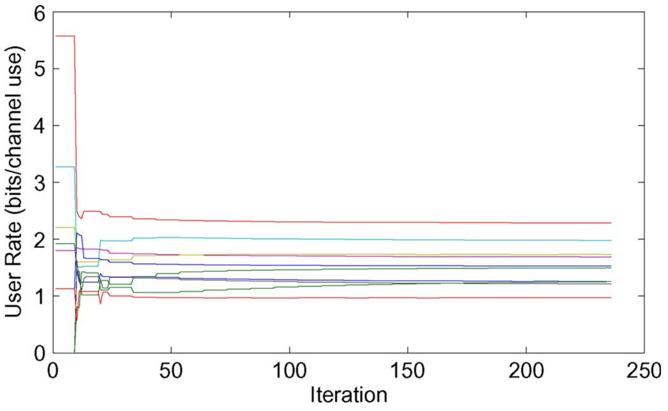

Fig. 3. Rate of each user $\mathbf{R}^*(\boldsymbol{\mu}(t))$ in each iteration of our proposed algorithm.

IV. Numerical Examples

We consider the α-fairness problem for the parallel Gaussian BC with ten users, and the number of channels ranges from 10 to 60. Both the total transmission power and the variance of the additive white Gaussian noise at each channel of each receiver are assumed to be one. The power gain of each user in each channel is an independent exponential random variable with mean one, which is a typical Rayleigh fading scenario. The simulation results for sum rate and minimum rate performance of the users are plotted in Figs. 1 and 2, respectively. Each point is obtained by averaging more than 500 simulation runs. From these graphs, we can see that the sum rate decreases with α, whereas the minimum rate increases with α, revealing that a larger value for α improves fairness in the sense that it reduces rate difference among users, at the expense of lower sum rate.

Next, we compare the convergence of our algorithm with the subgradient approach in [10]. The same system model as previously mentioned with ten users and ten parallel channels is used, and the proportional fairness problem is considered. In our approach, bisection search is used to search for the arg min in (8). We initialize the close interval of the bisection search of μ_n to be $[0, \mu_{\max}]$, with $\mu_{\max} = 100$. The upper bound $\mu_{\max}$ is doubled until $\partial q/\partial \mu_n(\mu_{\max}, \boldsymbol{\mu}_{-n})$ is larger than 0.001. The bisection search terminates when the partial derivative $\partial q/\partial \mu_n$ is less than 0.001 in magnitude. In the subgradient approach, we initialize all dual variables μ_n to one and adopt *diminishing step size*. Let $a(t)$ be the step size at the tth iteration. In the simulations, $a(t)$ is chosen as $1/\sqrt{t}$. It can be proven that $\lim_{t\to\infty} a(t) = 0$ and $\sum_{t=1}^{\infty} a(t) = \infty$, which are the requirements of diminishing step sizes.

The evolution of user rates obtained by the proposed algorithm is compared with those obtained by the subgradient approach. We execute both algorithms over many simulation runs and find that the evolution curves are all similar. Therefore, here we only consider one typical simulation instance. The user rates yielded by our algorithm and the subgradient approach are plotted in Figs. 3 and 4, respectively. Notice that, in the implementation of our algorithm, $\mathbf{R}_{\max}(t)$ is used as the user rates, but in the subgradient approach, the user rates are given by $\mathbf{R}^*(\boldsymbol{\mu}(t))$. In our algorithm, one iteration means a Gauss–Seidel update of one μ_n, whereas in the subgradient approach, one iteration means a subgradient update of the vector $\boldsymbol{\mu}$, which means all μ_n's are updated in one iteration. In our algorithm, each iteration requires a bisection search for an optimal μ_n, and each iteration of a bisection search requires one execution of the weighted sum rate maximization algorithm. In the subgradient approach, each subgradient update of $\boldsymbol{\mu}$ only needs one execution of the weighted sum rate maximization algorithm. To have a fair comparison of the efficiency of both algorithms, we need the distribution of the number of iterations of bisection search, which is plotted in Fig. 5. In addition, we compare the execution time of both algorithms by comparing the number of executions of the weighted sum rate maximization algorithm. The number of executions of our algorithm is given by the product of the results implied by Figs. 3 and 5.

