Multi-Cell MIMO Downlink with Cell Cooperation and Fair Scheduling: a Large-System Limit Analysis

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

Abstract

We consider the downlink of a cellular network with multiple cells and multi-antenna base stations, including a realistic distance-dependent pathloss model, 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 so far. We propose an analytic solution based on the combination of large random matrix results and convex optimization. The proposed method is computationally much more efficient than Monte Carlo simulation and provides surprisingly accurate approximations for the actual finite-dimensional systems, even for a small number of users and base station antennas. Numerical examples include 2-cell linear and three-sectored 7-cell planar layouts, with no inter-cell cooperation, sector cooperation, or full inter-cell cooperation.

Index Terms

Asymptotic analysis, fairness scheduling, inter-cell cooperation, large-system limit, multi-cell MIMO downlink, weighted sum rate maximization.

H. Huh and G. Caire are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA. (e-mail: hhuh, 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, reftm, inkyu@korea.ac.kr)

The material in this paper was presented in part at the 2010 IEEE International Communications Conference (ICC), Cape Town, South Africa, May 2010 and will be presented in part at the 2010 International Symposium on Information Theory (ISIT), Austin, Texas, June 2010.

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]–[7] 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 [7] where the optimality of Gaussian Dirty-Paper Coding (DPC) is shown, subject to a general convex input covariance constraint.

In a multi-cell 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 the multi-cell system was introduced by Wyner [8]. In both one-dimensional linear cellular array and two-dimensional hexagonal cellular pattern, only interference from adjacent cells is considered with a single scaling factor, and the uplink capacity is obtained in a closed form in the case of full joint processing of all cells and no fading. Wyner's setting was extended in several works [9]–[13]. Single cell processing and joint two-cell processing was investigated by treating the inter-cell interference as Gaussian noise in [9] and the flat-fading channel case with full joint cell processing was treated in [10]. 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 [11]–[13] and references therein).

Although the Wyner model captures some fundamental aspects of the multi-cell problem, its rather unrealistic assumption for the pathloss makes the system essentially symmetric with respect to any user. More realistically, 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 [14], and therefore they are in fundamentally asymmetric conditions. It follows that characterizing the sum-capacity (or achievable sum-throughput, under some suboptimal scheme) is rather meaningless from a system performance viewpoint, unless some appropriate notion of *fairness* is taken into account. In fact, if the sum-throughput is the only objective, the resulting rate and power optimization under distance-dependent pathloss would lead to the solution of serving only the users close to their base station (BS), while leaving the users at the cell edge to starve.

As a matter of fact, "fairness" is a fundamental aspect in cellular networks. The problem of *downlink scheduling* subject to some fairness criterion has been widely studied (see for example [15]–[17] 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 increasing *network utility function*

is maximized [18]. By choosing the shape of the network utility function, a desired fairness criterion can be enforced. The framework of *stochastic network optimization* [16] 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 [19]–[29].

In this paper, we propose an alternative approach based on the "large-system limit." We leverage results on large random matrices [30]–[34], in order to characterize the system achievable rate region in the limit where both the number of antennas per BS 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 BSs in the same cluster operate as a distributed antenna array (full cooperation) and inter-cluster interference is treated as Gaussian noise (no inter-cluster cooperation). As special cases, we recover conventional cellular systems (no inter-cell cooperation) and the case of full cooperation. 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 solve this problem in almost closed form (up to the numerical solution of a fixed-point equation). The solution is particularly simple when each cooperation cluster satisfies certain symmetry conditions that will be discussed later on. The proposed method is much more efficient than Monte Carlo simulation and, somehow surprisingly, it provides results that match very closely the performance of finite-dimensional systems, even for very small dimension.

The remainder of this paper is organized as follows. In Section II, we present the MU-MIMO downlink system model with cell cluster 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 M BSs with γN antennas each, and KN single-antenna user terminals, distributed in the cellular coverage region. Users are divided into K co-located "user groups" of equal size N. Users in the same group are statistically equivalent: they experience the same pathloss from all BSs and their small-scale fading channel coefficients are 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, but the wavelength is sufficiently small so that they all have essentially the same distance-dependent pathloss. Users in different groups observe generally different pathlosses, depending on their relative positions with respect to the BSs.

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 [35]. 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 multi-cell MU-MIMO downlink is described by

$$\mathbf{y}_k = \sum_{m=1}^{M} \alpha_{m,k} \mathbf{H}_{m,k}^{\mathsf{H}} \mathbf{x}_m + \mathbf{n}_k \tag{1}$$

where $\mathbf{y}_k = [y_{k,1} \dots y_{k,N}]^{\mathsf{T}} \in \mathbb{C}^N$ denotes the received signal vector for the k-th user group, $\alpha_{m,k}$ and $\mathbf{H}_{m,k}$ denote the 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, $\mathbf{x}_m = [x_{m,1} \dots x_{m,\gamma N}]^{\mathsf{T}} \in \mathbb{C}^{\gamma N}$ is the signal vector transmitted by the m-th BS, and $\mathbf{n}_k = [n_{k,1} \dots n_{k,N}]^{\mathsf{T}} \in \mathbb{C}^N$ 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 $\mathrm{tr}\left(\mathrm{Cov}(\mathbf{x}_m)\right) \leq P_m$, where $P_m > 0$ denotes the total transmit power of the m-th BS.

We assume that the BSs are grouped into cooperation clusters. Each cluster acts effectively as a distributed MU-MIMO system, with a distributed transmit antenna array formed by all antennas of all BS in the cluster. 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, we consider ideal joint

processing of all BSs in the same cluster. Inter-Cluster Interference (ICI) is treated as additional Gaussian noise. Let L denote the number of cooperation clusters. We define the BS partition $\{\mathcal{M}_1,\ldots,\mathcal{M}_L\}$ of the set $\{1,\ldots,M\}$ and the corresponding user group partition $\{\mathcal{K}_1,\ldots,\mathcal{K}_L\}$ of the set $\{1,\ldots,K\}$, where $\mathcal{M}_\ell$ and $\mathcal{K}_\ell$ denote the set of BSs and user groups forming the ℓ-th cooperation cluster. 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. Hence, the ICI plus noise variance at any user terminal in group $k \in \mathcal{K}_\ell$ is given by

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

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). Therefore, from now on, we shall focus on a reference cluster (say, ℓ) 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 $m=1,\ldots,B$ and $k=1,\ldots,A$, respectively. Also, we define the modified path coefficients $\beta_{m,k}=\frac{\alpha_{m,k}}{\sigma_k}$ and the cluster channel matrix

$$\widetilde{\mathbf{H}}=\begin{bmatrix}\beta_{1,1}\mathbf{H}_{1,1} & \cdots & \beta_{1,A}\mathbf{H}_{1,A}\\ \vdots & \ddots & \vdots\\ \beta_{B,1}\mathbf{H}_{B,1} & \cdots & \beta_{B,A}\mathbf{H}_{B,A}\end{bmatrix}. \tag{3}$$

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

$$\mathbf{y}=\widetilde{\mathbf{H}}^{\mathsf{H}}\mathbf{x}+\mathbf{v} \tag{4}$$

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 ℓ for notation simplicity).

