

# Multi-Cell MIMO Downlink With Cell Cooperation and Fair Scheduling: A Large-System Limit Analysis

Hoon Huh, *Member, IEEE*, Sung-Hyun Moon, *Student Member, IEEE*, Young-Tae Kim, *Student Member, IEEE*, Inkyu Lee, *Senior Member, IEEE*, and Giuseppe Caire, *Fellow, IEEE*

**Abstract**—We consider the downlink of a cellular network with multiple cells and multi-antenna base stations. Our model includes distance-dependent pathloss, arbitrary clusters of cooperating cells, and general “fairness” requirements. Beyond Monte Carlo simulation, no efficient computation method to evaluate the ergodic throughput of such systems has been presented, yet. Furthermore, for systems of practical size with tens of cells and hundreds of users per cell, even simulation becomes challenging. We develop an analytic framework based on the combination of results from large random matrix theory and convex optimization. This allows computationally efficient calculation of the system performance in the so-called “large system limit”, i.e., in the limit of a large number of antennas per base station and a large number of users per cell, while the ratio of antennas per user is kept constant. In particular, the system ergodic throughput, subject to per-base station power constraints and to general fairness criteria, is obtained via the iterative solution of a system of fixed-point equations. Comparisons with finite-dimensional simulation results show that the large-system analysis provides remarkably accurate approximations for the actual finite-dimensional systems, even for a small number of users and base station antennas.

**Index Terms**—Asymptotic analysis, fairness scheduling, inter-cell cooperation, large-system limit, multicell MIMO downlink, weighted sum rate maximization.

## I. INTRODUCTION

MULTIUSER MIMO (MU-MIMO) technology is expected to play a key role in future wireless cellular networks [1], [2]. The MIMO Gaussian Broadcast Channel (BC) model [3]–[8] serves as the information theoretic foundation for various MU-MIMO downlink schemes. In particular, the MIMO Gaussian BC capacity region with zero common message rate was characterized in [8] where the optimality of Gaussian Dirty-Paper Coding (DPC) is shown, subject to a general convex input covariance constraint.

Manuscript received June 10, 2010; revised May 02, 2011; accepted July 14, 2011. Date of current version December 07, 2011. The work of G. Caire and H. Huh was supported in part by the National Science Foundation under Grant CCF 0917343. The work of I. Lee was supported by the National Research Foundation of Korea (NRF) under a grant funded by the Korean government (MEST) (2010-0017909).

H. Huh and G. Caire are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: hhuh@usc.edu; caire@usc.edu).

S.-H. Moon, Y.-T. Kim, and I. Lee are with the School of Electrical Engineering, Korea University, Seoul, Korea (e-mail: shmoon@korea.ac.kr; reftm@korea.ac.kr; inkyu@korea.ac.kr).

Communicated by G. Taricco, Associate Editor for Communications.

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/TIT.2011.2170123

In a multicell scenario, depending on the level of inter-cell cooperation, we are in the presence of a MIMO broadcast and interference channel, which is not yet fully understood in an information theoretic sense. A simple and analytically tractable model for a multicell system with full inter-cell cooperation was introduced by Wyner [9] for the uplink, and extended by uplink-downlink duality [6] to the downlink in [10]. Wyner model considers either a 1-D linear cellular array and a 2-D hexagonal cellular pattern, where only interference from adjacent cells is accounted for, and where the channel path coefficients are either 1 (between a base-station and the users inside its cell) or  $\alpha \in [0, 1]$  (between a base-station and the users in the nearest adjacent cells). In Wyner’s model, the system capacity takes on an elegant single-letter closed-form expression in the case of full joint processing of all cells and no fading. Wyner’s setting was extended in several works [10]–[14]. Single cell processing and joint two-cell processing were investigated by treating the inter-cell interference as Gaussian noise in [11] and the flat-fading channel case with full joint cell processing was treated in [12]. This model was modified and extended to take into account various issues such as soft hand-off and limited inter-cell cooperation due to constrained backhaul capacity (see [10], [13], [14] and references therein).

As often happens in science and engineering, there is a tradeoff between the model simplicity, the possibility of obtaining simple closed-form analysis, and the relevance of the analysis results to cases of practical interest. Even restricting to Wyner’s two-valued channel path coefficients model, the simplicity of the original Wyner’s result disappears when including fading and/or some forms of partial inter-cell cooperation, as it can be seen from the above referenced literature. Wyner’s model suffers from two major shortcomings, that somehow limit the relevance of the system design conclusions based on its analysis: 1) The restriction of interference to adjacent cells only leads to the conclusion that even a simple reuse scheme without cell joint processing yields an interference unlimited system. For example, in the linear cellular layout, we can activate all even-numbered cells at even time-slots and odd-numbered cells at odd time slots, and obtain a system with zero inter-cell interference even without cell cooperation. 2) The restricting of the path coefficients to the values  $\{1, \alpha\}$ , as said before, yields a symmetric system with respect to any user. Therefore, maximizing the sum rate is equivalent to maximizing each individual user rate, since every user achieves by symmetry an equal share of the total rate.

More realistically, the interference from nonadjacent cells is a very important limiting factor of the system performance,

especially when cell cooperation is limited to clusters of size much smaller than the whole system due to practical implementation complexity, as currently considered in various “Coordinated Multi-Point” (CoMP) schemes in the standardization of fourth generation cellular systems [15]. Furthermore, users in different locations of the cellular coverage region are subject to distance-dependent pathloss that may have more than 30 dB of dynamic range [16]. Therefore, the system is fundamentally asymmetric with respect to different users. This asymmetry with respect to different users locations yields that the system sum-capacity (or achievable sum-throughput, under some sub-optimal scheme) fails to capture the actual system performance. For example, the sum-throughput may be maximized by allocating zero power and rate to users in disadvantaged locations (cell or cluster edges), and focusing the system resource only on the users in the neighborhood of the base-stations. Such system performance would be unacceptable, since it would imply that users in bad locations aren’t served for a long time. Instead, actual systems must serve *all* users by sharing their resources according to some fairness criterion. Therefore, the characterization of the system performance must take into account some appropriate notion of *fairness*.

The problem of *downlink scheduling* subject to some fairness criterion has been widely studied (see for example [17]–[19] and references therein). The goal of fairness scheduling is to make the system operate at some point of its ergodic achievable rate region, such that a suitable concave and componentwise non-decreasing *network utility function* is maximized [20]. By choosing the shape of the network utility function, a desired fairness criterion can be enforced. The framework of *stochastic network optimization* [18] can be leveraged in order to systematically devise scheduling algorithms that perform arbitrarily close to the optimal achievable fairness point, even when the explicit computation of the achievable ergodic rate region is hopelessly complicated. The fairness operating point is given as the time-averaged rate obtained by applying a dynamic scheduling algorithm on a slot-by-slot basis. Hence, its analytical characterization is generally very difficult and the system performance is typically evaluated by letting the scheduling algorithm evolve in time and computing the time-averaged rates by Monte Carlo simulation [21]–[31].

In this paper, we seek a mid-way approach in between the extreme simplification of the problem (as in Wyner’s model) and the full-scale system simulation, as in the works referenced above. For this purpose, we develop a “large-system limit” analysis. We leverage results on large random matrices [32]–[36], in order to characterize the system achievable rate region in the limit where both the number of antennas per base-station and the number of users per cell grow to infinity with a fixed ratio. Our model encompasses arbitrary user locations and distance-dependent pathloss and considers arbitrary inter-cell cooperation clusters, where the base-stations in the same cluster operate as a distributed antenna array (intra-cluster full cooperation) and inter-cluster interference is treated as Gaussian noise (no inter-cluster cooperation).<sup>1</sup> As special cases, we recover

conventional cellular systems (no inter-cell cooperation) and the case of full cooperation (all base-stations in a single cluster). In the large-system limit, the channel randomness disappears and the MU-MIMO system becomes a deterministic network. It follows that the performance of dynamic fairness scheduling can be calculated by solving a “static” convex optimization problem. By incorporating the large random matrix results into the convex optimization solution, we obtain the system capacity subject to any desired concave and componentwise non-decreasing *network utility function* of the users’ ergodic rates. The solution is obtained by solving, iteratively, a system of fixed-point equations. We also show that the solution becomes particularly simple when each cooperation cluster satisfies certain symmetry conditions, discussed later on. The proposed method is much more efficient than Monte Carlo simulation and, not surprisingly, it provides a very accurate approximation for the performance of finite-dimensional systems, even for small dimension. The accuracy of the large-system analysis with respect to the corresponding finite-dimensional systems is a well-known phenomenon, widely studied in the simpler case of single-user point-to-point MIMO channels [39], and depends on the fact that the quantities of interest in the analysis, namely, the log-determinant of signal plus noise covariance matrices, converge very rapidly to their asymptotic limit as the system dimensions grow.

The remainder of this paper is organized as follows. In Section II, we present the MU-MIMO downlink system model with inter-cell cooperation and formulate the fairness scheduling problem. We develop the numerical solution for the input covariance maximizing the weighted average sum rate in the large-system limit in Section III. In Section IV, we use these results in order to obtain a semi-analytic method to calculate the optimal ergodic fairness rate point in the asymptotic regime. In Section V, the asymptotic rates are shown in 2-cell linear and 7-cell planar models and are compared with finite-dimensional simulation results obtained by the combination of DPC and the actual dynamic scheduling scheme based on stochastic optimization. Concluding remarks are presented in Section VI.

## II. PROBLEM SETUP

We consider a cellular system formed by  $M$  base-stations (BSs), each equipped with multiple antennas. Each cell contains single-antenna user terminals. For the sake of the large-system analysis developed in this work, we shall scale the number of antennas per BS and the number of users per cell to infinity, such that their ratio remains constant. Furthermore, for analytical simplicity, we consider a spatial discretization of the users positions, such that the whole users population is partitioned into  $K$  co-located groups of  $N$  users each. When we scale the system to its large-system limit, we let the parameter  $N$  (number of users in each user group) go to infinity. We let  $\gamma > 0$  indicate the ratio between the number of BS antennas and the number of users in each group, therefore, the number of BS antennas is equal to  $\gamma N$ .

Users in the same group are statistically equivalent: since they are co-located, they experience the same set of pathloss coefficients with respect to the BSs, and their small-scale fading

<sup>1</sup>Although not considered in this work, we should mention that inter-cluster interference can also be handled by using *Interference Alignment* [37], [38].

channel coefficients are mutually independent and identically-distributed (i.i.d.). In practice, it is reasonable to assume that co-located users are separated by a sufficient number of wavelength such that they undergo i.i.d. small-scale fading (due to local scattering effects and small movements of the users around their nominal position), but the wavelength is sufficiently small so that they all have essentially the same distance-dependent pathloss from any given BS. Users in different groups observe generally different sets of pathlosses, depending on their relative positions with respect to the BSs. We remark here that the assumption of  $K$  co-located user groups is inessential for our analysis to hold, but it significantly simplifies the fixed-point equations that we have to solve iteratively to obtain the desired results (see Sections III-B and III-C). It is possible to extend our results to continuous user spatial distributions (e.g., uniform over each cell). However, in order to solve the resulting fixed-point equations such as (22), we have to discretize the domain of certain integrals, and therefore in practice we go back to the discretized case. Here, we avoid this complication by assuming a discrete user distribution from the very beginning.

We assume a block-fading model where the channel coefficients are constant over time-frequency “slots” determined by the channel coherence time and bandwidth, and change according to some well-defined ergodic process from slot to slot. In contrast, the distance-dependent pathloss coefficients are constant in time. This is representative of a typical situation where the distance between BSs and users changes significantly over a time-scale of the order of tens of seconds, while the small-scale fading decorrelates completely within a few milliseconds [40]. The slot index shall be denoted by  $t$ , but we will omit  $t$  for notation simplicity whenever possible. We shall make explicit reference to the time slot when discussing the dynamic fairness scheduling policy in Section IV.

One channel use of the multicell MU-MIMO downlink is described by

$$\mathbf{y}_k = \sum_{m=1}^M \alpha_{m,k} \mathbf{H}_{m,k}^H \mathbf{x}_m + \mathbf{n}_k, \quad k = 1, \dots, K \quad (1)$$

where  $\mathbf{y}_k = [y_{k,1} \dots y_{k,N}]^T \in \mathbb{C}^N$  denotes the received signal vector for the  $k$ -th user group,  $\alpha_{m,k}$  and  $\mathbf{H}_{m,k}$  denote the distance-dependent pathloss and a  $\gamma N \times N$  channel matrix collecting the small-scale channel fading coefficients from the  $m$ -th BS to the  $k$ -th user group, respectively,  $\mathbf{x}_m = [x_{m,1} \dots x_{m,\gamma N}]^T \in \mathbb{C}^{\gamma N}$  is the signal vector transmitted by the  $m$ -th BS, and  $\mathbf{n}_k = [n_{k,1} \dots n_{k,N}]^T \in \mathbb{C}^N$  denotes the AWGN at the user receivers in the  $k$ -th user group. The elements of  $\mathbf{n}_k$  and of  $\mathbf{H}_{m,k}$  are i.i.d.  $\sim \mathcal{CN}(0, 1)$ . We assume a per-BS average power constraint expressed by  $\text{tr}(\text{Cov}(\mathbf{x}_m)) \leq P_m$ , where  $P_m > 0$  denotes the total transmit power of the  $m$ -th BS.