As shown in Fig. 5, most of the time, the bisection search requires fewer than 20 iterations. Fig. 3 indicates that convergence occurs somewhere between the 150th and the 200th iteration. Combining the results, we find that the number of executions of the weighted sum maximization algorithm is in the range of 3000–4000. Fig. 4 indicates that the subgradient approach requires about 5000 executions, which is less efficient. Moreover, our approach provides a flexible tradeoff between running time and solution accuracy. It can be observed that the user rates provided by our algorithm have much smaller fluctuations before convergence. It appears to be stable after about 50 iterations or, equivalently, about 1000 executions of weighted sum rate maximization. In other words, our algorithm can be stopped any time after that, without affecting solution quality too much. It can thus be applied to channels with shorter coherence time at the expense of slight decrease in solution accuracy. Indeed, the minimum user rate provided by the subgradient approach is still zero bit per channel

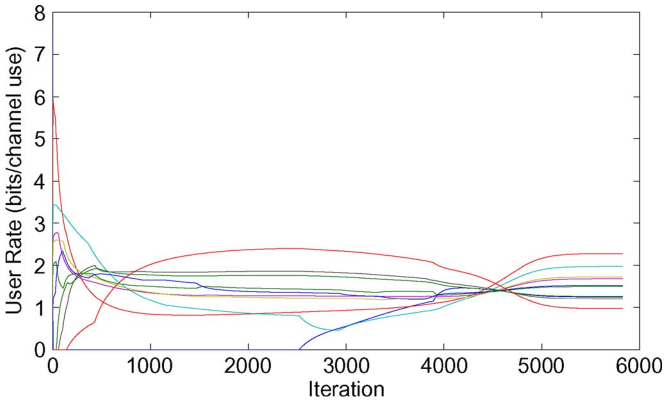

Fig. 4. Rate of each user $\mathbf{R}^*(\boldsymbol{\mu}(t))$ in each iteration of the dual decomposition approach in [10].

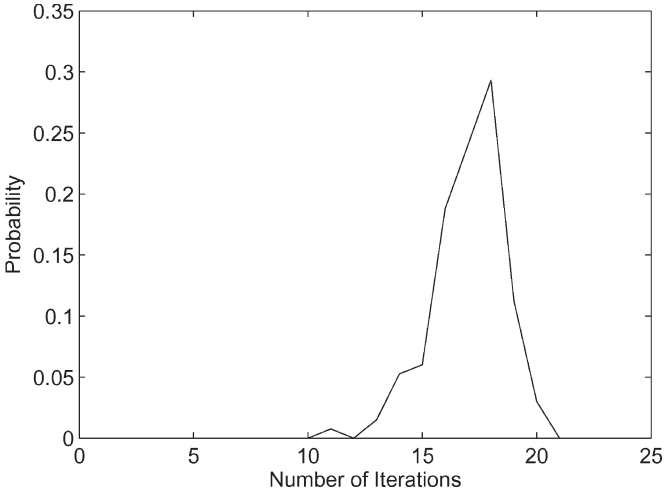

Fig. 5. Distribution of the number of iterations of each bisection search.

use, even after 2500 executions. Our algorithm clearly improves the minimum-rate performance more rapidly.

V. Conclusion

We have proposed a framework for joint utility maximization and shown that the problem can be numerically solved by decomposition into two layers. Our work thus provides a bridge to link the physical-layer issue and the upper layer issue, which facilitates a simple and efficient way to allocate resources. The effectiveness of our proposed framework has been demonstrated by application to the parallel Gaussian BC. The proposed method is a centralized algorithm and can serve as a benchmark for other distributed algorithms such as the subgradient method.

Appendix

We provide the statements of the three theorems used in the main text.

Theorem 6 (Convergence of Gauss–Seidel Algorithm [12, Prop. 6]): Suppose that $\mathcal{X} \subseteq \mathbb{R}^N$ is the Cartesian product $\prod_{n=1}^N \mathcal{X}_n$, with $\mathcal{X}_n \subseteq \mathbb{R}$, where $\mathcal{X}_n \subset \mathbb{R}$ is a nonempty closed and convex set. Let $f(\mu_1, \ldots, \mu_N)$ be a real-valued function such that it is pseudoconvex on $\mathcal{X}$, and the level set $\{\boldsymbol{\mu} \in \mathcal{X} : f(\boldsymbol{\mu}) \leq f(\boldsymbol{\mu}_0)\}$ is compact for any $\boldsymbol{\mu}_0 \in \mathcal{X}$. Then, the sequence generated by the Gauss–Seidel algorithm

$$\mu_n(t+1) = \arg \min_{\xi \in \mathcal{X}_n} f\left(\mu_1(t+1), \ldots \right.$$
$$\left. \mu_{n-1}(t+1), \xi, \mu_{n+1}(t), \ldots, \mu_N(t)\right)$$

has limit points, and each limit point is a global minimizers f.