It is well-known that the boundary of the capacity region of the MIMO BC (4) for fixed channel matrix $\widetilde{\mathbf{H}}$ and given per-group-of-antennas power constraints $\{P_1,\ldots,P_B\}$ can be characterized by the solution of a min-max weighted sum-rate problem [36]–[38]. 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 $R_k(\widetilde{\mathbf{H}})=\frac{1}{N}\sum_{i=1}^{N}R_{k,i}(\widetilde{\mathbf{H}})$ 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 $\widetilde{\mathbf{H}}$. Since this changes from slot to slot, instantaneous quantities also change accordingly. We let π denote the permutation that sorts the weights in increasing order $W_{\pi_1} \leq \ldots \leq W_{\pi_A}$ and use the subscript $[k:A]$ to indicate quantities involving user groups from π_k to π_A. In particular, we let $\widetilde{\mathbf{H}}_{k:A} = \left[\widetilde{\mathbf{H}}_{\pi_k} \ldots \widetilde{\mathbf{H}}_{\pi_A}\right]$ and $\mathbf{Q}_{k:A} = \text{diag}\left(\mathbf{Q}_{\pi_k}, \ldots, \mathbf{Q}_{\pi_A}\right)$, where $\widetilde{\mathbf{H}}_k$ is the k-th $\gamma BN \times N$ slice of $\widetilde{\mathbf{H}}$ in (3), and where $\mathbf{Q}_k = \text{diag}(q_{k,1}, \ldots, q_{k,N})$ is a $N \times N$ non-negative definite diagonal matrix.

The rate point $\{R_1(\widetilde{\mathbf{H}}), \ldots, R_A(\widetilde{\mathbf{H}})\}$ corresponding to weights $\{W_1, \ldots, W_A\}$ is obtained as solution of the max-min problem [36]–[38]

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

for the instantaneous per-user rate of each group

$$R_{\pi_k}(\widetilde{\mathbf{H}}) = \frac{1}{N} \log \frac{\left|\boldsymbol{\Sigma}(\boldsymbol{\lambda}) + \widetilde{\mathbf{H}}_{k:A} \mathbf{Q}_{k:A} \widetilde{\mathbf{H}}_{k:A}^{\mathsf{H}}\right|}{\left|\boldsymbol{\Sigma}(\boldsymbol{\lambda}) + \widetilde{\mathbf{H}}_{k+1:A} \mathbf{Q}_{k+1:A} \widetilde{\mathbf{H}}_{k+1:A}^{\mathsf{H}}\right|} \tag{6}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\lambda})$ 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 $m = 1, \ldots, B$ and the maximization with respect to $\mathbf{Q}$ is subject to the trace constraint

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

The variables $\boldsymbol{\lambda} = \{\lambda_m\}$ are the Lagrange multipliers corresponding to the per-group-of-antennas power constraints. The rate $R_{\pi_k}(\widetilde{\mathbf{H}})$ in (6) can be interpreted as the instantaneous per-user rate of user group π_k in the dual vector Multiple Access Channel (MAC) with worst-case noise defined by

$$\mathbf{r} = \sum_{k=1}^{A} \widetilde{\mathbf{H}}_k \mathbf{s}_k + \mathbf{z} \tag{8}$$

where $\text{Cov}(\mathbf{s}_k) = \mathbf{Q}_k$ and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\lambda}))$. In this "dual MAC" interpretation, group sum-rate expression (6) corresponds to group-wise successive interference cancellation, where user groups are decoded successively in the order of $\pi_1, \pi_2, \ldots, \pi_A$, and users in each group are jointly decoded. Also, notice that users in group π_k in general do not achieve individually the rate $R_{\pi_k}(\widetilde{\mathbf{H}})$ on every slot. Rather, this rate is the aggregate sum-rate of all users in group π_k, normalized by N, i.e., the mean user rate of group π_k for given $\widetilde{\mathbf{H}}$.

Efficient interior-point methods to solve (5) are given, for example, in [36]–[38]. 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, let $R_k(\widetilde{\mathbf{H}}, W_1, \ldots, W_A)$ denote the k-th user group rate at the solution of (5). Then, an inner bound to the ergodic capacity region is given by

$$\underline{\mathcal{C}}(P_1, \ldots, P_B) = \mathrm{coh} \bigcup_{W_1,\ldots,W_A \geq 0} \Big\{ \mathbf{R} : 0 \leq R_{k,i} \leq \mathbb{E}\left[R_k(\widetilde{\mathbf{H}}, W_1, \ldots, W_A) \right], \\ \forall k = 1, \ldots, A, \ \forall i = 1, \ldots, N \Big\} \tag{9}$$

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 $\underline{\mathcal{C}}(P_1, \ldots, P_B)$ 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 $N \to \infty$, 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:

$$\begin{aligned} &\text{maximize} \ \ g(\mathbf{R}) \\ &\text{subject to} \ \ \mathbf{R} \in \underline{\mathcal{C}}(P_1, \ldots, P_B) \end{aligned} \tag{10}$$

Different fairness criteria can be enforced by choosing appropriately the function $g(\cdot)$ [18]. For example, proportional fairness [15], [39], [40] 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 characterization of the ergodic rate point $\mathbf{R}^\star$ achieving the optimum in (10) is in general extremely complicated. However, by applying the general stochastic optimization framework of [16] (see also [17], 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 $\mathbf{R}^\star$. The rest of this paper is dedicated to finding an efficient method to directly compute $\mathbf{R}^\star$, 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

$$\begin{aligned} \text{maximize} \quad & \sum_{k=1}^{A} W_{\pi_k} \frac{1}{N} \mathbb{E}\left[\log \frac{\left| \mathbf{\Sigma}(\boldsymbol{\lambda}) + \widetilde{\mathbf{H}}_{k:A} \mathbf{Q}_{k:A} \widetilde{\mathbf{H}}_{k:A}^{\mathsf{H}} \right|}{\left| \mathbf{\Sigma}(\boldsymbol{\lambda}) + \widetilde{\mathbf{H}}_{k+1:A} \mathbf{Q}_{k+1:A} \widetilde{\mathbf{H}}_{k+1:A}^{\mathsf{H}} \right|} \right] \\ \text{subject to} \quad & \text{tr}(\mathbf{Q}) \leq Q \end{aligned} \tag{11}$$

where we define $Q \triangleq \sum_{m=1}^{B} \lambda_m P_m$. 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

$$F_{\mathbf{W},\boldsymbol{\lambda}}(\mathbf{Q}) = \sum_{k=1}^{A} \Delta_k \frac{1}{N} \mathbb{E}\left[\log \left| \mathbf{\Sigma}(\boldsymbol{\lambda}) + \widetilde{\mathbf{H}}_{k:A} \mathbf{Q}_{k:A} \widetilde{\mathbf{H}}_{k:A}^{\mathsf{H}} \right| \right] - W_{\pi_A} \frac{1}{N} \log |\mathbf{\Sigma}(\boldsymbol{\lambda})| \tag{12}$$

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: 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 $q_{k,1}, \ldots, q_{k,N}$. That is, for any $k = 1, \ldots, A$, 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

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

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 $q_{k,i}^\star = Q_k/N$ for all $i = 1, \ldots, N$, for some values $Q_1, \ldots, Q_A$. ■

Using Lemma 1, we restrict the optimization in (12) to block-diagonal matrices $\mathbf{Q}$ with constant diagonal blocks $\mathbf{Q}_k = \frac{Q_k}{N}\mathbf{I}$. The following lemma shows that we can restrict to strictly positive $\{\lambda_m\}$:

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

Proof: The dual variable λ_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 $\lambda_m^\star = 0$ for some m. Then,