We consider a multicell cooperative system where cells are partitioned into cooperation clusters. Each cluster acts effectively as a distributed MU-MIMO system, by jointly processing all the antennas for all BSs in the cluster, and by serving the users in the groups located in the cells forming the cluster. At each user receiver in a given cluster, the interference coming from other clusters, referred to as Inter-Cluster Interference

(ICI), is treated as additional Gaussian noise. Since the clusters form a *partition* of the cells set, no user group can be served by more than one cluster.<sup>2</sup>

Let  $L$  denote the number of cooperation clusters. We define the partition  $\{\mathcal{M}_1, \dots, \mathcal{M}_L\}$  of the BSs set  $\{1, \dots, M\}$ , and the corresponding partition  $\{\mathcal{K}_1, \dots, \mathcal{K}_L\}$  of the user groups set  $\{1, \dots, K\}$ , where  $\mathcal{M}_\ell$  and  $\mathcal{K}_\ell$  denote the set of BSs and user groups forming the  $\ell$ -th cooperation cluster. For example, consider the three-sectored hexagonal 7 cell network described in details in the numerical example of Section V. This network has one BS and 4 user groups in each of the 21 sectors. In the results of Section V we consider three possible cluster arrangements: (a) trivial clusters formed by a single sector; (b) clusters formed by the three sector within each cell; (c) a single cluster formed by all 21 sectors. The case of clusters containing a single BS (e.g., case (a) in the example) is referred to as “no cooperation”, since no joint processing across BSs is used. At the other extreme, the case of a single cluster containing all BSs (e.g., case (c) in the example) is referred to as “full cooperation”. The reader can refer to Fig. 5 in order to easily visualize the different cluster arrangements.

We assume that clusters are selfish and use all available transmit power, with no consideration for the ICI that they may cause to other clusters. Also, we assume that each cluster has *perfect channel state information* for all the users associated with the cluster, and has *statistical information* (i.e., known distributions but not the instantaneous values) relative to signals from other clusters. Within these channel state information assumptions, the ICI plus noise variance at any user terminal in group  $k \in \mathcal{K}_\ell$  is given by

$$\begin{aligned} \sigma_k^2 &= \mathbb{E} \left[ \frac{1}{N} \left\| \sum_{m \notin \mathcal{M}_\ell} \alpha_{m,k} \mathbf{H}_{m,k}^H \mathbf{x}_m + \mathbf{n}_k \right\|^2 \right] \\ &= 1 + \sum_{m \notin \mathcal{M}_\ell} \alpha_{m,k}^2 P_m. \end{aligned} \quad (2)$$

From well-known worst-case noise results [42], since users in each cluster have only statistical information about the ICI, treating the noise plus ICI as a white complex circularly symmetric Gaussian noise with variance  $\sigma_k^2$  yields an achievable inner bound to the achievable rate region of the clusters. Therefore, from the viewpoint of cluster  $\mathcal{M}_\ell$ , the system is equivalent to a single-cell MU-MIMO downlink with per-group-of-antennas power constraint, where each antenna group corresponds to each BS’s antennas, and with AWGN power at the user receivers given by (2). Hence, from now on,

<sup>2</sup>A more elaborate strategy considers multiple intertwined cluster patterns, operating over different frequency subbands. In this way, any given user group can be served by some cluster on one frequency subband and by some another cluster on another frequency band. By properly designing the cluster patterns, users located in a disadvantaged position (e.g., at the edge of a cluster) in a particular subband are simultaneously located in a good position (e.g., near the center of another cluster) on some other subband. This allows to mitigate “bottleneck” user groups and therefore to obtain larger system throughput subject to fairness, since user groups are *symmetrized* across the different frequency subbands. Intertwined clustering schemes have been proposed and investigated through Monte Carlo simulation in [41]. The analysis technique presented here can be extended to such schemes, although we don’t do this explicitly for the sake of brevity.

we shall focus on a reference cluster (say,  $\ell$ ) and simplify our notation. We let  $B = |\mathcal{M}_\ell|$  and  $A = |\mathcal{K}_\ell|$  denote the number of BS and user groups in the cluster, and enumerate the BS and the user groups forming the cluster as  $m = 1, \dots, B$  and  $k = 1, \dots, A$ , respectively. Also, we define the modified path coefficients  $\beta_{m,k} = \frac{\alpha_{m,k}}{\sigma_k}$  and the cluster channel matrix

$$\tilde{\mathbf{H}} = \begin{bmatrix} \beta_{1,1}\mathbf{H}_{1,1} & \cdots & \beta_{1,A}\mathbf{H}_{1,A} \\ \vdots & \ddots & \vdots \\ \beta_{B,1}\mathbf{H}_{B,1} & \cdots & \beta_{B,A}\mathbf{H}_{B,A} \end{bmatrix}. \quad (3)$$

Hence, one channel use of the reference cluster downlink is given by

$$\mathbf{y} = \tilde{\mathbf{H}}^H \mathbf{x} + \mathbf{v} \quad (4)$$

where  $\mathbf{y} = \mathbb{C}^{AN}$ ,  $\mathbf{x} = \mathbb{C}^{\gamma BN}$ , and  $\mathbf{v} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$  (we drop subscript  $\ell$  for notation simplicity).

It is well-known that the boundary of the capacity region of the MIMO BC (4) for fixed channel matrix  $\tilde{\mathbf{H}}$  and given per-group-of-antennas power constraints  $\{P_1, \dots, P_B\}$  can be characterized by the solution of a min-max weighted sum-rate problem [43]–[45]. For reasons that will be clear when discussing the scheduling policy in Section IV, we restrict ourselves to the case of identical weights for all statistically equivalent users, i.e., for the case that users in the same group have the same weight for their individual rates. We let  $W_k$  and  $R_k(\tilde{\mathbf{H}}) = \frac{1}{N} \sum_{i=1}^N R_{k,i}(\tilde{\mathbf{H}})$  denote the weight for user group  $k$  and the corresponding instantaneous per-user rate, respectively. In this paper, we refer to as “instantaneous” the quantities that depend on the realization of the channel matrix  $\tilde{\mathbf{H}}$ . Since this changes from slot to slot, instantaneous quantities also change accordingly. We let  $\pi$  denote the permutation that sorts the weights in increasing order  $W_{\pi_1} \leq \dots \leq W_{\pi_A}$  and use the subscript  $[k : A]$  to indicate quantities involving user groups from  $\pi_k$  to  $\pi_A$ .<sup>3</sup> Then, we let  $\tilde{\mathbf{H}}_k$  denote the  $k$ -th  $\gamma BN \times N$  slice of  $\tilde{\mathbf{H}}$  in (3), and  $\mathbf{Q}_k = \text{diag}(q_{k,1}, \dots, q_{k,N})$  denote an  $N \times N$  non-negative definite diagonal matrix of the dual uplink users’ transmit powers (see below). Furthermore, we define  $\mathbf{Q} = \text{diag}(\mathbf{Q}_1, \dots, \mathbf{Q}_A)$  and, for given permutation  $\pi$ , we let  $\tilde{\mathbf{H}}_{k:A} = [\tilde{\mathbf{H}}_{\pi_k} \dots \tilde{\mathbf{H}}_{\pi_A}]$  and  $\mathbf{Q}_{k:A} = \text{diag}(\mathbf{Q}_{\pi_k}, \dots, \mathbf{Q}_{\pi_A})$ . The rate point  $\{R_k(\tilde{\mathbf{H}}, W_1, \dots, W_A)\}$  on the boundary of the instantaneous capacity region corresponding to weights  $\{W_1, \dots, W_A\}$  is obtained as solution of the min-max problem [43]–[45]

$$\min_{\lambda \geq 0} \max_{\mathbf{Q} \geq 0} \sum_{k=1}^A W_{\pi_k} R_{\pi_k}(\tilde{\mathbf{H}}) \quad (5)$$

where the instantaneous per-user rate of each group takes on the expression

$$R_{\pi_k}(\tilde{\mathbf{H}}) = \frac{1}{N} \log \frac{|\Sigma(\lambda) + \tilde{\mathbf{H}}_{k:A} \mathbf{Q}_{k:A} \tilde{\mathbf{H}}_{k:A}^H|}{|\Sigma(\lambda) + \tilde{\mathbf{H}}_{k+1:A} \mathbf{Q}_{k+1:A} \tilde{\mathbf{H}}_{k+1:A}^H|} \quad (6)$$

<sup>3</sup>It is understood that matrices with subscripts  $k_1 : k_2$  with  $k_1 > k_2$  are empty (equal to all-zero matrices of appropriate dimension).

where  $\Sigma(\lambda)$  is a  $\gamma BN \times \gamma BN$  block-diagonal matrix with  $\gamma N \times \gamma N$  constant diagonal blocks  $\lambda_m \mathbf{I}_{\gamma N}$ , for  $m = 1, \dots, B$  and the maximization with respect to  $\mathbf{Q}$  is subject to the trace constraint

$$\text{tr}(\mathbf{Q}) \leq \sum_{m=1}^B \lambda_m P_m. \quad (7)$$

The variables  $\lambda = \{\lambda_m\}$  are the Lagrange multipliers corresponding to the per-group-of-antennas power constraints.

The rate  $R_{\pi_k}(\tilde{\mathbf{H}})$  in (6) can be interpreted as the instantaneous per-user rate of user group  $\pi_k$  in the dual vector Multiple Access Channel (MAC) with worst-case noise defined by

$$\mathbf{r} = \sum_{k=1}^A \tilde{\mathbf{H}}_k \mathbf{s}_k + \mathbf{z} \quad (8)$$

where  $\text{Cov}(\mathbf{s}_k) = \mathbf{Q}_k$  and  $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \Sigma(\lambda))$ . In this “dual uplink” interpretation, group sum-rate expression (6) corresponds to group-wise successive interference cancellation, where user groups are decoded successively in the order of  $\pi_1, \pi_2, \dots, \pi_A$ , and users in each group are jointly decoded. Also, notice that users in group  $\pi_k$  in general do not achieve individually the rate  $R_{\pi_k}(\tilde{\mathbf{H}})$  on every slot. Rather, this rate is the aggregate sum-rate of all users in group  $\pi_k$ , normalized by  $N$ , i.e., the mean user rate of group  $\pi_k$  for given  $\tilde{\mathbf{H}}$ . Efficient interior-point methods to solve (5) are given, for example, in [43]–[45]. Yet, the solution of this problem is numerically fairly involved, especially for large dimensions.

Consistently with the assumption of fixed coefficients  $\{\beta_{m,k}\}$  and ergodic block-fading for the small-scale fading coefficients  $\{\mathbf{H}_{m,k}\}$ , the *ergodic capacity region* of the MU-MIMO downlink channel (4) is given by the set of all achievable *average* rates, where averaging is with respect to the small-scale fading coefficients. In particular, an inner bound to the ergodic capacity region is given by

$$\begin{aligned} \underline{\mathcal{C}}(P_1, \dots, P_B) = \\ \text{coh} \bigcup_{W_1, \dots, W_A \geq 0} \{ \mathbf{R} : 0 \leq R_{k,i} \leq \mathbb{E}[R_k(\tilde{\mathbf{H}}, W_1, \dots, W_A)], \\ \forall k = 1, \dots, A, \forall i = 1, \dots, N \} \end{aligned} \quad (9)$$

where “coh” indicates the closure of the convex hull. The achievability of the above region is clear: all users  $i$  in group  $k$  are statistically equivalent and therefore they can achieve the same ergodic rate. Notice that  $\underline{\mathcal{C}}(P_1, \dots, P_B)$  is generally an inner bound because of the restriction of the weights in (5) to be identical for all users in the same group. We will see later that, for fairness scheduling in the limit of  $N \rightarrow \infty$ , this limitation becomes immaterial.

At this point we can formulate the fairness scheduling problem. Let  $g(\mathbf{R})$  denote a strictly increasing and concave network utility of the ergodic user rates. While the channel fading coefficients change from time slot to time slot according to some ergodic process, the optimal scheduling policy allocates dynamically the transmit powers and the DPC precoding

order in order to let the system operate at the ergodic rate point solution of

$$\begin{aligned} \max \quad & g(\mathbf{R}) \\ \text{s.t.} \quad & \mathbf{R} \in \underline{\mathcal{C}}(P_1, \dots, P_B). \end{aligned} \quad (10)$$

Different fairness criteria can be enforced by choosing appropriately the function  $g(\cdot)$  [20]. For example, proportional fairness [17], [46], [47] is obtained by letting  $g(\mathbf{R}) = \sum_{k,i} \log R_{k,i}$  and max-min fairness is obtained by letting  $g(\mathbf{R}) = \min_{k,i} R_{k,i}$ .

We notice here that an analytical computation of the ergodic rate point  $\mathbf{R}^*$  achieving the optimum in (10) is difficult in general, and has escaped clean closed form solutions even in much simpler cases. However, by applying the general stochastic optimization framework of [18] (see also [19], more specifically targeted to the MU-MIMO downlink), explicit scheduling policies can be designed such that the limit of the time-averaged user rates converges to  $\mathbf{R}^*$ . The rest of this paper is dedicated to finding an efficient method to directly compute  $\mathbf{R}^*$ , by exploiting large random matrix theory and convex optimization.