Theorem 7 (Danskin [15]): Let $\mathcal{X} \in \mathbb{R}^N$, $\mathcal{Y} \subset \mathbb{R}^M$, and $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function, such that, for each $\mathbf{y} \in \mathcal{Y}$, the function $\phi(\cdot, \mathbf{y})$ is convex in the first argument. Suppose that, for each $\mathbf{x} \in \mathcal{X}$, the maximum $\max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$ is finite and attained at a unique point, which is denoted by $\mathbf{y}_\mathbf{x}$. Then, the maximum value function $f(\mathbf{x}) = \phi(\mathbf{x}, \mathbf{y}_\mathbf{x})$ is convex. Furthermore, if ϕ is differentiable, then f is differentiable, and

$$\frac{\partial f}{\partial x_n}(\mathbf{x}_0) = \frac{\partial \phi}{\partial x_n}(\mathbf{x}_0, \mathbf{y}_{\mathbf{x}_0})$$

for $n = 1, 2, \ldots, N$. The right-hand side in the preceding equation is the partial derivative of ϕ with respect to x_n evaluated at $\mathbf{x} = \mathbf{x}_0$ and $\mathbf{y} = \mathbf{y}_{\mathbf{x}_0}$.

Proof of Danskin's theorem can be found in [13, p. 245].

Theorem 8 (Special Case of Berge's Maximum Theorem [16, p. 116]): Using the notation as in Danskin's theorem, if $\mathcal{Y} \subset \mathbb{R}^M$ is a compact set, then both $\phi(\mathbf{x}, \mathbf{y}_\mathbf{x})$ and $\mathbf{y}_\mathbf{x}$ are continuous functions.

References

[1] D. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," in *Proc. IEEE ISIT*, Jun. 1997, p. 27.

[2] G. Wunder and T. Michel, "Optimal resource allocation for parallel Gaussian broadcast channels: Minimum rate constraints and sum power minimization," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4817–4822, Dec. 2007.

[3] K. Seong, D. D. Yu, Y. Kim, and J. M. Cioffi, "Optimal resource allocation via geometric programming for OFDM broadcast and multiple access channels," in *Proc. IEEE GLOBECOM*, Dec. 2006, pp. 4817–4822.

[4] F. Kelly, "Charging and rate control for elastic traffic," *Eur. Trans. Telecommun.*, vol. 8, no. 1, pp. 33–37, Jan./Feb. 1997.

[5] C. W. Sung, K. W. Shum, and C. Y. Ng, "Fair resource allocation for the Gaussian broadcast channel with ISI," *IEEE Trans. Commun.*, vol. 57, no. 5, pp. 1381–1389, May 2009.

[6] D. Tse and S. Hanly, "Multi-access fading channels—Part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2796–2815, Nov. 1998.

[7] C. Y. Ng, K. W. Shum, C. W. Sung, and T. M. Lok, "Rate allocation for cooperative orthogonal-division channels with dirty-paper coding," *IEEE Trans. Commun.*, vol. 58, no. 10, pp. 2949–2959, Oct. 2010.

[8] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

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

[10] J. Brehmer and W. Utschick, "A decomposition of the downlink utility maximization problem," in *Proc. 41st Annu. CISS*, 2007, pp. 437–441.

[11] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[12] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss-Seidel method under convex constraints," *Oper. Res. Lett.*, vol. 26, no. 3, pp. 127–136, Apr. 2000.

[13] D. P. Bertsekas, A. Nedi, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA: Athena Scientific, 2003.

[14] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[15] J. M. Danskin, *Theory of Max-Min and Its Application in Weapon Allocation Problems*. New York: Springer-Verlag, 1967.

[16] C. Berge, *Topological Spaces*. New York: Dover, 1997.