$|\Sigma(\boldsymbol{\lambda}^\star)| = 0$ in (12) and $G_{\mathbf{W}}(\boldsymbol{\lambda}^\star)$ goes to positive infinity, which clearly cannot be the solution of the minimization with respect to $\boldsymbol{\lambda}$ in (5). Therefore, the optimal $\lambda_m^\star$ is strictly positive for all $m = 1, \cdots, B$. ■

Then we can define $\overline{\mathbf{H}}_k \stackrel{\Delta}{=} \frac{1}{\sqrt{N}}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\lambda})\widetilde{\mathbf{H}}_k$ and rewrite the objective function with some abuse of notation as

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

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

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

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

$$\frac{\partial\mathcal{L}}{\partial Q_{\pi_j}} = \sum_{k=1}^{j}\frac{\Delta_k}{N}\mathbb{E}\left[\mathrm{tr}\left(\overline{\mathbf{H}}_{\pi_j}^{\mathsf{H}}\left[\mathbf{I} + \sum_{\ell=k}^{A}\overline{\mathbf{H}}_{\pi_\ell}\overline{\mathbf{H}}_{\pi_\ell}^{\mathsf{H}}Q_{\pi_\ell}\right]^{-1}\overline{\mathbf{H}}_{\pi_j}\right)\right] \leq \xi \tag{15}$$

for $j = 1,\ldots,A$, 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 [41], the trace in (15) can be rewritten in a more convenient form

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

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

$$\mathbf{r}_{[k:A]} = \sum_{\ell=k}^{A}\sqrt{Q_{\pi_\ell}}\,\overline{\mathbf{H}}_{\pi_\ell}\mathbf{s}_\ell + \mathbf{z} \tag{17}$$

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

$$\begin{aligned}\mathsf{mmse}_{k:A}^{(j)} &= \frac{1}{N}\mathrm{tr}\left(\mathbf{I} - Q_{\pi_j}\overline{\mathbf{H}}_{\pi_j}^{\mathsf{H}}\left[\overline{\mathbf{H}}_{\pi_j}\overline{\mathbf{H}}_{\pi_j}^{\mathsf{H}}Q_{\pi_j} + \boldsymbol{\Theta}_{k:A/j}\right]^{-1}\overline{\mathbf{H}}_{\pi_j}\right)\\ &= \frac{1}{N}\mathrm{tr}\left(\left[\mathbf{I} + Q_{\pi_j}\overline{\mathbf{H}}_{\pi_j}^{\mathsf{H}}\boldsymbol{\Theta}_{k:A/j}^{-1}\overline{\mathbf{H}}_{\pi_j}\right]^{-1}\right)\end{aligned} \tag{18}$$

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

$$\xi = \frac{1}{Q} \sum_{\ell=1}^{A} \sum_{k=1}^{\ell} \Delta_k (1 - \mathbb{E}[\mathsf{mmse}_{k:A}^{(\ell)}]) \tag{19}$$

Finally, we arrive at the conditions

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

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

$$Q \sum_{k=1}^{j} \frac{\Delta_k}{N} \mathbb{E}\left[\mathrm{tr}\left(\overline{\mathbf{H}}_{\pi_j}^{\mathsf{H}} \mathbf{\Theta}_{k:A/j}^{-1} \overline{\mathbf{H}}_{\pi_j}\right)\right] \leq \sum_{\ell=1}^{A} \sum_{k=1}^{\ell} \Delta_k (1 - \mathbb{E}[\mathsf{mmse}_{k:A}^{(\ell)}]) \tag{21}$$

must hold. Eventually, we have proved the following result:

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

In finite dimension, an iterative algorithm that provably converges to the solution can be obtained as a simple modification of [33, 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 $N \to \infty$, 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 $\boldsymbol{\lambda}$ is needed.

B. Limit for $N \to \infty$

In this section, we consider problem (11) in the limit for $N \to \infty$, making use of the asymptotic random matrix results of [32]. 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) [30], [31]. 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 $\overline{\mathbf{H}}$ 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:

- $\mathcal{Q}(t)$: (dual uplink) transmit power profile such that $\mathcal{Q}(t) = Q_{\pi_k}$ for $\frac{k-1}{A} \leq t < \frac{k}{A}$.
- $\mathcal{G}(r,t)$: channel gain profile of the matrix $\overline{\mathbf{H}}$ such that $\mathcal{G}(r,t) = \beta^2_{m,\pi_k}/\lambda_m$ 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 $\Upsilon_{k:A}(t) = \mathsf{mmse}^{(j)}_{k:A}$ 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 $N \to \infty$, for each $k = 1, \ldots, A$, the SINR functions $\Gamma_{k:A}(t)$ satisfy the fixed-point equation

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

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: We apply [32, Lemma 1] to the matrix $\mathbf{I} + \sum_{\ell=k}^A \overline{\mathbf{H}}_{\pi_\ell} \overline{\mathbf{H}}^{\mathsf{H}}_{\pi_\ell} Q_{\pi_\ell}$ where $\overline{\mathbf{H}}_{k:A} = [\overline{\mathbf{H}}_{\pi_k}, \ldots, \overline{\mathbf{H}}_{\pi_A}]$ has independent non-identically distributed components. The variance profile in [32, Lemma 1] is defined as the limit of the variance of the elements of the matrix $\overline{\mathbf{H}}_{k:A}$, multiplied by the number of rows, γBN. With our normalization, the elements of each (m,ℓ)-th block of $\overline{\mathbf{H}}_{k:A}$ of size $\gamma N \times N$ have variance $\frac{\beta^2_{m,\pi_\ell}/\lambda_m}{N}$. Therefore, the variance profile for the application of [32, 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). ∎

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 $\Gamma_{k:A}(t) = \Gamma^{(j)}_{k:A}$ 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

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

Combining (23) with the already mentioned modification of the iterative algorithm of [33, 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 $k = 1, \ldots, A$. 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 $k = 1, \ldots, A$.
2) For $i = 0, 1, 2, \ldots$, iterate until the following solution settles:

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

for $j = 1, \ldots, A$, 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

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

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 $Q_k^\star = Q_k(\infty)$ 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

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

In the limit for $N \to \infty$, we can use the asymptotic analytical expression for the mutual information given in [34]. Adapting [34, Result 1] to our case, we obtain

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

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

$$u_m = \left(1 + \sum_{\ell=k}^{A} Q^\star_{\pi_\ell}(\beta^2_{m,\pi_\ell}/\lambda_m) v_\ell\right)^{-1}, \quad m = 1, \ldots, B,$$

$$v_\ell = \left(1 + \gamma \sum_{m=1}^{B} Q^\star_{\pi_\ell}(\beta^2_{m,\pi_\ell}/\lambda_m) u_m\right)^{-1}, \quad \ell = k, \ldots, A. \tag{28}$$

The proof follows from [34] based on the Girko's theorem [31] (see also [30]). 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 [37] or the infeasible-start Newton method of [38], [42]. 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 $\boldsymbol{\lambda}^\star$ must have strictly positive components by Lemma 2. Therefore, at the optimal point we have $\frac{\partial G_{\mathbf{W}}}{\partial \lambda_m}\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^\star} = 0$ for all $m = 1, \cdots, B$. It follows that the solution can be approached by gradient descent iterations where the gradient can be estimated by numerical differentiation [43]. Let $\boldsymbol{\epsilon}_m$ be a γBN-length vector for which the elements $(m-1)\gamma N + 1, \cdots, m\gamma N$ are ϵ 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 λ_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 [43]. 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 $\sum_k Q_k \leq \sum_m P_m \stackrel{\Delta}{=} P_{\rm tot}$, where $P_{\rm tot}$ 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 $m = 1, \ldots, B$. 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 2 cells and K user 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

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

satisfying the above assumption.