### III. WEIGHTED AVERAGE SUM RATE MAXIMIZATION

In this section, we consider the solution of the following preliminary problem. For fixed  $\{\lambda_m\}$ , we consider the maximization of the weighted average sum-rate maximization

$$\begin{aligned} \max \quad & \sum_{k=1}^A \frac{W_{\pi_k}}{N} \mathbb{E} \left[ \log \frac{|\boldsymbol{\Sigma}(\boldsymbol{\lambda}) + \tilde{\mathbf{H}}_{k:A} \mathbf{Q}_{k:A} \tilde{\mathbf{H}}_{k:A}^{\text{H}}|}{|\boldsymbol{\Sigma}(\boldsymbol{\lambda}) + \tilde{\mathbf{H}}_{k+1:A} \mathbf{Q}_{k+1:A} \tilde{\mathbf{H}}_{k+1:A}^{\text{H}}|} \right] \\ \text{s.t.} \quad & \text{tr}(\mathbf{Q}) \leq Q \end{aligned} \quad (11)$$

where we define  $Q \triangleq \sum_{m=1}^B \lambda_m P_m$ . The maximization in (11) is with respect to the diagonal elements  $\{q_{k,i}\}$  of  $\mathbf{Q}$ , which can be functions of the channel statistics, but not of their instantaneous values (that is,  $\mathbf{Q}$  does not depend on  $\tilde{\mathbf{H}}$ , but only on its statistics). Letting  $\Delta_k \triangleq W_{\pi_k} - W_{\pi_{k-1}}$  with  $W_{\pi_0} = 0$ , the objective function in (11) can be written as

$$\begin{aligned} F_{\mathbf{W},\boldsymbol{\lambda}}(\mathbf{Q}) = & \sum_{k=1}^A \Delta_k \frac{1}{N} \mathbb{E} \left[ \log \left| \boldsymbol{\Sigma}(\boldsymbol{\lambda}) + \tilde{\mathbf{H}}_{k:A} \mathbf{Q}_{k:A} \tilde{\mathbf{H}}_{k:A}^{\text{H}} \right| \right] \\ & - W_{\pi_A} \frac{1}{N} \log |\boldsymbol{\Sigma}(\boldsymbol{\lambda})|. \end{aligned} \quad (12)$$

In this form, problem (11) is clearly convex, since  $F_{\mathbf{W},\boldsymbol{\lambda}}(\mathbf{Q})$  in (12) is a concave function of  $\mathbf{Q}$ . In addition, we can prove the following symmetry result:

*Lemma 1:* The optimal  $\mathbf{Q}$  in (11) allocates equal power to the users in the same group. ■

*Proof:* See Appendix A. ■

Using Lemma 1, we restrict the optimization in (12) to block-diagonal matrices  $\mathbf{Q}$  with constant diagonal blocks  $\mathbf{Q}_k = \frac{Q_k}{N} \mathbf{I}$ . Furthermore, we have:

*Lemma 2:* The optimal  $\boldsymbol{\lambda}^*$  for the min-max problem (5) are strictly positive, i.e.,  $\boldsymbol{\lambda}^* > \mathbf{0}$ .

*Proof:* See Appendix A. ■

Using Lemma 2, we define  $\bar{\mathbf{H}}_k \triangleq \frac{1}{\sqrt{N}} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\lambda}) \tilde{\mathbf{H}}_k$  and rewrite the objective function with some abuse of notation as

$$F_{\mathbf{W},\boldsymbol{\lambda}}(Q_1, \dots, Q_A) = \sum_{k=1}^A \frac{\Delta_k}{N} \mathbb{E} \left[ \log \left| \mathbf{I} + \sum_{\ell=k}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^{\text{H}} Q_{\pi_\ell} \right| \right] \quad (13)$$

where the trace constraint (11) becomes  $\sum_{k=1}^A Q_k \leq Q$ .

#### A. Solution for Finite $N$

The Lagrangian function of problem (11) is given by

$$\mathcal{L}(Q_1, \dots, Q_A; \xi) = F_{\mathbf{W},\boldsymbol{\lambda}}(Q_1, \dots, Q_A) - \xi \left( \sum_{k=1}^A Q_k - Q \right). \quad (14)$$

Using the differentiation rule  $\partial \log \|\mathbf{X}\| = \text{tr}(\mathbf{X}^{-1} \partial \mathbf{X})$ , we write the KKT conditions as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial Q_{\pi_j}} = & \sum_{k=1}^j \frac{\Delta_k}{N} \mathbb{E} \left[ \text{tr} \left( \bar{\mathbf{H}}_{\pi_j}^{\text{H}} \left[ \mathbf{I} + \sum_{\ell=k}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^{\text{H}} Q_{\pi_\ell} \right]^{-1} \bar{\mathbf{H}}_{\pi_j} \right) \right] \leq \xi \end{aligned} \quad (15)$$

for  $j = 1, \dots, A$ , where equality must hold at the optimal point for all  $j$  such that  $Q_{\pi_j} > 0$ . After some algebra and the application of the Sherman-Morrison-Woodbury matrix inversion lemma [48], the trace in (15) can be rewritten in a more convenient form

$$\begin{aligned} & \frac{1}{N} \text{tr} \left( \bar{\mathbf{H}}_{\pi_j}^{\text{H}} \left[ \mathbf{I} + \sum_{\ell=k}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^{\text{H}} Q_{\pi_\ell} \right]^{-1} \bar{\mathbf{H}}_{\pi_j} \right) \\ & = \frac{1}{N} \text{tr} \left( \bar{\mathbf{H}}_{\pi_j}^{\text{H}} \boldsymbol{\Theta}_{k:A \setminus j}^{-1} \bar{\mathbf{H}}_{\pi_j} \left[ \mathbf{I} + Q_{\pi_j} \bar{\mathbf{H}}_{\pi_j} \boldsymbol{\Theta}_{k:A \setminus j}^{-1} \bar{\mathbf{H}}_{\pi_j} \right]^{-1} \right) \\ & = \frac{1 - \text{mmse}_{k:A}^{(j)}}{Q_{\pi_j}} \end{aligned} \quad (16)$$

where we let  $\boldsymbol{\Theta}_{k:A \setminus j} = \mathbf{I} + \sum_{\ell=k, \ell \neq j}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^{\text{H}} Q_{\pi_\ell}$  and where we define  $\text{mmse}_{k:A}^{(j)}$  as follows: consider the observation model

$$\mathbf{r}_{[k:A]} = \sum_{\ell=k}^A \sqrt{Q_{\pi_\ell}} \bar{\mathbf{H}}_{\pi_\ell} \mathbf{s}_\ell + \mathbf{z} \quad (17)$$

where  $\mathbf{s}_k, \mathbf{s}_{K+1}, \dots, \mathbf{s}_A$  and  $\mathbf{z}$  are Gaussian independent vectors with i.i.d. components  $\sim \mathcal{CN}(0, 1)$ . Then,  $\text{mmse}_{k:A}^{(j)}$  denotes the per-component MMSE for the estimation of  $\mathbf{s}_j$  from  $\mathbf{r}_{[k:A]}$ , for fixed (known) matrices  $\bar{\mathbf{H}}_{\pi_k}, \dots, \bar{\mathbf{H}}_{\pi_A}$ . Explicitly, we have

$$\begin{aligned} \text{mmse}_{k:A}^{(j)} & = \frac{1}{N} \text{tr} \left( \mathbf{I} - Q_{\pi_j} \bar{\mathbf{H}}_{\pi_j}^{\text{H}} \left[ \bar{\mathbf{H}}_{\pi_j} \bar{\mathbf{H}}_{\pi_j}^{\text{H}} Q_{\pi_j} + \boldsymbol{\Theta}_{k:A \setminus j} \right]^{-1} \bar{\mathbf{H}}_{\pi_j} \right) \\ & = \frac{1}{N} \text{tr} \left( \left[ \mathbf{I} + Q_{\pi_j} \bar{\mathbf{H}}_{\pi_j}^{\text{H}} \boldsymbol{\Theta}_{k:A \setminus j}^{-1} \bar{\mathbf{H}}_{\pi_j} \right]^{-1} \right). \end{aligned} \quad (18)$$

Using (16) in (15) and solving for the Lagrange multiplier, we find

$$\xi = \frac{1}{Q} \sum_{\ell=1}^A \sum_{k=1}^{\ell} \Delta_k \left( 1 - \mathbb{E} \left[ \text{mmse}_{k:A}^{(\ell)} \right] \right). \quad (19)$$

Finally, we arrive at the conditions

$$Q_{\pi_j} = Q \frac{\sum_{k=1}^j \Delta_k \left( 1 - \mathbb{E} \left[ \text{mmse}_{k:A}^{(j)} \right] \right)}{\sum_{\ell=1}^A \sum_{k=1}^{\ell} \Delta_k \left( 1 - \mathbb{E} \left[ \text{mmse}_{k:A}^{(\ell)} \right] \right)} \quad (20)$$

for all  $j$  such that  $Q_{\pi_j} > 0$ , and, using the KKT conditions and (19), we find that the inequality

$$\begin{aligned} & Q \sum_{k=1}^j \frac{\Delta_k}{N} \mathbb{E} \left[ \text{tr} \left( \mathbf{H}_{\pi_j}^H \mathbf{\Theta}_{k:A \setminus j}^{-1} \mathbf{H}_{\pi_j} \right) \right] \\ & \leq \sum_{\ell=1}^A \sum_{k=1}^{\ell} \Delta_k \left( 1 - \mathbb{E} \left[ \text{mmse}_{k:A}^{(\ell)} \right] \right) \end{aligned} \quad (21)$$

must hold for all  $j$  such that  $Q_{\pi_j} = 0$ . Eventually, we have proved the following result:

*Theorem 1:* The solution  $Q_1^*, \dots, Q_A^*$  of problem (11) is given as follows. For all  $j$  for which (21) is satisfied, then  $Q_{\pi_j}^* = 0$ . Otherwise, the positive  $Q_{\pi_j}^*$  satisfy (20). ■

In finite dimension, an iterative algorithm that provably converges to the solution can be obtained as a simple modification of [35, Algorithm 1]. The amount of calculation is tremendous because the average MMSE terms must be computed by Monte Carlo simulation. Since our emphasis is on the solution in the limit for  $N \rightarrow \infty$ , we omit these details and focus on the infinite-dimensional case in Section III-B. In addition, we have not yet addressed the outer minimization with respect to the Lagrange multipliers  $\{\lambda_m\}$ . We postpone this issue to Section III-D where we discuss system symmetry conditions for which the solution under the per-BS power constraint coincides with the laxer per-cluster sum power constraint. In this case, we can let  $\lambda_m = 1$  for all  $m$ , and no minimization with respect to  $\lambda$  is needed.

### B. Limit for $N \rightarrow \infty$

In this section, we consider problem (11) in the limit for  $N \rightarrow \infty$ , making use of the asymptotic random matrix results of [34]. In this regime, the instantaneous per-user rates in (5) converge to their expected values by well-known convergence results of the empirical distribution of the log-determinants in the rate expression (6) [32], [33]. Hence, in the large-system regime, the solution of (11) coincides with that of (5), for fixed channel pathloss coefficients  $\{\beta_{m,k}\}$ , transmit power constraints  $\{P_m\}$ , weights  $\{W_k\}$  and Lagrange multipliers  $\{\lambda_m\}$ . We will use this fact in Section IV, where we will examine a general dynamic fairness scheduling policy for the actual (finite dimensional) system, and study its performance in the large-system regime.

We introduce the normalized row and column indices  $r$  and  $t$ , taking values in  $[0, 1)$ , and the aspect ratio of the matrix  $\mathbf{H}$  given by the ratio of the number of columns over the number of rows and given by  $\nu = \frac{A}{\gamma B}$ . Then, we define the following piecewise constant functions:

- $Q(t)$ : (dual uplink) transmit power profile such that  $Q(t) = Q_{\pi_k}$  for  $\frac{k-1}{A} \leq t < \frac{k}{A}$ .
- $\mathcal{G}(r, t)$ : channel gain profile of the matrix  $\mathbf{H}$  such that  $\mathcal{G}(r, t) = \beta_{m,\pi_k}^2 / \lambda_m$  for  $\frac{m-1}{B} \leq r < \frac{m}{B}$  and  $\frac{k-1}{A} \leq t < \frac{k}{A}$ .
- $\Upsilon_{k:A}(t)$ : average per-component MMSE profile of the observation model (17), such that  $\Upsilon_{k:A}(t) = \text{mmse}_{k:A}^{(j)}$  for  $\frac{k-1}{A} \leq t < 1$ .
- $\Gamma_{k:A}(t)$ : signal-to-interference-plus-noise ratio (SINR) profile corresponding to  $\Upsilon_{k:A}(t)$  such that  $\Gamma_{k:A}(t) = 1 / \Upsilon_{k:A}(t) - 1$ .

In the limit of large  $N$ , these functions satisfy equations given by the following lemma:

*Lemma 3:* As  $N \rightarrow \infty$ , for each  $k = 1, \dots, A$ , the SINR functions  $\Gamma_{k:A}(t)$  satisfy the fixed-point equation

$$\Gamma_{k:A}(t) = \int_0^1 \frac{\gamma B \mathcal{G}(r, t) Q(t) dr}{1 + \nu \int_{(k-1)/A}^1 \frac{\gamma B \mathcal{G}(r, \tau) Q(\tau) d\tau}{1 + \Gamma_{k:A}(\tau)}}. \quad (22)$$

Also, the asymptotic  $\Upsilon_{k:A}(t)$  is given in terms of the asymptotic  $\Gamma_{k:A}(t)$  as  $\Upsilon_{k:A}(t) = 1 / (1 + \Gamma_{k:A}(t))$ .

*Proof:* See Appendix A. ■

Since all functions involved in Lemma 3 are piecewise constant (although the lemma applies in more generality), we can give a more explicit expression directly in terms of the discrete set of values of these functions. Replacing  $\Gamma_{k:A}(t) = \Gamma_{k:A}^{(j)}$  for all  $\frac{j-1}{A} \leq t < \frac{j}{A}$  with  $j \geq k$  in (22) and solving for the integrals of piecewise constant functions, we obtain

$$\begin{aligned} \Gamma_{k:A}^{(j)} &= \sum_{m=1}^B \int_{\frac{m-1}{B}}^{\frac{m}{B}} \frac{\gamma B \mathcal{G}(r, t) Q(t) dr}{1 + \nu \sum_{\ell=k}^A \int_{\frac{\ell-1}{A}}^{\frac{\ell}{A}} \frac{\gamma B \mathcal{G}(r, \tau) Q(\tau) d\tau}{1 + \Gamma_{k:A}^{(\ell)}}} \\ &= \gamma \sum_{m=1}^B \frac{\left( \beta_{m,\pi_j}^2 / \lambda_m \right) Q_{\pi_j}}{1 + \sum_{\ell=k}^A \frac{\left( \beta_{m,\pi_\ell}^2 / \lambda_m \right) Q_{\pi_\ell}}{1 + \Gamma_{k:A}^{(\ell)}}}. \end{aligned} \quad (23)$$

Combining (23) with the already mentioned modification of the iterative algorithm of [35, Algorithm 1], we obtain a procedure to compute the maximum weighted average sum rate of problem (11), for fixed weights  $\{W_k\}$  and Lagrange multipliers  $\{\lambda_m\}$ . This is summarized by Algorithm 1 below (for notation simplicity, the algorithm is written assuming  $\pi_k = k$  for all  $k = 1, \dots, A$ . We can always reduce to this case after a simple reordering of the weights).

---

#### Algorithm 1 Input power optimization for weighted average sum rate maximization

---

- 1) Initialize  $Q_k(0) = Q/A$  for all  $k = 1, \dots, A$ .
- 2) For  $i = 0, 1, 2, \dots$ , iterate until the following solution settles:

$$Q_j(i+1) = Q \frac{\sum_{k=1}^j \Delta_k \left( 1 - \Upsilon_{k:A}^{(j)}(i) \right)}{\sum_{\ell=1}^A \sum_{k=1}^{\ell} \Delta_k \left( 1 - \Upsilon_{k:A}^{(\ell)}(i) \right)} \quad (24)$$

for  $j = 1, \dots, A$ , where  $\Upsilon_{k:A}^{(j)}(i) = 1 / (1 + \Gamma_{k:A}^{(j)}(i))$ , and  $\Gamma_{k:A}^{(j)}(i)$  is obtained as the solution of the system of

fixed-point equations (23), also obtained by iteration, for powers  $Q_k = Q_k(i), \forall k$ .

- 3) Denote by  $\Gamma_{k:A}^{(j)}(\infty)$ ,  $\Upsilon_{k:A}^{(j)}(\infty)$  and by  $Q_j(\infty)$  the fixed-points reached by the iteration at step 2). If the condition

$$Q \sum_{k=1}^j \Delta_k \Gamma_{k:A}^{(j)}(\infty) \leq \sum_{\ell=1}^A \sum_{k=1}^{\ell} \Delta_k \left(1 - \Upsilon_{k:A}^{(\ell)}(\infty)\right) \quad (25)$$

is satisfied for all  $j$  such that  $Q_j(\infty) = 0$ , then stop. Otherwise, go back to the initialization step, set  $Q_j(0) = 0$  for  $j$  corresponding to the lowest value of  $\sum_{k=1}^j \Delta_k \Gamma_{k:A}^{(j)}(\infty)$ , and repeat steps 2) and 3) starting from the new initial condition.

### C. Computation of the Asymptotic Rates

After the powers  $Q_k^* = Q_k(\infty)$  have been obtained from Algorithm 1, it remains to compute the corresponding average per-user rates. The average rate of users in group  $k$  is given by

$$R_{\pi_k} = \frac{1}{N} \mathbb{E} \left[ \log \left| \mathbf{I} + \sum_{\ell=k}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^H Q_{\pi_\ell} \right| \right] - \frac{1}{N} \mathbb{E} \left[ \log \left| \mathbf{I} + \sum_{\ell=k+1}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^H Q_{\pi_\ell} \right| \right]. \quad (26)$$

In the limit for  $N \rightarrow \infty$ , we can use the asymptotic analytical expression for the mutual information given in [36]. Adapting [36, Result 1] to our case, we obtain

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{N} \mathbb{E} \left[ \log \left| \mathbf{I} + \sum_{\ell=k}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^H Q_{\pi_\ell} \right| \right] &= \\ \sum_{\ell=k}^A \log \left( 1 + \gamma Q_{\pi_\ell}^* \sum_{m=1}^B (\beta_{m,\pi_\ell}^2 / \lambda_m) u_m \right) &+ \\ + \gamma \sum_{m=1}^B \log \left( 1 + \sum_{\ell=k}^A (\beta_{m,\pi_\ell}^2 / \lambda_m) Q_{\pi_\ell}^* v_\ell \right) &- \\ - \gamma \sum_{\ell=k}^A \sum_{m=1}^B (\beta_{m,\pi_\ell}^2 / \lambda_m) Q_{\pi_\ell}^* u_m v_\ell & \quad (27) \end{aligned}$$

where for each  $k = 1, \dots, A$ ,  $\{u_m : m = 1, \dots, B\}$  and  $\{v_\ell : \ell = k, \dots, A\}$  are the unique solutions to the system of fixed-point equations

$$\begin{aligned} u_m &= \left( 1 + \sum_{\ell=k}^A Q_{\pi_\ell}^* (\beta_{m,\pi_\ell}^2 / \lambda_m) v_\ell \right)^{-1}, \quad m = 1, \dots, B, \\ v_\ell &= \left( 1 + \gamma \sum_{m=1}^B Q_{\pi_\ell}^* (\beta_{m,\pi_\ell}^2 / \lambda_m) u_m \right)^{-1}, \quad \ell = k, \dots, A. \end{aligned} \quad (28)$$

The proof follows from [36] based on the Girko's theorem [33] (see also [32]). Although (27) is not in a closed form,  $\{u_m\}$  and  $\{v_\ell\}$  in (28) can be solved by fixed-point iterations with  $A+B$  variables. These converge very quickly to the solution to any desired degree of numerical accuracy.

### D. System Symmetry

So far, we have considered the solution of the maximization in (11) for fixed  $\{\lambda_m\}$ . However, we are interested in the solution of (5) including the per-BS power constraint, that requires minimization with respect to  $\{\lambda_m\}$ . In finite dimension and for fixed channel matrix, the min-max problem can be solved by the subgradient-based iterative method of [44] or the infeasible-start Newton method of [45], [49]. A direct application of these algorithms to the large system limit requires asymptotic expressions for the subgradient with respect to  $\{\lambda_m\}$  or the *KKT matrix*, respectively. These quantities contain the second order derivatives of the Lagrangian function with respect to  $\{Q_k\}$  and  $\{\lambda_m\}$ , which do not appear to be amenable for easily computable asymptotic limits.

A general method for the minimization with respect to  $\{\lambda_m\}$  can be obtained as follows. Let  $G_{\mathbf{W}}(\boldsymbol{\lambda})$  denote the solution of (11). This is a convex function of  $\boldsymbol{\lambda}$  and the minimizing  $\boldsymbol{\lambda}^*$  must have strictly positive components by Lemma 2. Therefore, at the optimal point we have  $\frac{\partial G_{\mathbf{W}}}{\partial \lambda_m} \Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = 0$  for all  $m = 1, \dots, B$ . It follows that the solution can be approached by gradient descent iterations where the gradient can be estimated by numerical differentiation [50]. Let  $\boldsymbol{\epsilon}_m$  be a  $\gamma BN$ -length vector for which the elements  $(m-1)\gamma N + 1, \dots, m\gamma N$  are  $\epsilon$  for some  $\epsilon > 0$  and the other elements are zero. Then the approximation for the partial derivative of  $G_{\mathbf{W}}(\boldsymbol{\lambda})$  with respect to  $\lambda_m$  is given by  $\frac{G_{\mathbf{W}}(\boldsymbol{\lambda} + \boldsymbol{\epsilon}_m) - G_{\mathbf{W}}(\boldsymbol{\lambda} - \boldsymbol{\epsilon}_m)}{2\epsilon}$  with  $O(\epsilon^2)$  error term [50]. Both  $G_{\mathbf{W}}(\boldsymbol{\lambda} + \boldsymbol{\epsilon}_m)$  and  $G_{\mathbf{W}}(\boldsymbol{\lambda} - \boldsymbol{\epsilon}_m)$  are computed by Algorithm 1.

From the above argument it follows that the general case where minimization with respect to  $\{\lambda_m\}$  is required does not present any conceptual difficulty beyond the fact that it may be numerically cumbersome. Of course, a simple upper bound consists of relaxing the per-BS power constraint to a sum-power constraint in the reference cluster. Notice that the solution and the value of the objective function is invariant to a common scaling of the Lagrange multipliers. Therefore, we can assume  $\frac{1}{B} \sum_{m=1}^B \lambda_m = 1$  without loss of generality. Letting  $\lambda_m = 1$  for all  $m$  yields the laxer sum-power constraint  $\sum_k Q_k \leq \sum_m P_m \triangleq P_{\text{tot}}$ , where  $P_{\text{tot}}$  denotes the total transmit power of the cluster. This choice yields an upper-bound to the capacity region of the cluster (under the constraint of treating ICI as noise) and therefore also provides an upper-bound to the whole system achievable region under the assumption that all BSs transmit at their maximum power.

Next, we present a system symmetry condition under which the sum-power and the per-BS power solutions coincide. Assume the same BS power constraint  $P_m = P$  for all  $m = 1, \dots, B$ . Then, let  $A' = A/B$  assuming that  $B$  divides  $A$ . In particular, this is true when we have the same number of user groups in each cell of the cluster. Finally, assume that the  $B \times A$  matrix of the coefficients  $\boldsymbol{\beta} = \{\beta_{m,k}\}$  can be partitioned into  $A'$  submatrices of size  $B \times B$  such that each submatrix has the property that all rows are permutations of the first row, and all columns are permutations of the first column. Since this requirement reminds the condition for strongly symmetric discrete memoryless channels, we shall refer to these submatrices as "strongly symmetric blocks". To fix ideas, consider Fig. 1 showing a linear cellular layout with two cells and  $K$  user

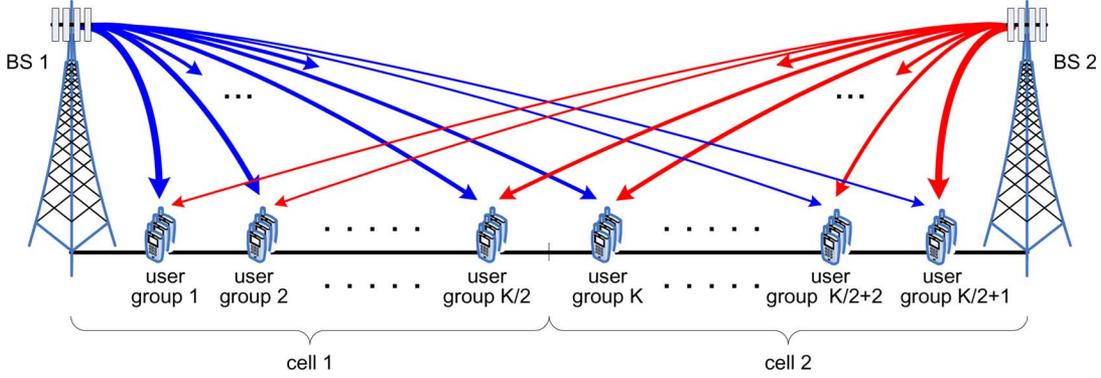


Fig. 1. Linear cellular layout with two cells and  $K$  symmetric user groups.

groups. Let  $K = 8$  and assume distance-dependent pathloss coefficients yielding the matrix

$$\boldsymbol{\beta} = \begin{bmatrix} a & b & b & a & f & e & d & c \\ f & e & d & c & a & b & b & a \end{bmatrix}$$

for some positive numbers  $a, b, c, d, e, f$ . We notice that this matrix can be decomposed into the  $A' = 4$  strongly symmetric blocks

$$\begin{bmatrix} a & f \\ f & a \end{bmatrix}, \begin{bmatrix} b & e \\ e & b \end{bmatrix}, \begin{bmatrix} b & d \\ d & b \end{bmatrix}, \begin{bmatrix} a & c \\ c & a \end{bmatrix}$$

satisfying the above assumption.

When these symmetry conditions hold, the user groups corresponding to the same strongly symmetric block (e.g., user groups pairs (1, 5), (2, 6), (3, 7) and (4, 8) in the example) are statistically equivalent, in the sense that they see exactly the same landscape of channel coefficients from all the BSs forming the cluster. In this case, as it will be clear in Section IV, we can restrict the weighted sum-rate maximization in (5), (11) to the case where the weights  $W_k$  are identical for all user groups in the same strongly symmetric block. Without loss of generality, let us enumerate the user groups such that the  $b$ -th symmetric block contains user groups with indices  $k = (b-1)B + 1, \dots, bB$ , with corresponding constant weights  $W_k = W'_b$ . Then, the objective function (13) takes on the form

$$F_{\mathbf{W}, \lambda}(Q_1, \dots, Q_A) = \sum_{b=1}^{A'} \frac{\Delta'_b}{N} \times \mathbb{E} \left[ \log \left| \mathbf{I} + \boldsymbol{\Sigma}^{-1}(\lambda) \tilde{\mathbf{H}}_{(b-1)B+1:A} \mathbf{Q}_{(b-1)B+1:A} \tilde{\mathbf{H}}_{(b-1)B+1:A}^H \right| \right] \quad (29)$$

where  $\Delta'_b = W'_{b+1} - W'_b$  for  $b = 1, \dots, A'$ , with  $W'_0 = 0$ , and where

$$\mathbf{Q} = \frac{1}{N} \text{diag}(\underbrace{Q_1, \dots, Q_1}_N, \underbrace{Q_2, \dots, Q_2}_N, \dots, \underbrace{Q_A, \dots, Q_A}_N)$$

with trace constraint  $\sum_{k=1}^A Q_k \leq BP = P_{\text{tot}}$ . We have the following result:

**Theorem 2:** Under the above system symmetry conditions, the minimization in the min-max problem (5) in the limit of  $N \rightarrow \infty$  is achieved for  $\lambda_m = 1$  for all  $m = 1, \dots, B$ .

*Proof:* See Appendix A. ■

As a corollary, we have that under the symmetry conditions of Theorem 2 the per-BS power constraint and the total power constraint solutions coincide.