When these symmetry condition 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's enumerate the user groups such that the b-th symmetric block contains user groups with indices $k = (b-1)B + 1, \ldots, bB$, with corresponding constant weights $W_k = W'_b$. Then, the objective function (13) takes on the form:

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

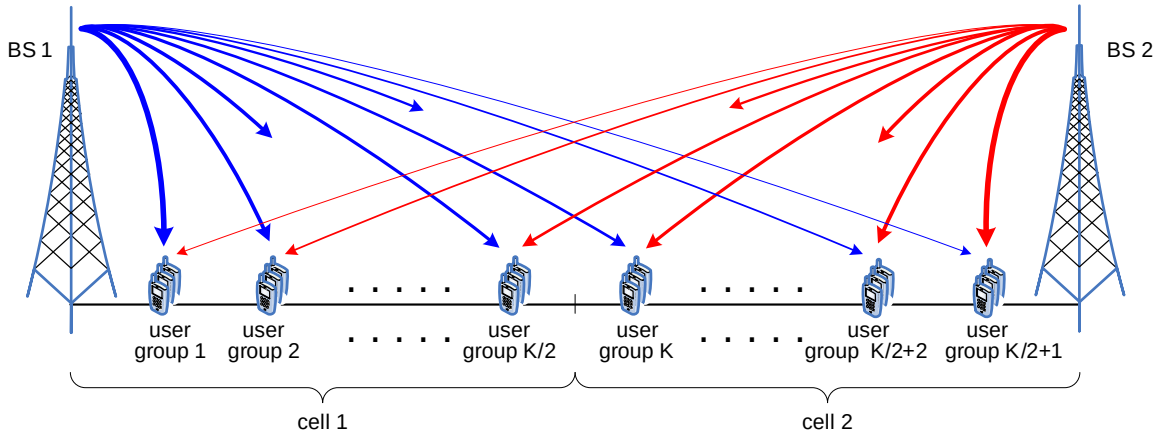


Fig. 1. A linear cellular layout with two cells and K symmetric user groups.

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

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

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 $N \to \infty$ is achieved for $\lambda_m = 1$ for all $m = 1, \ldots, B$.

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

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

are the same, for all row index $i = 1, \ldots, \gamma BN$. This means that the matrix of the element variances $\{\mathsf{P}_{i,j}\}$ corresponding to $\widetilde{\mathbf{H}}_{(b-1)B+1:A}$ is *row-regular* (see definition in [32, Definition 5]). Under the row-regularity condition, it follows that $\widetilde{\mathbf{H}}_{(b-1)B+1:A}$ in the limit of $N \to \infty$ is statistically equivalent to a matrix $\check{\mathbf{H}}_{(b-1)B+1:A} = \mathbf{G}_{(b-1)B+1:A}\mathbf{T}^{1/2}_{(b-1)B+1:A}$, 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 $\lim_{N\to\infty} \mathcal{F}_{i,b}^{(N)}(z)$. 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 $\mathbf{\Pi}$ denote a $\gamma BN \times \gamma BN$ block-permutation matrix, that permutes the B blocks of consecutive positions of length γN in the index vector $\{1,\ldots,\gamma BN\}$. Using the above asymptotic statistical equivalence, in the limit of large N we can write, for any $\{\lambda_m\}$ and $\{Q_k\}$,

$$\begin{aligned}
F_{\mathbf{W},\boldsymbol{\lambda}}(Q_1,\ldots,Q_A) &\overset{(a)}{=} \sum_{b=1}^{A'} \frac{\Delta'_b}{N}\mathbb{E}\left[\frac{1}{B!}\sum_{\mathbf{\Pi}} \log\left|\mathbf{I}+\mathbf{\Sigma}^{-1}(\boldsymbol{\lambda})\mathbf{\Pi}\check{\mathbf{H}}_{(b-1)B+1:A}\mathbf{Q}_{(b-1)B+1:A}\check{\mathbf{H}}^{\mathsf{H}}_{(b-1)B+1:A}\mathbf{\Pi}^{\mathsf{T}}\right|\right] \\
&= \sum_{b=1}^{A'} \frac{\Delta'_b}{N}\mathbb{E}\left[\frac{1}{B!}\sum_{\mathbf{\Pi}} \log\left|\mathbf{I}+\mathbf{\Pi}^{\mathsf{T}}\mathbf{\Sigma}^{-1}(\boldsymbol{\lambda})\mathbf{\Pi}\check{\mathbf{H}}_{(b-1)B+1:A}\mathbf{Q}_{(b-1)B+1:A}\check{\mathbf{H}}^{\mathsf{H}}_{(b-1)B+1:A}\right|\right] \\
&= \sum_{b=1}^{A'} \frac{\Delta'_b}{N}\mathbb{E}\left[\frac{1}{B!}\sum_{\mathbf{\Pi}} \log\left|\mathbf{I}+\left(\mathbf{\Pi}^{\mathsf{T}}\mathbf{\Sigma}(\boldsymbol{\lambda})\mathbf{\Pi}\right)^{-1}\check{\mathbf{H}}_{(b-1)B+1:A}\mathbf{Q}_{(b-1)B+1:A}\check{\mathbf{H}}^{\mathsf{H}}_{(b-1)B+1:A}\right|\right] \\
&\overset{(b)}{\geq} \sum_{b=1}^{A'} \frac{\Delta'_b}{N}\mathbb{E}\left[\log\left|\mathbf{I}+\left(\frac{1}{B!}\sum_{\mathbf{\Pi}}\mathbf{\Pi}^{\mathsf{T}}\mathbf{\Sigma}(\boldsymbol{\lambda})\mathbf{\Pi}\right)^{-1}\check{\mathbf{H}}_{(b-1)B+1:A}\mathbf{Q}_{(b-1)B+1:A}\check{\mathbf{H}}^{\mathsf{H}}_{(b-1)B+1:A}\right|\right] \\
&\overset{(c)}{=} \sum_{b=1}^{A'} \frac{\Delta'_b}{N}\mathbb{E}\left[\log\left|\mathbf{I}+\check{\mathbf{H}}_{(b-1)B+1:A}\mathbf{Q}_{(b-1)B+1:A}\check{\mathbf{H}}^{\mathsf{H}}_{(b-1)B+1:A}\right|\right] \qquad (30)
\end{aligned}$$

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 $N \to \infty$ the instantaneous rates in (5) converge to the average rates in (11), the theorem is proved. ■

IV. Fairness scheduling

Downlink opportunistic scheduling is currently used by "high data rate" third-generation cellular systems such as EV-DO [39] and HSDPA [40]. 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 $\mathbf{R}^{\star}$, the solution of (10).

The scheduling policy can be systematically designed by using the stochastic optimization approach of [16], [17], 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. In order to illustrate the scheduling mechanism we will denote *instantaneous* quantities as dependent on the slot index t. In short, the policy ensures that the virtual queue of each user (k,i) (i.e., user i in group k) is *strongly stable* (see [16, Definition 3.1]). This implies that the arrival rate $\Lambda_{k,i}$ is strictly less than the average service rate $R_{k,i} = \mathbb{E}[R_{k,i}(\widetilde{\mathbf{H}}(t))]$. Then, the desired ergodic rate point $\mathbf{R}^\star$ can be approached if the virtual queues are fed by virtual arrival processes $A_{k,i}(t)$ with arrival rates $\Lambda_{k,i} = \mathbb{E}[A_{k,i}(t)]$ sufficiently close to the desired values $R^\star_{k,i}$. The interesting feature of this approach is that it is possible to generate such virtual arrival processes adaptively, even if the values $R^\star_{k,i}$ 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

$$U_{k,i}(t+1) = \Big[U_{k,i}(t) - R_{k,i}(\widetilde{\mathbf{H}}(t))\Big]_+ + A_{k,i}(t) \tag{31}$$

We consider the scheduling policy given as follows:

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

$$\begin{aligned} &\text{maximize} \quad \sum_{k=1}^{A}\sum_{i=1}^{N} U_{k,i}(t) R_{k,i}(\widetilde{\mathbf{H}}(t)) \\ &\text{subject to} \quad \mathrm{Cov}(\mathbf{x}_m) \le P_m \end{aligned} \tag{32}$$

- The virtual queues are updated according to (31), where the arrival processes are given by $A_{k,i}(t) = a^\star_{k,i}$, where the vector $\mathbf{a}^\star$ is the solution of the maximization problem:

$$\begin{aligned} &\text{maximize} \quad Vg(\mathbf{a}) - \sum_{k=1}^{A}\sum_{i=1}^{N} a_{k,i} U_{k,i}(t) \\ &\text{subject to} \quad 0 \le a_{k,i} \le A_{\max} \end{aligned} \tag{33}$$

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

The parameters V and $A_{\max}$ determine the accuracy and the rate of convergence of the time-average rates to their expected values. It can be shown [16], [17] that, for fixed sufficiently large parameters $A_{\max}$, the

gap between the long-time average rates $\lim_{t\to\infty}\sum_{\tau=0}^{t-1}\frac{1}{t}R_{k,i}(\widetilde{\mathbf{H}}(\tau))$ and the optimal ergodic rates $R_{k,i}^{\star}$ 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 $\underline{\mathcal{C}}(P_1,\ldots,P_B)$ 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.[1] For example, the α-fairness utility function proposed in [18] 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 $\mathbf{R}^{\star}=\{R_{k,i}^{\star}\}$ solution of (10) must satisfy, for all i,

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

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.

[1]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 $\{\beta_{k,m}\}$.

A. Lagrangian optimization

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

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \max_{\mathbf{r},\mathbf{Q},\pi} \quad & g(\mathbf{r}) \\ \text{subject to} \quad & r_{\pi_k} \leq \frac{1}{N}\mathbb{E}\left[\log \frac{\left|\mathbf{I} + \sum_{\ell=k}^{A} \overline{\mathbf{H}}_{\pi_\ell} \overline{\mathbf{H}}_{\pi_\ell}^{\mathsf{H}} Q_{\pi_\ell}\right|}{\left|\mathbf{I} + \sum_{\ell=k+1}^{A} \overline{\mathbf{H}}_{\pi_\ell} \overline{\mathbf{H}}_{\pi_\ell}^{\mathsf{H}} Q_{\pi_\ell}\right|}\right], \\ & \mathrm{tr}(\mathbf{Q}) \leq Q, \quad \boldsymbol{\lambda} \geq 0 \end{aligned} \tag{34}$$

The Lagrange function for (34) is given by

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda}, \mathbf{r}, \mathbf{Q}, \pi, \mathbf{W}) &= g(\mathbf{r}) - \sum_{k=1}^{A} W_{\pi_k} \left(r_{\pi_k} - \frac{1}{N}\mathbb{E}\left[\log \frac{\left|\mathbf{I} + \sum_{\ell=k}^{A} \overline{\mathbf{H}}_{\pi_\ell} \overline{\mathbf{H}}_{\pi_\ell}^{\mathsf{H}} Q_{\pi_\ell}\right|}{\left|\mathbf{I} + \sum_{\ell=k+1}^{A} \overline{\mathbf{H}}_{\pi_\ell} \overline{\mathbf{H}}_{\pi_\ell}^{\mathsf{H}} Q_{\pi_\ell}\right|}\right] \right) \\ &= \underbrace{g(\mathbf{r}) - \sum_{k=1}^{A} W_k r_k}_{f_{\mathbf{W}}(\mathbf{r})} + \underbrace{\sum_{k=1}^{A} W_{\pi_k} \frac{1}{N}\mathbb{E}\left[\log \frac{\left|\mathbf{I} + \sum_{\ell=k}^{A} \overline{\mathbf{H}}_{\pi_\ell} \overline{\mathbf{H}}_{\pi_\ell}^{\mathsf{H}} Q_{\pi_\ell}\right|}{\left|\mathbf{I} + \sum_{\ell=k+1}^{A} \overline{\mathbf{H}}_{\pi_\ell} \overline{\mathbf{H}}_{\pi_\ell}^{\mathsf{H}} Q_{\pi_\ell}\right|}\right]}_{h_{\mathbf{W}}(\boldsymbol{\lambda},\mathbf{Q},\pi)} \end{aligned} \tag{35}$$

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 a function of $\mathbf{r}$ only, denoted by $f_{\mathbf{W}}(\mathbf{r})$, and a function of $\boldsymbol{\lambda}, \mathbf{Q}$ and π only, denoted by $h_{\mathbf{W}}(\mathbf{Q}, \pi)$. The Lagrange dual function for the problem (35) is given by

$$\begin{aligned} \mathcal{G}(\mathbf{W}) &= \min_{\boldsymbol{\lambda}} \max_{\mathbf{r},\mathbf{Q},\pi} \quad \mathcal{L}(\boldsymbol{\lambda}, \mathbf{r}, \mathbf{Q}, \pi, \mathbf{W}) \\ &= \underbrace{\max_{\mathbf{r}} \quad f_{\mathbf{W}}(\mathbf{r})}_{(a)} + \underbrace{\min_{\boldsymbol{\lambda}} \max_{\mathbf{Q},\pi} \quad h_{\mathbf{W}}(\boldsymbol{\lambda}, \mathbf{Q}, \pi)}_{(b)} \end{aligned} \tag{36}$$

and it is obtained by the decoupled maximization in (a) (with respect to $\mathbf{r}$) and the min-max in (b) (with respect to $\boldsymbol{\lambda}, \mathbf{Q}, \pi$). Notice that problems (a) and (b) correspond to the static forms of (33) and (32), 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

$$\min_{\mathbf{W} \geq \mathbf{0}} \quad \mathcal{G}(\mathbf{W}) \tag{37}$$

Eventually, the solution of (37) can be found via inner-outer iterations as follows:

Inner Problem: For given $\mathbf{W}$, we solve (36) with respect to $\boldsymbol{\lambda}$, $\mathbf{r}$, $\mathbf{Q}$ and π. 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 $N \to \infty$, this problem is solved by Algorithm 1 for fixed $\boldsymbol{\lambda} > 0$. If the system satisfies the symmetry conditions of Theorem 2 hold, then we let $\lambda_m = 1$ and no minimization with respect to λ_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 constraint 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 $\boldsymbol{\lambda}^*$, π^*, $\mathbf{Q}^*$ and $\mathbf{r}^*(\mathbf{W})$ denote [2] the solution of the inner problem for fixed $\mathbf{W}$. For any $\mathbf{W}'$, we have

$$\begin{aligned}\mathcal{G}(\mathbf{W}') &= \max_{\mathbf{r}} \; f_{\mathbf{W}'}(\mathbf{r}) + \max_{\mathbf{Q}} \; h_{\mathbf{W}'}(\boldsymbol{\lambda}^*, \mathbf{Q}, \pi^*) \\ &\geq f_{\mathbf{W}'}(\mathbf{r}^*(\mathbf{W})) + h_{\mathbf{W}'}(\boldsymbol{\lambda}^*, \mathbf{Q}^*, \pi^*) \\ &= \mathcal{G}(\mathbf{W}) + \sum_{k=1}^{A} \left(W_k' - W_k\right) \left(R_k^*(\mathbf{W}) - r_k^*(\mathbf{W})\right) \end{aligned} \tag{38}$$