#### IV. FAIRNESS SCHEDULING

Downlink opportunistic scheduling is currently used by “high data rate” third-generation cellular systems such as EV-DO [46] and HSDPA [47]. It is expected that in the next generation of systems based on MIMO-OFDM, such as IEEE 802.16m [1] and LTE-Advanced [2], such strategies will be integrated with the MU-MIMO physical layer. In such systems, each cooperation cluster runs a downlink scheduler that computes a set of rate weight coefficients and, at each scheduling time slot  $t$ , solves the maximization of the instantaneous weighted rate-sum subject to the per-BS power constraint, as in (5). The result of this maximization provides the power and rate allocation and the corresponding downlink precoder parameters (i.e., the beamforming vectors and the DPC encoding order) to be used in the current slot. The scheduler weights are recursively computed such that the time-averaged user rates converge to the desired ergodic rate point  $\mathbf{R}^*$ , the solution of (10).

The scheduling policy can be systematically designed by using the stochastic optimization approach of [18], [19], based on the idea of “virtual queues”. Notice that we do not consider *exogenous* arrivals: consistently with the classical information theoretic setting, we assume that an arbitrarily large number of information bits are to be transmitted to the users in each cluster (infinitely backlogged system). The virtual queues defined here are only a tool to recursively compute the weights of the scheduling policy.

For the sake of clarity, we provide here a short review of the scheduling algorithm for the general finite-dimensional system. We denote *instantaneous* quantities as dependent on the slot index  $t$ . The scheduling policy ensures that the virtual queue of each user  $(k, i)$  (i.e., user  $i$  in group  $k$ ) is *strongly stable* (see [18, Definition 3.1]). This implies that the arrival rate  $\Lambda_{k,i}$  is strictly less than the average service rate  $R_{k,i} = \mathbb{E}[R_{k,i}(\mathbf{H}(t))]$ . Then, the desired ergodic rate point  $\mathbf{R}^*$  can be approached if the virtual queues are fed by virtual arrival processes  $A_{k,i}(t)$  with rates  $\Lambda_{k,i} = \mathbb{E}[A_{k,i}(t)]$  sufficiently close to the desired values  $R_{k,i}^*$ . The interesting feature of this approach is that it is possible to generate such virtual arrival processes greedily and adaptively, even if the values  $R_{k,i}^*$  are unknown a priori, and may be very difficult to be calculated directly.

Let  $U_{k,i}(t)$  denote the virtual queue backlog for user  $i$  in group  $k$  at time slot  $t$ , evolving according to the stochastic difference equation

$$U_{k,i}(t+1) = [U_{k,i}(t) - R_{k,i}(\tilde{\mathbf{H}}(t))]_{+} + A_{k,i}(t) \quad (30)$$

where  $[\cdot]_{+} = \max(\cdot, 0)$ . We consider the scheduling policy given as follows:

- At each time slot  $t$ , solve the weighted sum-rate maximization problem

$$\begin{aligned} \max \quad & \sum_{k=1}^A \sum_{i=1}^N U_{k,i}(t) R_{k,i}(\tilde{\mathbf{H}}(t)) \\ \text{s.t.} \quad & \text{Cov}(\mathbf{x}_m) \leq P_m. \end{aligned} \quad (31)$$

- The virtual queues are updated according to (30), where the arrival processes are given by  $A_{k,i}(t) = a_{k,i}^*$ , where the vector  $\mathbf{a}^*$  is the solution of the maximization problem

$$\begin{aligned} \max \quad & Vg(\mathbf{a}) - \sum_{k=1}^A \sum_{i=1}^N a_{k,i} U_{k,i}(t) \\ \text{s.t.} \quad & 0 \leq a_{k,i} \leq A_{\max} \end{aligned} \quad (32)$$

for suitable  $V > 0$  and  $A_{\max} > 0$ .

The parameters  $V$  and  $A_{\max}$  determine the accuracy and the convergence speed of the time-average rates to their expected values. It can be shown [18], [19] that, for fixed sufficiently large parameters  $A_{\max}$ , the gap between the long-time average rates  $\lim_{t \rightarrow \infty} \sum_{\tau=0}^{t-1} \frac{1}{t} R_{k,i}(\tilde{\mathbf{H}}(\tau))$  and the optimal ergodic rates  $R_{k,i}^*$  decrease as  $O(1/V)$  while the expected backlog of the virtual queues increases as  $O(V)$ .

After reviewing the above background on scheduling and stochastic optimization, we are ready to make some observations that are instrumental for the performance computation in the large-system limit. Due to the statistical equivalence of users in the same group, the ergodic rate points with  $R_{k,i} = R_k$  (independent of  $i$ ) are achievable. In particular, the boundary of the system ergodic capacity region and of the region  $\mathcal{C}(P_1, \dots, P_B)$  in (9) coincide for all rate points meeting this condition. It is meaningful to assume that the network utility function  $g(\mathbf{R})$  is invariant with respect to permutations of the rates of statistically equivalent users. In fact, all statistically equivalent users should be treated equally in the long-term average sense.<sup>4</sup> For example, the  $\alpha$ -fairness utility function proposed in [20] satisfies this condition. In this case, it is immediate to show that the function  $g(\mathbf{R})$  is maximized by some rate point with equal rates over each user group or, if the symmetry conditions of Theorem 2 hold, over all groups in the same strongly symmetric block. Hence, in large-system limit the point  $\mathbf{R}^* = \{R_{k,i}^*\}$  solution of (10) must satisfy, for all  $i$ ,

$$R_{\pi_k, i}^* = \lim_{N \rightarrow \infty} \frac{1}{N} \mathbb{E} \left[ \log \frac{|\mathbf{I} + \sum_{\ell=k}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^H Q_{\pi_\ell}|}{|\mathbf{I} + \sum_{\ell=k+1}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^H Q_{\pi_\ell}|} \right]$$

<sup>4</sup>Here, we assume that all users have equal priority. For example, they are all delay-tolerant data users with no particular individual priority: users differ only by their location in the cluster, which determines their channel coefficients  $\{\beta_{k,m}\}$ .

where the term on the right-hand side is the average per-user rate given by the solution of (11) for some choice of the weights  $\{W_k\}$  and Lagrange multipliers  $\{\lambda_m\}$ . It is well-known that, for a deterministic network, the dynamic scheduling policy described before coincides with the Lagrangian dual optimization with outer subgradient iteration, where the Lagrangian dual variables play the role of the virtual queues backlogs in the dynamic setting. In the large-system limit, the channel uncertainty disappears and the MU-MIMO system ‘‘freezes’’ to a deterministic limit. Using the large-system limit solution of (11) presented in Section III, the solution of the fairness scheduling problem (10) can be addressed directly, using Lagrangian duality.

### A. Lagrangian Optimization

We rewrite (10) using the auxiliary variables  $\mathbf{r} = [r_1, \dots, r_A]$  and using the definition of the ergodic rate region (9) as

$$\begin{aligned} \min_{\lambda} \max_{\mathbf{r}, \mathbf{Q}, \pi} \quad & g(\mathbf{r}) \\ \text{s.t.} \quad & r_{\pi_k} \leq \frac{1}{N} \mathbb{E} \left[ \log \frac{|\mathbf{I} + \sum_{\ell=k}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^H Q_{\pi_\ell}|}{|\mathbf{I} + \sum_{\ell=k+1}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^H Q_{\pi_\ell}|} \right], \\ & \text{tr}(\mathbf{Q}) \leq Q, \lambda \geq 0. \end{aligned} \quad (33)$$

The Lagrange function for (33) is given by

$$\begin{aligned} \mathcal{L}(\lambda, \mathbf{r}, \mathbf{Q}, \pi, \mathbf{W}) = & g(\mathbf{r}) - \underbrace{\sum_{k=1}^A W_k r_k}_{f_{\mathbf{W}}(\mathbf{r})} \\ & + \underbrace{\sum_{k=1}^A W_{\pi_k} \frac{1}{N} \mathbb{E} \left[ \log \frac{|\mathbf{I} + \sum_{\ell=k}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^H Q_{\pi_\ell}|}{|\mathbf{I} + \sum_{\ell=k+1}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^H Q_{\pi_\ell}|} \right]}_{h_{\mathbf{W}}(\lambda, \mathbf{Q}, \pi)} \end{aligned} \quad (34)$$

where  $\mathbf{W}$  is the vector of dual variables corresponding to the auxiliary variable constraints (rate constraints). The Lagrange function can be decomposed into a sum of functions of  $\mathbf{r}$  only, denoted by  $f_{\mathbf{W}}(\mathbf{r})$ , and a function of  $\lambda, \mathbf{Q}$  and  $\pi$  only, denoted by  $h_{\mathbf{W}}(\lambda, \mathbf{Q}, \pi)$ . The Lagrange dual function for the problem (34) is given by

$$\begin{aligned} \mathcal{G}(\mathbf{W}) = & \min_{\lambda} \max_{\mathbf{r}, \mathbf{Q}, \pi} \mathcal{L}(\lambda, \mathbf{r}, \mathbf{Q}, \pi, \mathbf{W}) \\ = & \underbrace{\max_{\mathbf{r}} f_{\mathbf{W}}(\mathbf{r})}_{(a)} + \underbrace{\min_{\lambda} \max_{\mathbf{Q}, \pi} h_{\mathbf{W}}(\lambda, \mathbf{Q}, \pi)}_{(b)} \end{aligned} \quad (35)$$

and it is obtained by the decoupled maximization in (a) (with respect to  $\mathbf{r}$ ) and the min-max in (b) (with respect to  $\lambda, \mathbf{Q}, \pi$ ). Notice that problems (a) and (b) correspond to the static forms of (32) and (31), respectively, after identifying  $\mathbf{r}$  with the virtual arrival rates  $\mathbf{A}(t)$  and  $\mathbf{W}$  with the virtual queue backlogs  $\mathbf{U}(t)$ . Finally, we can solve the dual problem defined as

$$\min_{\mathbf{W} \geq 0} \mathcal{G}(\mathbf{W}). \quad (36)$$

Eventually, the solution of (36) can be found via inner-outer iterations as follows:

**Inner Problem:** For given  $\mathbf{W}$ , we solve (35) with respect to  $\lambda$ ,  $\mathbf{r}$ ,  $\mathbf{Q}$  and  $\pi$ . This can be further decomposed into:

- Subproblem (a): Since  $f_{\mathbf{W}}(\mathbf{r})$  is concave in  $\mathbf{r} \geq 0$ , the optimal  $\mathbf{r}^*$  readily obtained by imposing the KKT conditions.
- Subproblem (b): Taking the limit of  $N \rightarrow \infty$ , this problem is solved by Algorithm 1 for fixed  $\lambda > 0$ . If the system satisfies the symmetry conditions of Theorem 2 hold, then we let  $\lambda_m = 1$  and no minimization with respect to  $\lambda_m$  is needed. If these conditions do not hold, the outer minimization can be solved by the gradient descent method with the approximated gradient. Otherwise, letting  $\lambda_m = 1$  yields an upper bound on the achievable network utility function, corresponding to the relaxation of the per-BS power constraints to the sum-power constraint.

**Outer Problem:** the minimization of  $\mathcal{G}(\mathbf{W})$  with respect to  $\mathbf{W} \geq \mathbf{0}$  can be obtained by subgradient adaptation. Let  $\lambda^*$ ,  $\pi^*$ ,  $\mathbf{Q}^*$  and  $\mathbf{r}^*(\mathbf{W})$ <sup>5</sup> denote the solution of the inner problem for fixed  $\mathbf{W}$ . For any  $\mathbf{W}'$ , we have

$$\begin{aligned} \mathcal{G}(\mathbf{W}') &= \max_{\mathbf{r}} f_{\mathbf{W}'}(\mathbf{r}) + \max_{\mathbf{Q}} h_{\mathbf{W}'}(\lambda^*, \mathbf{Q}, \pi^*) \\ &\geq f_{\mathbf{W}'}(\mathbf{r}^*(\mathbf{W})) + h_{\mathbf{W}'}(\lambda^*, \mathbf{Q}^*, \pi^*) \\ &= \mathcal{G}(\mathbf{W}) + \sum_{k=1}^A (W'_k - W_k)(R_k^*(\mathbf{W}) - r_k^*(\mathbf{W})) \end{aligned} \quad (37)$$

where  $R_k^*(\mathbf{W})$  denotes the  $k$ -th group rate resulting from the solution of the inner problem with weights  $\mathbf{W}$ , which is efficiently calculated by Algorithm 1 in the large-system regime. A subgradient for  $\mathcal{G}(\mathbf{W})$  is given by the vector with components  $R_k^*(\mathbf{W}) - r_k^*(\mathbf{W})$ . Eventually, the dual variables  $\mathbf{W}$  can be updated at the  $n$ -th outer iteration according to

$$W_k(n+1) = W_k(n) - \mu(n)(R_k^*(\mathbf{W}(n)) - r_k^*(\mathbf{W}(n))) \quad (38)$$

for all  $k$ , and for some step size  $\mu(n) > 0$  which can be determined by an efficient algorithm such as the back-tracking line search method [51] or Ellipsoid method [52]. In the numerical example of Section V, we use the back-tracking line search method. It should be noticed that by setting  $\mu(n) = 1$  this subgradient update plays the role of the virtual queue update in the dynamic scheduling policy of (30). But in this optimization, the objective function converges to a single optimal point by the iterations and, by adjusting the step size  $\mu(n)$  with the above methods, the convergence can be attained very fast.

As an application example of the above general optimization, we focus on the two special cases of *proportional fairness scheduling* (PFS) and *hard-fairness scheduling* (HFS), also known as max-min fairness scheduling.

### B. Proportional Fairness Scheduling

The network utility function for PFS is given as

$$g(\mathbf{r}) = \sum_{k=1}^A \log(r_k). \quad (39)$$

<sup>5</sup>It is useful to explicitly point out the dependence of  $\mathbf{W}$  only for  $\mathbf{r}^*(\mathbf{W})$ , since this appears in the subgradient expression, although it is clear that  $\lambda^*$ ,  $\pi^*$  and  $\mathbf{Q}^*$  also in general depend on  $\mathbf{W}$ .

In this case, the KKT conditions for the inner subproblem (a) yield the solution

$$r_k^*(\mathbf{W}) = 1/W_k, \quad \forall k \quad (40)$$

(notice that  $r_k$  must be positive for all  $k$  otherwise the objective function is  $-\infty$ ). As mentioned before, the dual variables play the role of the virtual queue backlogs in the dynamic scheduling policy, while the auxiliary variables  $\mathbf{r}$  correspond to the virtual arrival rates. From (40), we see that at the  $n$ -th outer iteration these variables are related by  $W_k(n) = \frac{1}{r_k^*(\mathbf{W}(n))}$ . As observed at the beginning of Section IV, the virtual arrival rates of the dynamic scheduling policy are designed in order to be close to the ergodic rates  $\mathbf{R}^*$  at the optimal fairness point. It follows that the usual intuition of PFS, according to which the scheduler weights are inversely proportional to the long-term average user rates, is recovered.

### C. Hard Fairness Scheduling

In case of HFS, the scheduler maximizes the minimum ergodic user rate. The network utility function is given by

$$g(\mathbf{r}) = \min_{k=1, \dots, A} r_k. \quad (41)$$

This objective function is not strictly concave and differentiable everywhere. Therefore, it is convenient to rewrite subproblem (a) by introducing an auxiliary variable  $\gamma$ , as follows:

$$\begin{aligned} \max_{\gamma, \mathbf{r} \geq 0} \quad & \gamma - \sum_{k=1}^A W_k r_k \\ \text{subject to} \quad & r_k \geq \gamma, \forall k. \end{aligned} \quad (42)$$

The solution must satisfy  $r_k = \gamma$  for all  $k$ , leading to

$$\max_{\gamma > 0} \left( 1 - \sum_{k=1}^A W_k \right) \gamma. \quad (43)$$

Since the original maximization in (33) is bounded while (43) may be unbounded, we must have that  $\sum_{k=1}^A W_k = 1$  and  $\gamma$  must take on some appropriate value that enforces this condition. The subgradient iteration for the weights  $\mathbf{W}$ , using  $r_k^*(\mathbf{W}(n)) = \gamma^*(\mathbf{W}(n))$ , becomes

$$W_k(n+1) = W_k(n) - \mu(R_k^*(\mathbf{W}(n)) - \gamma^*(\mathbf{W}(n))), \quad \forall k. \quad (44)$$

Summing up the update equations over  $k = 1, \dots, A$  and using the conditions that  $\sum_{k=1}^A W_k(n) = 1$  for all  $n$ , we obtain

$$r_k^*(\mathbf{W}(n)) = \gamma^*(\mathbf{W}(n)) = \frac{1}{A} \sum_{j=1}^A R_j^*(\mathbf{W}(n)), \quad \forall k. \quad (45)$$

## V. NUMERICAL RESULTS

In this section, we present some examples of the multicell model considered in this paper and (when possible) compare the numerical results using the proposed large-system analysis with the results of Monte Carlo simulation applied to an actual finite-

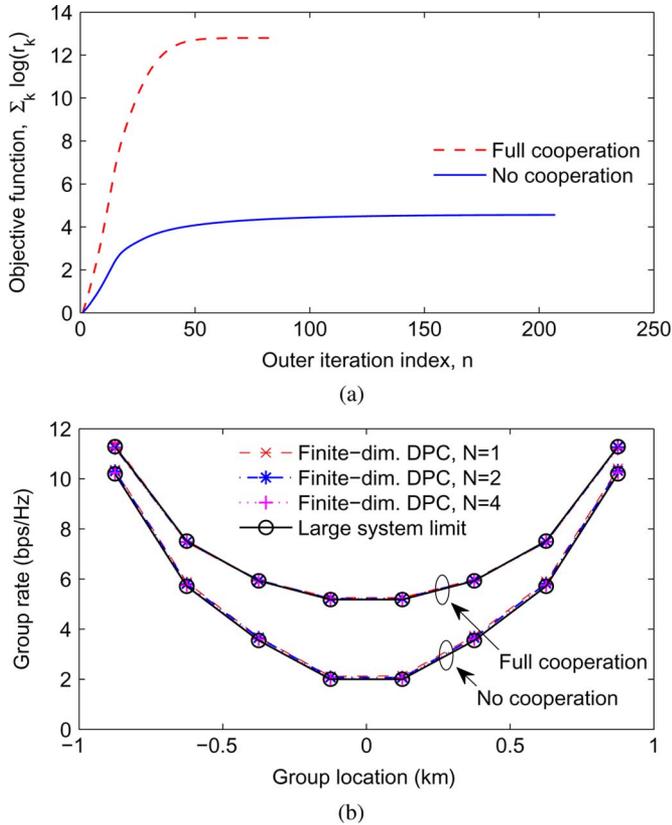


Fig. 2. Proportional fairness scheduling with  $\gamma = 4$  and  $K = 8$  in the 2-cell model. (a) Convergence of the utility function. (b) Individual group rates.

dimensional system subject to the dynamic fairness scheduling policy outlined at the beginning of Section IV.

The examples involve a 1-D 2-cell model ( $M = 2$ ) and a 2-D three-sectored 7-cell model ( $M = 21$ ). In both cases, the system parameters and pathloss model are based on the mobile WiMAX system evaluation specification [16] with cell radius 1.0 km and no shadowing assumption. The 2-cell model, shown in Fig. 1, considers two one-sided BSs with  $\gamma = 4$ , located at  $\pm 1$  km, and  $K = 8$  user groups equally spaced between the BSs. We consider the case of full BS cooperation (i.e., one cluster comprising the two BSs) and no cooperation (i.e., two clusters of one BS each) with a symmetric partition of users, yielding  $L = 2$  clusters with  $\mathcal{K}_1 = \{1, 2, 3, 4\}$  and  $\mathcal{K}_2 = \{5, 6, 7, 8\}$ .

Fig. 2 illustrates the convergence of the utility function and individual group rates under PFS. In Fig. 2(a), the PFS objective functions in the no cooperation and full cooperation cases are shown to converge to the respective optimal PFS points. Not surprisingly, the full cooperative system achieves significantly higher value of the objective function (sum of the log-rates). In Fig. 2(b), we compare the asymptotic rates in the large-system limit with the achievable rates obtained by using Monte Carlo simulation in finite dimension. In finite dimension we considered  $N = 1, 2$ , or 4 and the same parameters of the infinite-dimensional case. The channel vectors are randomly generated and change at every  $t$  in an i.i.d. fashion, and the instantaneous rates are allocated by using the DPC with the water-filling algorithm [43] combined with the dynamic scheduling policy [19]

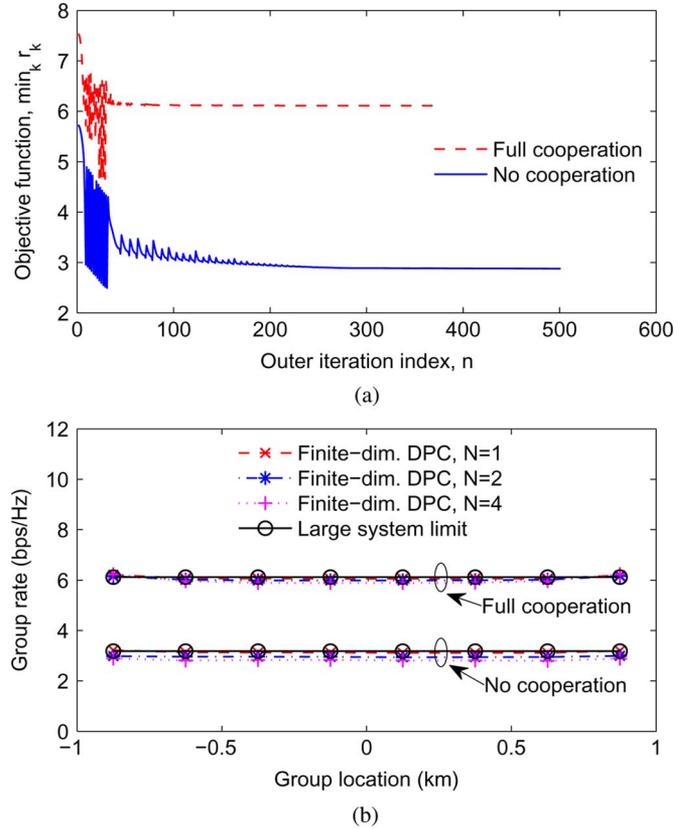


Fig. 3. Hard fairness scheduling with  $\gamma = 4$  and  $K = 8$  in the 2-cell model. (a) Convergence of the utility function. (b) Individual group rates.

outlined in Section IV. Remarkably, the finite-dimensional simulation produced nearly the same rates for the considered values of  $N$  and these rates also almost overlap with the large-system asymptotic results even for very small  $N$ . Notice that the dynamic scheduling policy should provide multiuser diversity gain and in general should achieve higher rates than the large-system limit, which is not able to exploit the dynamic fluctuations of the small-scale fading due to “channel hardening”. However, it appears that in the regime where the pathloss is dominant over the randomness of the multiantenna channels and the number of users is not much larger than the number of BS antennas, the multiuser diversity gain is negligible and the asymptotic analysis generates results very close to the simulations with dynamic scheduling and DPC.

Fig. 3 shows the convergence behavior of the utility function and individual group under HFS. In the HFS case, all the users achieve the same individual rate which is slightly higher than the smallest rate of the PFS case. Also, the agreement with of the individual user rates with the finite-dimensional simulation is noticeable.

Using the proposed asymptotic analysis, validated in the simple 2-cell model, we can obtain ergodic rate distributions for much larger systems, for which a full-scale simulation would be very demanding. We considered a 2-D cell layout where seven hexagonal cells form a network and each cell consists of three 120-degree sectors. As depicted in Fig. 4(a), three BSs are co-located at the center of each cell such that each BS handles one sector in no cooperation case. Each sector is split into the 4 diamond-shaped equal-area grids and one user group is placed

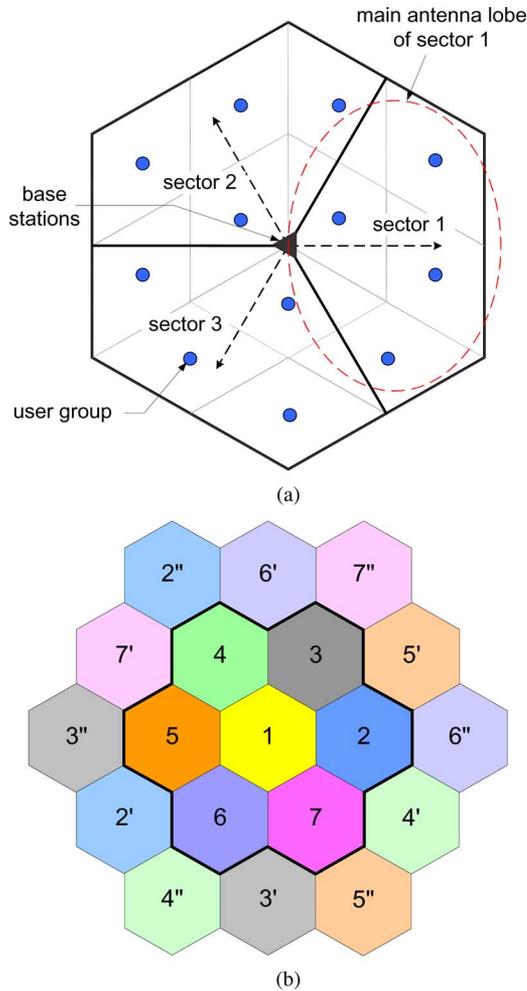


Fig. 4. Two-dimensional three-sectored 7-cell model. (a) 3-sectored cell configuration. (b) Wrap-around torus topology.

at the center of each grid. Therefore there are total  $M = 21$  BSs and  $K = 84$  user groups in the network. In addition, we assume a wrap-around torus topology as shown in Fig. 4(b), such that each cell is virtually surrounded by the other six cells and all the cells have the symmetric ICI distribution. The antenna orientation and pattern follows [53] and the non-ideal spatial antenna gain pattern (overlapping between sectors in the same cell) generates interference even between sectors of the same cell. Fig. 6 shows the user rate distribution with PFS under three levels of cooperation (a) no cooperation; (b) three-sector cooperation within each cell; and (c) full cooperation over 7-cell network, as illustrated in Fig. 5. From the asymptotic rate results, it is shown that in case (b), the cooperation gain over the case (a) is primarily obtained for the users around cell centers, while the cooperation gain is attained over the whole cellular coverage area in case (c).