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

$$W_k(n+1) = W_k(n) - \mu(n) \left(R_k^*(\mathbf{W}(n)) - r_k^*(\mathbf{W}(n))\right), \quad \forall\, k \tag{39}$$

for some step size $\mu(n) > 0$ which can be determined by a efficient algorithm such as the back-tracking line search method [44] or Ellipsoid method [45]. 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 (31). 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

[2]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 $\boldsymbol{\lambda}^*, \pi^*$ and $\mathbf{Q}^*$ also in general depend on $\mathbf{W}$.

fairness scheduling.

B. Proportional fairness scheduling

The network utility function for PFS is given as

$$g(\mathbf{r}) = \sum_{k=1}^{A} \log(r_k) \tag{40}$$

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

$$r_k^*(\mathbf{W}) = 1/W_k, \ \ \forall\, k \tag{41}$$

(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 (41), 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 $\mathbf{R}^\star$ 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 user ergodic rate. The network utility function is given by

$$g(\mathbf{r}) = \min_{k=1,\ldots,A} r_k. \tag{42}$$

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

$$\begin{aligned} \max_{\gamma, \mathbf{r} \geq 0} \quad & \gamma - \sum_{k=1}^{A} W_k r_k \\ \text{subject to} \quad & r_k \geq \gamma, \ \ \forall\, k \end{aligned} \tag{43}$$

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

$$\max_{\gamma > 0} \quad (1 - \sum_{k=1}^{A} W_k)\gamma. \tag{44}$$

Since the original maximization in (34) is bounded while (44) may be unbounded, we must have that $\sum_{k=1}^{A} W_k = 1$ and γ 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

$$W_k(n+1) = W_k(n) - \mu \left(R_k^*(\mathbf{W}(n)) - \gamma^*(\mathbf{W}(n)) \right), \quad \forall\, k \tag{45}$$

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

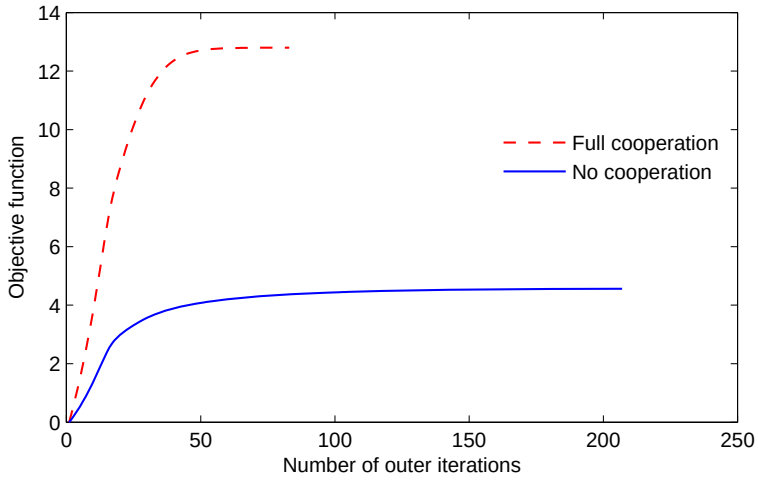
$$r_k^*(\mathbf{W}(n)) = \gamma^*(\mathbf{W}(n)) = \frac{1}{A} \sum_{j=1}^{A} R_j^*(\mathbf{W}(n)), \quad \forall\, k \tag{46}$$

V. Numerical Results

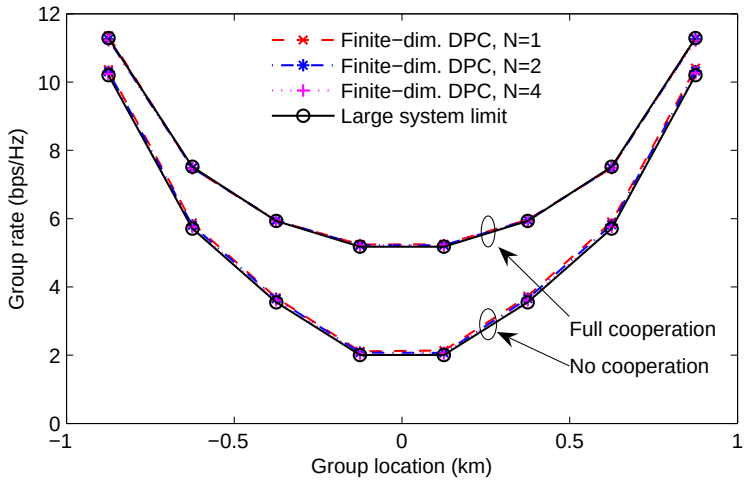
In this section we present some examples of the multi-cell model considered in this paper and compare (when possible) the numerical results using the proposed large-system analysis with the results of Monte Carlo simulation applied to an actual finite-dimensional system subject to the dynamic fairness scheduling policy outlined at the beginning of Section IV.

The examples involve a one-dimensional 2-cell model ($M = 2$) and a two-dimensional 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 [14] 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 ± 1 km, and $K = 8$ user groups equally spaced between the BSs. We consider the case of full BS cooperation and no cooperation 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 [36] combined with the dynamic scheduling policy [17] 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 the large-system asymptotic results even for very small N. Notice that the dynamic scheduling policy should provide multi-user diversity gain and in general should achieve higher



(a) Convergence of the utility function

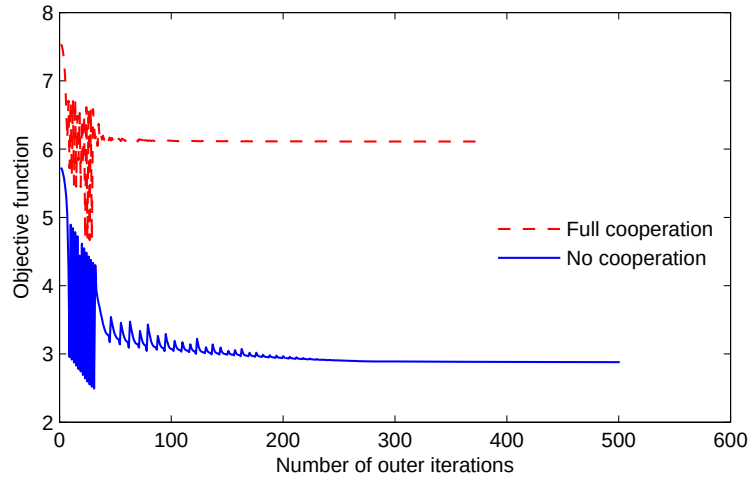


(b) Individual group rates

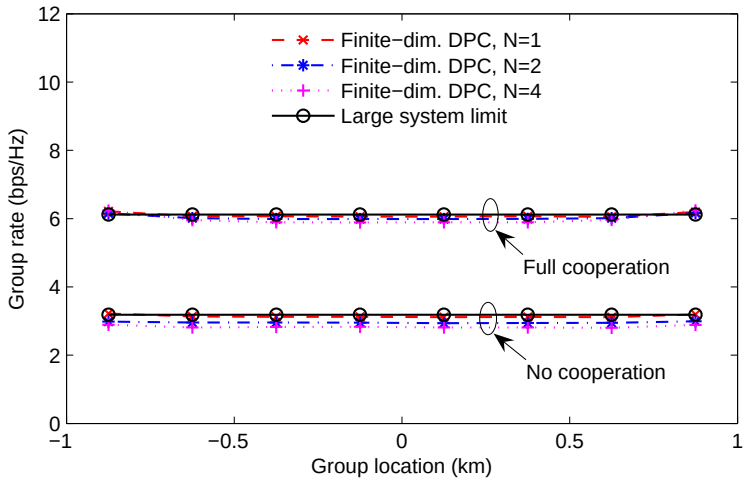
Fig. 2. Proportional fairness scheduling with $\gamma = 4$ and $K = 8$ in the 2-cell model.

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 multi-antenna channels and the number of users is not much larger than the number of BS antennas, the multi-user 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



(a) Convergence of the utility function

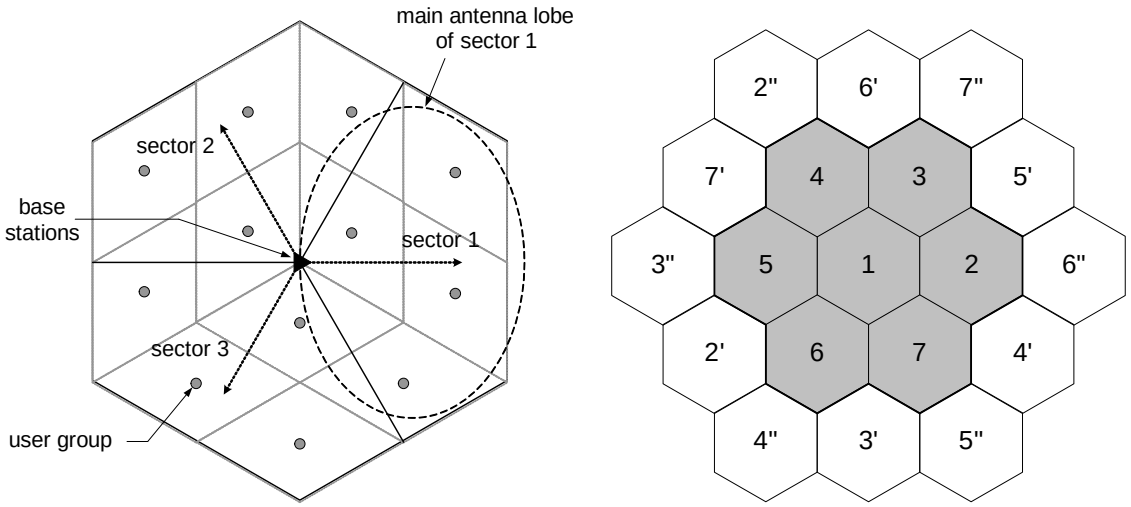


(b) Individual group rates

Fig. 3. Hard fairness scheduling with $\gamma = 4$ and $K = 8$ in the 2-cell model.

the PFS case. Also, the agreement with of the individual user rates with the finite dimensional simulation is remarkable.

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 two-dimensional cell layout where 7 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



(a) 3-sectored cell configuration

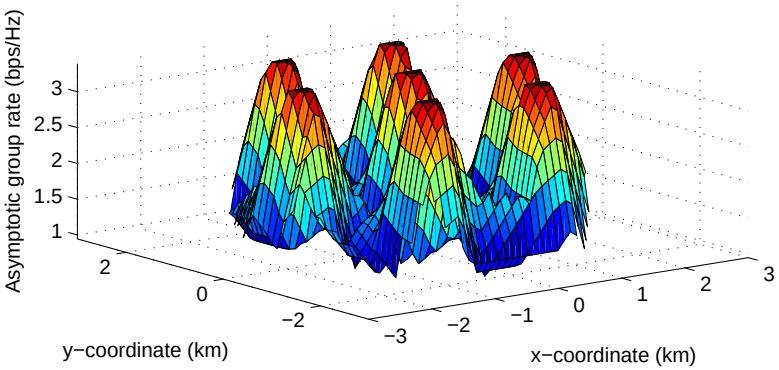
(b) Wrap-around torus topology

Fig. 4. Two-dimensional three-sectored 7-cell model.

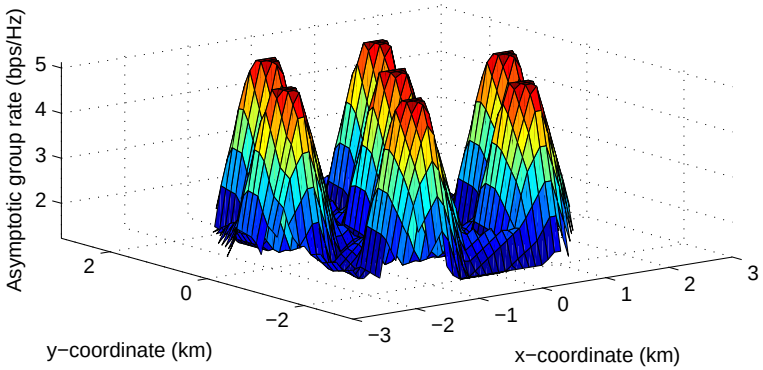
diamond-shaped equal-area grids and one user group is placed at the center of each grid. Therefore there are total $M = 21$ BSs and $K = 84$ user groups in the network. In addition, we assume a wrap-around torus topology as shown in Fig. 4(b), such that each cell is virtually surrounded by the other 6 cells and all the cells have the symmetric ICI distribution. The antenna orientation and pattern follows [46] and the non-ideal spatial antenna gain pattern (overlapping between sectors in the same cell) generates ICI even between sectors in the same cell with no cooperation. This model is relevant for a macro-cell network where both the ICI and the effective inter-cell cooperation are due to neighboring cells. Fig. 5 shows the user rate distribution under three levels of cooperation, (a) no cooperation ($L = 21$ single-sector clusters), (b) cooperation among the co-located 3 sector BSs ($L = 7$ clusters formed by three sectors of each cell), and (c) full cooperation over 7-cell network ($L = 1$). 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. Conclusions

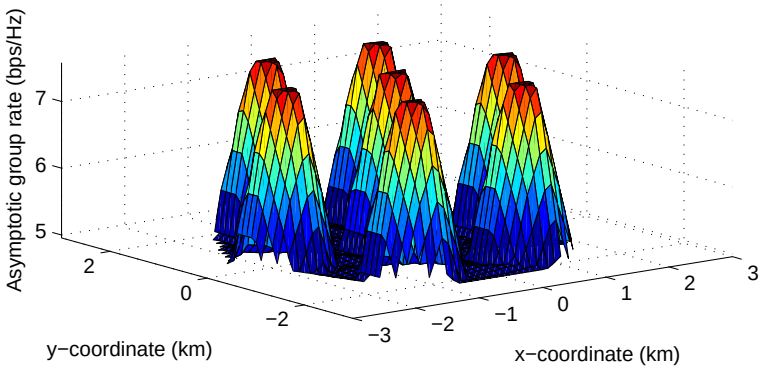
We considered the downlink of a multi-cell MU-MIMO cellular system where the pathloss and inter-cell interference make the users' channel statistics unequal. In this case, it is important to evaluate the system performance subject to some form of fairness. Downlink scheduling that make the system operate at a desired point of the long-term average achievable rate region is an important issue, widely studied and widely applied in practice [15], [39], [40]. This is classically formulated as the maximization of a



(a) No cooperation



(b) Cooperation among co-located sectors



(c) Full cooperation over 7 cells

Fig. 5. Ergodic user rate distribution in the 7-cell model.

concave network utility function over the achievable ergodic rate region of the system. We also considered an inter-cell cooperation scheme for which groups of cells operates jointly, as a distributed multi-antenna transmitter, and have perfect channel state information for all users in their cluster and only statistical information on the inter-cluster interference. Under the constraint that inter-cluster interference is treated as noise, this model is quite general.