## VI. CONCLUSION

We considered the downlink of a multicell MU-MIMO cellular system where the pathloss and inter-cell interference make the users' channel statistics unequal. In this case, it is important to evaluate the system performance subject to some form of fairness, formulated as the maximization of a concave network utility function over the achievable ergodic rate region

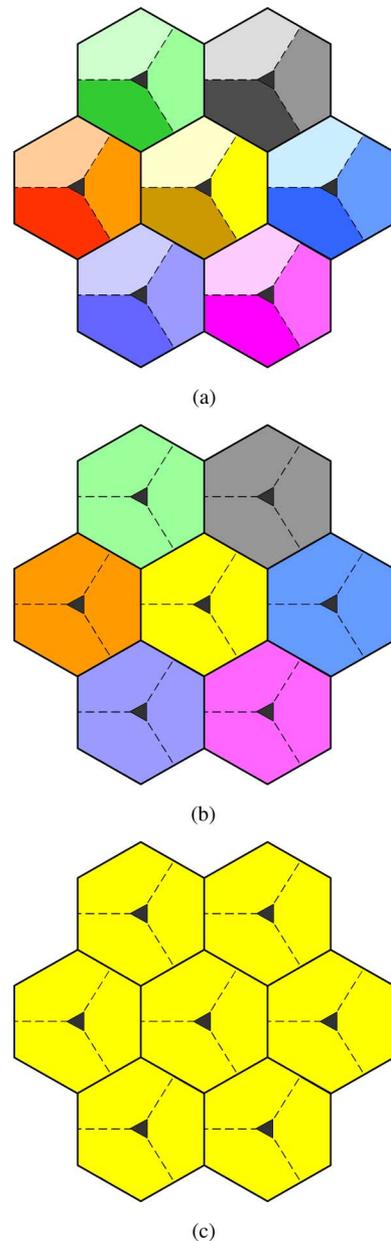


Fig. 5. Cluster forms for three level of cooperation with different colors denoting different clusters. (a) No cooperation. (b) Three-sector cooperation within each cell. (c) Full cooperation over seven cells.

of the system. Downlink scheduling and resource allocation in order to maximize a desired network utility function is a widely studied issue in the literature, that has found several important practical applications [17], [46], [47]. Although dynamic scheduling policies are well-known form in the theory of stochastic network optimization, the calculation of the resulting ergodic throughput for a multiuser multicell MIMO downlink of practically relevant size, including tens of cells, hundreds of users per cell, and clusters of jointly processed cooperating base stations has been evaluated so far through very demanding system Monte Carlo simulation.

In this work, we used a large-system limit approach, where the number of base station antennas and the number of users at each location go to infinity with a fixed ratio. In this regime, we presented a semi-analytic method for the computation of

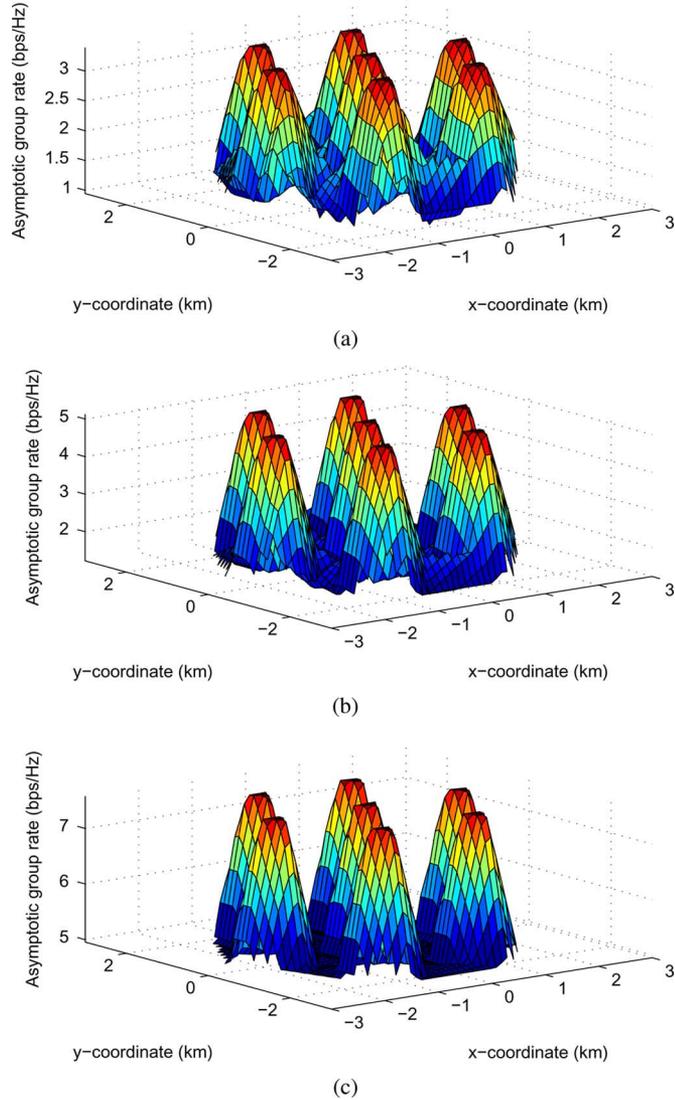
APPENDIX A  
 PROOFS


Fig. 6. Ergodic user rate distribution with PFS in the 7-cell model. (a) No cooperation. (b) Three-sector cooperation within each cell. (c) Full cooperation over seven cells.

the optimal fairness rate point, based on a combination of large random matrix results and Lagrangian optimization. We proved that the system optimization subject to per-BS power constraints coincides with the generally laxer sum-power constraint if the system satisfies certain symmetry conditions. In any case, our analysis yields an efficient computation method of the system throughput subject to general fairness criteria, through the iterative solution of a system of fixed-point equations. Numerical results show that the rates predicted by the large-system analysis are indeed remarkably close to the rates of the corresponding finite-dimensional systems obtained by Monte Carlo simulation. Overall, the results of this paper are useful in evaluating and optimizing multicell MU-MIMO systems, especially when the system dimension and network size are large. For example, using our method we can quickly quantify and compare the performance of different cooperative clustering arrangements, subject to different fairness criteria.

*Proof of Lemma 1:* Denote the utility function in (12) as a function of diagonal entries of  $\mathbf{Q}$  as  $f(\{q_{k,i} : 1 \leq k \leq A, 1 \leq i \leq N, \})$ . Since users in the same group are statistically equivalent and the function  $f(\cdot)$  is defined through an expectation with respect to the channel fading coefficients, it follows that  $f(\cdot)$  must be invariant with respect to permutations of the arguments  $q_{k,1}, \dots, q_{k,N}$ . That is, for any  $k = 1, \dots, A$ , and  $1 \leq i < j \leq N$ , the value of the function is invariant if the arguments  $q_{k,i}$  and  $q_{k,j}$  are exchanged. Suppose that  $\{q_{k,i}^*\}$  is the optimal input power allocation, solution of (11). Then, we have

$$\begin{aligned} f(\dots, q_{k,i}^*, \dots, q_{k,j}^*, \dots) &= f(\dots, q_{k,j}^*, \dots, q_{k,i}^*, \dots) \\ &\leq f\left(\dots, \frac{q_{k,i}^* + q_{k,j}^*}{2}, \dots, \frac{q_{k,i}^* + q_{k,j}^*}{2}, \dots\right) \end{aligned}$$

where the inequality follows from the concavity of  $f(\cdot)$  and Jensen's inequality. Under the optimality assumption, equality must hold and this implies that  $\frac{q_{k,i}^* + q_{k,j}^*}{2}$  is the optimal input power for both users  $i$  and  $j$  in group  $k$ . Extending this argument by induction, it follows that the optimal input power must be in the form  $q_{k,i}^* = Q_k/N$  for all  $i = 1, \dots, N$ , for some values  $Q_1, \dots, Q_A$ .

*Proof of Lemma 2:* The dual variable  $\lambda_m$  plays the role of the noise power at the antennas of the  $m$ -th BS in the dual MAC. Let  $G_{\mathbf{W}}(\boldsymbol{\lambda}) = \max_{\mathbf{Q}: \text{tr}(\mathbf{Q}) \leq Q} F_{\mathbf{W}, \boldsymbol{\lambda}}(\mathbf{Q})$  and suppose that  $\lambda_m^* = 0$  for some  $m$ . Then,  $|\Sigma(\boldsymbol{\lambda}^*)| = 0$  in (12) and  $G_{\mathbf{W}}(\boldsymbol{\lambda}^*)$  goes to positive infinity, which clearly cannot be the solution of the minimization with respect to  $\boldsymbol{\lambda}$  in (5). Therefore, the optimal  $\lambda_m^*$  is strictly positive for all  $m = 1, \dots, B$ .

*Proof of Lemma 3:* We apply [34, Lemma 1] to the matrix  $\mathbf{I} + \sum_{\ell=k}^A \bar{\mathbf{H}}_{\pi_\ell} \bar{\mathbf{H}}_{\pi_\ell}^H Q_{\pi_\ell}$  where  $\bar{\mathbf{H}}_{k:A} = [\bar{\mathbf{H}}_{\pi_k}, \dots, \bar{\mathbf{H}}_{\pi_A}]$  has independent non-identically distributed components. The variance profile in [34, Lemma 1] is defined as the limit of the variance of the elements of the matrix  $\bar{\mathbf{H}}_{k:A}$ , multiplied by the number of rows,  $\gamma BN$ . With our normalization, the elements of each  $(m, \ell)$ -th block of  $\bar{\mathbf{H}}_{k:A}$  of size  $\gamma N \times N$  have variance  $\frac{\beta_{m, \pi_\ell}^2 / \lambda_m}{N}$ . Therefore, the variance profile for the application of [34, Lemma 1] is given by  $\gamma B \mathcal{G}(r, t)$ , where  $\mathcal{G}(r, t)$  is the piecewise constant function defined above. Eventually, we arrive at (22).

*Proof of Theorem 2:* Let  $\mathbf{P}_{i,j} = \mathbb{E} \left[ \left[ \tilde{\mathbf{H}}_{i,j} \right]^2 \right]$  denote the variance of the  $(i, j)$ -th element of the channel matrix. For any  $b = 1, \dots, A'$ , the matrix  $\tilde{\mathbf{H}}_{(b-1)B+1:A}$  has the property that the empirical distribution of the element variances for all rows, i.e., the cumulative distribution functions

$$\mathcal{F}_{i,b}^{(N)}(z) \triangleq \frac{1}{(A - (b-1)B)N} \sum_{j=(b-1)B+1}^{AN} 1\{\mathbf{P}_{i,j} \leq z\}$$

are the same, for all row index  $i = 1, \dots, \gamma BN$ . This means that the matrix of the element variances  $\{\mathbf{P}_{i,j}\}$  corresponding to  $\tilde{\mathbf{H}}_{(b-1)B+1:A}$  is *row-regular* (see definition in [34, Definition 5]). Under the row-regularity condition, it follows that

$$\begin{aligned}
F_{\mathbf{W},\lambda}(Q_1, \dots, Q_A) &\stackrel{(a)}{=} \sum_{b=1}^{A'} \frac{\Delta'_b}{N} \mathbb{E} \left[ \frac{1}{B!} \sum_{\Pi} \log \left| \mathbf{I} + \Sigma^{-1}(\lambda) \Pi \check{\mathbf{H}}_{(b-1)B+1:A} \mathbf{Q}_{(b-1)B+1:A} \check{\mathbf{H}}_{(b-1)B+1:A}^{\mathbf{H}} \Pi^{\mathbf{T}} \right| \right] \\
&= \sum_{b=1}^{A'} \frac{\Delta'_b}{N} \mathbb{E} \left[ \frac{1}{B!} \sum_{\Pi} \log \left| \mathbf{I} + \Pi^{\mathbf{T}} \Sigma^{-1}(\lambda) \Pi \check{\mathbf{H}}_{(b-1)B+1:A} \mathbf{Q}_{(b-1)B+1:A} \check{\mathbf{H}}_{(b-1)B+1:A}^{\mathbf{H}} \right| \right] \\
&= \sum_{b=1}^{A'} \frac{\Delta'_b}{N} \mathbb{E} \left[ \frac{1}{B!} \sum_{\Pi} \log \left| \mathbf{I} + (\Pi^{\mathbf{T}} \Sigma(\lambda) \Pi)^{-1} \check{\mathbf{H}}_{(b-1)B+1:A} \mathbf{Q}_{(b-1)B+1:A} \check{\mathbf{H}}_{(b-1)B+1:A}^{\mathbf{H}} \right| \right] \\
&\stackrel{(b)}{\geq} \sum_{b=1}^{A'} \frac{\Delta'_b}{N} \mathbb{E} \left[ \log \left| \mathbf{I} + \left( \frac{1}{B!} \sum_{\Pi} \Pi^{\mathbf{T}} \Sigma(\lambda) \Pi \right)^{-1} \check{\mathbf{H}}_{(b-1)B+1:A} \mathbf{Q}_{(b-1)B+1:A} \check{\mathbf{H}}_{(b-1)B+1:A}^{\mathbf{H}} \right| \right] \\
&\stackrel{(c)}{=} \sum_{b=1}^{A'} \frac{\Delta'_b}{N} \mathbb{E} \left[ \log \left| \mathbf{I} + \check{\mathbf{H}}_{(b-1)B+1:A} \mathbf{Q}_{(b-1)B+1:A} \check{\mathbf{H}}_{(b-1)B+1:A}^{\mathbf{H}} \right| \right] \tag{46}
\end{aligned}$$

$\check{\mathbf{H}}_{(b-1)B+1:A}$  in the limit of  $N \rightarrow \infty$  is statistically equivalent to a matrix  $\check{\mathbf{H}}_{(b-1)B+1:A} = \mathbf{G}_{(b-1)B+1:A} \mathbf{T}_{(b-1)B+1:A}^{1/2}$ , where  $\mathbf{G}_{(b-1)B+1:A}$  is an i.i.d. matrix with zero-mean, unit-variance elements, and  $\mathbf{T}_{(b-1)B+1:A}$  is a non-negative diagonal matrix with asymptotic empirical spectral distribution given by  $\lim_{N \rightarrow \infty} \mathcal{F}_{i,b}^{(N)}(z)$ . In particular, the distribution of  $\check{\mathbf{H}}_{(b-1)B+1:A}$  is asymptotically unitary left-invariant, that is, for any unitary matrix  $\mathbf{U}$  independent of  $\check{\mathbf{H}}_{(b-1)B+1:A}$ , the matrices  $\mathbf{U} \check{\mathbf{H}}_{(b-1)B+1:A}$  and  $\check{\mathbf{H}}_{(b-1)B+1:A}$  are asymptotically identically distributed.