We focused on the large-system limit 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 the optimal fairness rate point, based on a combination of large random matrix results and Lagrangian optimization. The proposed method is particularly simple and efficient in the case where the system has certain symmetries. Otherwise, we can obtain a simple upper bound by relaxing the per-base station power constraint to the per-cluster sum-power constraint. Numerical results showed that the rates predicted by the large-system analysis are indeed remarkably close to the rates by Monte Carlo simulation of a corresponding finite-dimensional system. Overall, the results of this paper are useful in evaluating and optimizing the multi-cell MU-MIMO systems, especially when the system dimension and network size are large.

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., March 2010.

[3] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. on Inform. Theory*, vol. 49, pp. 1691–1706, July 2003.

[4] P. Viswanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. on Inform. Theory*, vol. 49, 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. on Inform. Theory*, vol. 49, pp. 2658–2668, Oct. 2003.

[6] W. Yu and J. Cioffi, "Sum capacity of Gaussian vector broadcast channels," *IEEE Trans. on Inform. Theory*, vol. 50, pp. 1875–1892, Sept. 2004.

[7] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. on Inform. Theory*, vol. 52, pp. 3936–3964, Sept. 2006.

[8] A. D. Wyner, "Shannon-theoretic approach to a Gaussian cellular multiple access channel," *IEEE Trans. on Inform. Theory*, vol. 40, pp. 1713–1727, Nov. 1994.

[9] S. Shamai (Shitz) and A. D. Wyner, "Information-theoretic considerations for symmetric, cellular, multiple-access fading channels – Part I & II," *IEEE Trans. on Inform. Theory*, vol. 43, pp. 1877–1894, Nov. 1997.

[10] O. Somekh and S. Shamai (Shitz), "Shannon-theoretic approach to a Gaussian cellular multiple-access channel with fading," *IEEE Trans. on Inform. Theory*, vol. 46, pp. 1401–1425, July 2000.

[11] O. Somekh, B. M. Zaidel, and S. Shamai (Shitz), "Sum rate characterization of joint multiple cell-site processing," *IEEE Trans. on Inform. Theory*, vol. 53, pp. 4473–4497, December 2007.

[12] 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. on Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Cannes, France, Sept. 2008.

[13] A. Sanderovich, O. Somekh, H. Poor, and S. Shamai (Shitz), "Uplink macro diversity of limited backhaul cellular network," *IEEE Trans. on Inform. Theory*, vol. 55, pp. 3457–3478, Aug. 2009.

[14] WiMAX Forum, "Mobile WiMAX – Part I: A technical overview and performance evaluation," Tech. Rep., Aug. 2006.

[15] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. on Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[16] L. Georgiadis, M. Neely, and L. Tassiulas, *Resource Allocation and Cross-Layer Control in Wireless Networks*. Foundations and Trends in Networking, 2006, vol. 1, no. 1.

[17] H. Shirani-Mehr, G. Caire, and M. J. Neely, "MIMO downlink scheduling with non-perfect channel state knowledge," *accepted for IEEE Trans. on Commun. (posted on arXiv:0904.1409 [cs.IT])*, 2009.

[18] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. on Networking*, vol. 8, pp. 556–567, Oct. 2000.

[19] H. Huang and R. Valenzuela, "Fundamental simulated performance of downlink fixed wireless cellular networks with multiple antennas," in *Proc. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Berlin, Germany, Sept. 2005.

[20] F. Boccardi and H. Huang, "Limited downlink network coordination in cellular networks," in *Proc. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Athens, Greece, Sept. 2007.

[21] H. Zhang, N. Mehta, A. Molisch, J. Zhang, and H. Dai, "Asynchronous interference mitigation in cooperative base station systems," *IEEE Trans. on Wireless Commun.*, vol. 7, pp. 155–165, Jan. 2008.

[22] G. Caire, S. Ramprashad, 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, Sept. 2008.

[23] 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.

[24] J. Zhang, R. Chen, J. Andrews, A. Ghosh, and R. Heath, "Networked MIMO with clustered linear precoding," *IEEE Trans. on Wireless Commun.*, vol. 8, pp. 1910–1921, April 2009.

[25] 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, June 2009.

[26] S. A. Ramprashad 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, Sept. 2009.

[27] 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, Sept. 2008.

[28] J.-B. Landre, A. Saadani, and F. Ortolan, "Realistic performance of HSDPA MIMO in macro-cell environment," in *Proc. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Tokyo, Japan, Sept. 2009.

[29] 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.

[30] A. M. Tulino and S. Verdu, *Random Matrix Theory and Wireless Communications*. Foundations and Trends in Communications and Information Theory, 2004, vol. 1, no. 1.

[31] V. L. Girko, *Theory of Random Determinants*. Kluwer Academic Publishers, Doordrecht and Boston, 1990.

[32] A. M. Tulino, A. Lozano, and S. Verdu, "Impact of antenna correlation on the capacity of multiantenna channels," *IEEE Trans. on Inform. Theory*, vol. 7, pp. 2491–2509, July 2005.

[33] ——, "Capacity-achieving input covariance for single-user multi-antenna channels," *IEEE Trans. on Wireless Commun.*, vol. 5, pp. 662–671, March 2006.

[34] 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. on Inform. Theory*, vol. 52, pp. 3264–3274, July 2006.

[35] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[36] W. Yu, "Sum-capacity computation for the Gaussian vector broadcast channel via dual decomposition," *IEEE Trans. on Inform. Theory*, vol. 52, pp. 754–759, Feb. 2006.

[37] 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. on Inform. Theory (ISIT)*, Seoul, Korea, June 2009.

[38] H. Huh, H. C. Papadopoulos, and G. Caire, "Multiuser MIMO transmitter optimization for inter-cell interference mitigation," *accepted for IEEE Trans. on Sig. Proc. (posted on arXiv:0909.1344v1[cs.IT])*, 2009.

[39] 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, July 2000.

[40] 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.

[41] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[42] W. Yu and T. Lan, "Transmitter optimization for the multi-antenna downlink with per-antenna power constraints," *IEEE Trans. on Sig. Proc.*, vol. 55, pp. 2646–2660, June 2007.

[43] W. Cheney and D. Kincaid, *Nemerical Mathematics and Computing*. Thomson Brooks/Cole, 2004.

[44] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[45] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. Athena Scientific, 1997.

[46] IEEE 802.16 broadband wireless access working group, "IEEE 802.16m evaluation methodology document (EMD)," IEEE 802.16m-08/004, Tech. Rep., Jan. 2009.