Let  $\Pi$  denote a  $\gamma BN \times \gamma BN$  block-permutation matrix, that permutes the  $B$  blocks of consecutive positions of length  $\gamma N$  in the index vector  $\{1, \dots, \gamma BN\}$ . Using the above asymptotic statistical equivalence, in the limit of large  $N$  we obtain the sequence of steps in (46), as shown at the top of the page, for any  $\{\lambda_m\}$  and  $\{Q_k\}$ , where (a) follows from the left-unitary invariance, (b) follows from Jensen's inequality, and (c) from the fact that, without loss of generality, we let  $\frac{1}{B} \sum_{m=1}^B \lambda_m = 1$ . This shows that, for asymptotically large  $N$  and under the given symmetry conditions of the channel coefficients and rate weights, the worst-case Lagrange multipliers for the weighted maximization of the average rates in (11) is  $\lambda_m = 1$ . Since for  $N \rightarrow \infty$  the instantaneous rates in (5) converge to the average rates in (11), the theorem is proved.

## REFERENCES

- [1] IEEE 802.16 broadband wireless access working group, IEEE 802.16m System Requirements IEEE 802.16m-07/002 Tech. Rep., Jan. 2010.
- [2] 3GPP technical specification group radio access network, Further Advancements for E-UTRA: LTE-Advanced Feasibility Studies in RAN WG4 3GPP TR 36.815, Tech. Rep., Mar. 2010.
- [3] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.
- [4] P. Viswanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.
- [5] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.
- [6] N. Jindal, S. Vishwanath, and A. J. Goldsmith, "On the duality of gaussian multiple-access and broadcast Channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 768–783, May 2004.
- [7] W. Yu and J. Cioffi, "Sum capacity of Gaussian vector broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1875–1892, Sep. 2004.
- [8] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.
- [9] A. D. Wyner, "Shannon-theoretic approach to a Gaussian cellular multiple access channel," *IEEE Trans. Inf. Theory*, vol. 40, no. 11, pp. 1713–1727, Nov. 1994.
- [10] O. Somekh, B. M. Zaidel, and S. Shamai (Shitz), "Sum rate characterization of joint multiple cell-site processing," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4473–4497, Dec. 2007.
- [11] S. Shamai (Shitz) and A. D. Wyner, "Information-theoretic considerations for symmetric, cellular, multiple-access fading channels—Part I & II," *IEEE Trans. Inf. Theory*, vol. 43, no. 11, pp. 1877–1894, Nov. 1997.
- [12] O. Somekh and S. Shamai (Shitz), "Shannon-theoretic approach to a Gaussian cellular multiple-access channel with fading," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 1401–1425, Jul. 2000.
- [13] S. Shamai (Shitz), O. Simeone, O. Somekh, A. Sanderovich, B. M. Zaidel, and H. V. Poor, "Information-theoretic implications of constrained cooperation in simple cellular models," in *Proc. IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Cannes, France, Sep. 2008.
- [14] A. Sanderovich, O. Somekh, H. Poor, and S. Shamai (Shitz), "Uplink macro diversity of limited backhaul cellular network," *IEEE Trans. on Inf. Theory*, vol. 55, no. 8, pp. 3457–3478, Aug. 2009.
- [15] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.
- [16] WiMAX Forum, Mobile WiMAX—Part I: A Technical Overview and Performance Evaluation Tech. Rep., Aug. 2006.
- [17] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.
- [18] L. Georgiadis, M. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Found. and Trends in Netw.*, vol. 1, no. 1, 2006.
- [19] H. Shirani-Mehr, G. Caire, and M. J. Neely, "MIMO downlink scheduling with non-perfect channel state knowledge," *IEEE Trans. Commun.*, vol. 58, no. 7, pp. 2055–2066, Jul. 2010.

- [20] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 10, pp. 556–567, Oct. 2000.
- [21] H. Huang and R. Valenzuela, "Fundamental simulated performance of downlink fixed wireless cellular networks with multiple antennas," in *Proc. IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Berlin, Germany, Sep. 2005.
- [22] F. Boccardi and H. Huang, "Limited downlink network coordination in cellular networks," in *Proc. IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Athens, Greece, Sep. 2007.
- [23] H. Zhang, N. Mehta, A. Molisch, J. Zhang, and H. Dai, "Asynchronous interference mitigation in cooperative base station systems," *IEEE Trans. Wireless Commun.*, vol. 7, no. 1, pp. 155–165, Jan. 2008.
- [24] G. Caire, S. Ramprasad, H. Papadopoulos, C. Pepin, and C.-E. Sundberg, "Multiuser MIMO downlink with limited inter-cell cooperation: Approximate interference alignment in time, frequency and space," in *Proc. Allerton Conf. on Commun., Control, and Computing*, Urbana-Champaign, IL, Sep. 2008.
- [25] P. Marsch and G. Fettweis, "On base station cooperation schemes for downlink network MIMO under a constrained backhaul," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, New Orleans, LA, Nov. 2008.
- [26] J. Zhang, R. Chen, J. Andrews, A. Ghosh, and R. Heath, "Networked MIMO with clustered linear precoding," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 1910–1921, Apr. 2009.
- [27] H. Huang, M. Trivellato, A. Hottinen, M. Shafi, P. Smith, and R. Valenzuela, "Increasing downlink cellular throughput with limited network MIMO coordination," *IEEE Trans. on Wireless Commun.*, vol. 8, pp. 2983–2989, Jun. 2009.
- [28] S. A. Ramprasad and G. Caire, "Cellular vs. network MIMO: A comparison including the channel state information overhead," in *Proc. IEEE Int. Symp. on Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Tokyo, Japan, Sep. 2009.
- [29] S. Parkvall, E. Dahlman, A. Furuskar, Y. Jading, M. Olsson, S. Wanstedt, and K. Zangi, "LTE-advanced—Evolving LTE towards IMT-advanced," in *Proc. IEEE Vehic. Tech. Conf. (VTC)*, Calgary, Alberta, Sep. 2008.
- [30] J.-B. Landre, A. Saadani, and F. Ortolan, "Realistic performance of HSDPA MIMO in macro-cell environment," in *Proc. IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Tokyo, Japan, Sep. 2009.
- [31] A. Farajidana, W. Chen, A. Damnjanovic, T. Yoo, D. Malladi, and C. Lott, "3GPP LTE downlink system performance," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Honolulu, HI, Nov. 2009.
- [32] A. M. Tulino and S. Verdu, "Random matrix theory and wireless communications," *Found. Trends in Commun. and Inf. Theory*, vol. 1, no. 1, 2004.
- [33] V. L. Girko, *Theory of Random Determinants*. Dordrecht/Boston: Kluwer Publishers, 1990.
- [34] A. M. Tulino, A. Lozano, and S. Verdu, "Impact of antenna correlation on the capacity of multiantenna channels," *IEEE Trans. Inf. Theory*, vol. 7, no. 7, pp. 2491–2509, Jul. 2005.
- [35] A. M. Tulino, A. Lozano, and S. Verdu, "Capacity-achieving input covariance for single-user multiantenna channels," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 662–671, Mar. 2006.
- [36] D. Aktas, M. N. Bacha, J. S. Evans, and S. V. Hanly, "Scaling results on the sum capacity of cellular networks with MIMO links," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 3264–3274, Jul. 2006.
- [37] S.-H. Park and I. Lee, "Degrees of freedom of multiple broadcast channels in the presence of inter-cell interference," *IEEE Trans. Commun.*, vol. 59, no. 5, pp. 1481–1487, May 2011.
- [38] C. Suh, M. Ho, and D. Tse, "Downlink interference alignment," *IEEE Trans. Commun.*, vol. 59, no. 1, pp. 1–11, Jan. 2011.
- [39] A. Moustakas, S. Simon, and A. Sengupta, "MIMO capacity through correlated channels in the presence of correlated interferers and noise: A (not so) large N analysis," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2545–2561, Oct. 2003.
- [40] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge University Press, 2005.
- [41] G. Caire, S. A. Ramprasad, and H. C. Papadopoulos, "Rethinking network MIMO: Cost of CSIT, performance analysis, and architecture comparisons," in *Proc. Inform. Theory and Appl. Workshop (ITA)*, San Diego, CA, Feb. 2010.
- [42] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?," *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.
- [43] W. Yu, "Sum-capacity computation for the Gaussian vector broadcast channel via dual decomposition," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 754–759, Feb. 2006.
- [44] L. Zhang, R. Zhang, Y.-C. Liang, Y. Xin, and H. V. Poor, "On gaussian MIMO BC-MAC duality with multiple transmit covariance constraints," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Seoul, Korea, Jun. 2009.
- [45] H. Huh, H. C. Papadopoulos, and G. Caire, "Multiuser MISO transmitter optimization for intercell interference mitigation," *IEEE Trans. Signal Proc.*, vol. 58, no. 8, pp. 4272–4285, Aug. 2010.
- [46] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, vol. 38, no. 7, pp. 70–77, Jul. 2000.
- [47] S. Parkvall, E. Englund, M. Lundevall, and J. Torsner, "Evolving 3G mobile systems: Broadband and broadcast services in WCDMA," *IEEE Commun. Mag.*, vol. 44, no. 2, pp. 30–36, Feb. 2006.
- [48] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge University Press, 1985.
- [49] W. Yu and T. Lan, "Transmitter optimization for the multi-antenna downlink with per-antenna power constraints," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2646–2660, Jun. 2007.
- [50] W. Cheney and D. Kincaid, *Numerical Mathematics and Computing*. Florence, KY: Thomson Brooks/Cole, 2004.
- [51] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.
- [52] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. Nashua, NH: Athena Scientific, 1997.
- [53] IEEE 802.16 broadband wireless access working group, IEEE 802.16m Evaluation Methodology Document (EMD) IEEE 802.16m-08/004, Tech. Rep., Jan. 2009.

**Hoon Huh** (S'09–M'11) received the B.S. and M.S. degrees from Seoul National University, Seoul, Korea, and the Ph.D. degree from the University of Southern California (USC), Los Angeles, CA, all in electrical engineering, in 1997, 1999, and 2011, respectively.

Since 1999, he has been with Samsung Electronics, Suwon, Korea, working on the physical-layer standards and the MODEM and scheduler algorithms for the 3GPP LTE, Mobile WiMAX, and CDMA2000 1xEV-DO systems. His research interests are multiple antenna techniques, resource allocation, and cooperative communications in the areas of information theory and wireless communications.

**Sung-Hyun Moon** (S'05) received the B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, Korea, in 2005 and 2007, respectively, where he is currently working toward the Ph.D. degree in the School of Electrical Engineering. During spring 2009, he visited the University of Southern California (USC), Los Angeles, CA, as a visiting student. His research interests include multiuser information theory and signal processing techniques for MIMO wireless cellular networks.

**Young-Tae Kim** (S'07) received the B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, Korea in 2006 and 2008, respectively, where he is currently working toward the Ph.D. degree in the school of Electrical Engineering. During spring 2009, he visited the University of Southern California (USC), Los Angeles, CA, as a visiting student. His research includes signal processing techniques for wireless communication systems. Mr. Kim received the silver paper award at the IEEE Seoul Section Paper Contest award in 2007.

**Inkyu Lee** (S'92–M'95–SM'01) received the B.S. degree (Hon.) in control and instrumentation engineering from Seoul National University, Seoul, Korea, in 1990, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1992 and 1995, respectively. From 1995 to 2001, he was a Member of Technical Staff at Bell Laboratories, Lucent Technologies, where he conducted research on high-speed wireless system designs. He later worked for Agere Systems (formerly Microelectronics Group of Lucent Technologies), Murray Hill, NJ, as a Distinguished Member of Technical Staff from 2001 to 2002. In September 2002, he joined the faculty of Korea University, Seoul, Korea, where he is currently a Professor in the School of Electrical Engineering. During 2009, he visited University of Southern California, LA, USA, as a visiting Professor. He has published over 70 journal papers in IEEE, and has 30 U.S. patents granted or pending. His research interests include digital communications, signal processing, and coding techniques applied for next generation wireless systems. Dr. Lee currently serves as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. Also, he has been a Chief Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (Special Issue on 4G Wireless Systems). He received the IT Young Engineer Award as the IEEE/IEEK joint award in 2006. Also he received the Best Paper Award at APCC in 2006 and IEEE VTC in 2009.

**Giuseppe Caire** (S'92–M'94–SM'03–F'05) was born in Torino, Italy, in 1965. He received the B.Sc. in electrical engineering from Politecnico di Torino (Italy), in 1990, the M.Sc. in Electrical Engineering from Princeton University in 1992 and the Ph.D. from Politecnico di Torino in 1994. He was a recipient of the AEI G. Someda Scholarship in 1991, has been with the European Space Agency (ESTEC, Noordwijk, The Netherlands) from May 1994 to February 1995, was a recipient of the COTRAO Scholarship in 1996 and of a CNR Scholarship in 1997. He has been visiting Princeton University in Summer 1997 and Sydney University in Summer 2000. He has been Assistant Professor in Telecommunications at the Politecnico di Torino, Associate Professor at the University of Parma, Italy, Professor with the Department of Mobile Communications at the Eurecom Institute, Sophia-Antipolis, France, and he is currently a professor of Electrical Engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, CA.

He served as Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS in 1998–2001 and as Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY in 2001–2003. He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, and the IEEE Communications Society & Information Theory Society Joint Paper Award in 2004 and in 2011. Giuseppe Caire is Fellow of IEEE since 2005. He has served in the Board of Governors of the IEEE Information Theory Society from 2004 to 2007, and as President of the IEEE Information Theory Society in 2011. His main research interests are in the field of communications theory, information theory, channel and source coding with particular focus on wireless communications.