Self-organizing Dynamic Fractional Frequency Reuse Through Distributed Inter-cell Coordination: The Case of Best-Effort Traffic

Alexander L. Stolyar
Bell Labs, Alcatel-Lucent
Murray Hill, NJ 07974
stolyar@research.bell-labs.com

Harish Viswanathan
Bell Labs, Alcatel-Lucent
Murray Hill, NJ 07974
harishv@research.bell-labs.com

May 29, 2008

Abstract

Improving cell edge data rates and self-optimization of the network are important objectives for next generation cellular systems. Towards realizing these goals, we propose algorithms that automatically create efficient, soft fractional frequency reuse (FFR) patterns for enhancing performance of orthogonal frequency division multiple access (OFDMA) based cellular systems for forward link best effort (BE) traffic, in particular for cell edge users. The Multi-sector Gradient (MGR) algorithm adjusts the transmit powers of the different sub-bands by systematically pursuing local maximization of the overall network utility. We show that the maximization can be done by sectors operating in a semi-autonomous way, with only some gradient information exchanged periodically by neighboring sectors. The Sector Autonomous (SA) algorithm adjusts its transmit powers in each sub-band independently in each sector using a non-trivial heuristic to achieve out-of-cell interference mitigation. This algorithm is completely autonomous and requires no exchange of information between sectors. Through extensive simulations, we demonstrate that both algorithms can improve the cell edge data throughputs significantly, by up to 66% in some cases for the MGR, while maintaining the overall sector throughput at the same level as that achieved by the traditional approach. The simulations also show that both algorithms lead the system to "self-organize" into efficient, soft frequency reuse patterns with no a priori frequency planning.

1 Introduction

Fourth generation cellular systems are currently being developed and will be deployed in a few years time. These systems target significantly higher sector capacities and higher per user data rates compared to third generation systems. In particular, one of the goals of these systems is to boost performance of users at the cell edge that typically suffer from significant out-of-cell interference. A variety of innovations including multiple input multiple output (MIMO) multi-antenna techniques and use of wider signaling bandwidths are being adopted to achieve the desired level of performance.

Having approached the information-theoretic limits of point-to-point communication through coding and MIMO techniques, further advances in cellular performance requires focusing the attention on eliminating interference efficiently. In particular, users at the cell edge will benefit significantly from interference reduction techniques. Two broad classes of techniques for interference mitigation are interference cancelation and interference avoidance. Interference cancelation relies on coding and signal processing at the transmitter or receiver to suppress interference. On the other hand, interference avoidance relies on intelligent resource allocation to eliminate interference.

One of the other important emerging trends in cellular systems is the increasingly diverse set of deployment scenarios. For many such deployment scenarios, base station location and orientations are largely unplanned or are governed by significant constraints that prevent optimal location and orientation. Examples include femto cells where end users deploy their own cells, and small out-door pico-cells where the base station location is governed by availability of digital subscriber line (DSL) access or optical fibre termination. Interference mitigation schemes become particularly important for such deployments. Furthermore, self-configuration and optimization of the network are critical for the success of such deployments.

In [10] we proposed a self-organizing interference avoidance scheme for constant bit rate traffic in orthogonal frequency division multiple access (OFDMA) systems through selfish optimization of resources by each sector, and demonstrated that efficient fractional frequency reuse (FFR) patterns could be achieved dynamically. In a similar vein, in this paper, we propose algorithms for improving the throughput performance for best effort traffic in OFDMA cellular systems through formation of FFR patterns automatically. We propose two different algorithms, namely the *Multi-cell Gradient* (MGR) that requires some information to be exchanged between neighboring sectors, and *Sector Autonomous* (SA) that is completely distributed and requires no exchange of information.

MGR algorithm adjusts the transmit powers of the different sub-bands by systematically pursuing local maximization of the overall network utility. We show that the maximization can be done semi-autonomously by each sector with only periodic exchange between interfering sectors of a few key ratios that naturally arise from the optimization approach. The computations are still distributed and performed independently in each sector. Since only neighboring base stations need to communicate and only a limited amount of information is exchanged between them, the proposed schemes can be implemented in practice.

SA algorithm, on the other hand, adjusts transmit powers in each sub-band independently in each sector using a non-trivial heuristic to achieve out-of-cell interference mitigation. This algorithm is completely autonomous and requires no exchange of information between sectors. Such an algorithm may be desirable when it is not possible to exchange any information between the relevant sectors. MGR, of course, outperforms the SA algorithm.

Both MGR and SA algorithms are only concerned with the power allocation (and reallocation) among the sub-bands by each sector, which is done on a relatively slow time scale. Given the power levels set by either algorithm, each sector can perform an opportunistic, channel-aware scheduling, taking advantage of the fast fading by proper assignment of users to sub-bands (on the fast time scale). In fact, this is the main scenario used in our simulations. We demonstrate through simulations that the performance of MGR and SA algorithms, when compared to that of the standard "universal reuse" (UNIVERSAL) approach where equal powers are assigned to each sub-band in each cell and channel-aware fast time scale scheduling is utilized within each sector, is significantly better especially in increasing cell edge user throughputs.

In addition to proposing novel algorithms, we also provide several related insights in this paper. Using a simple scenario, we illustrate that when the channel fading is present, *any* power allocation approach, even equal power allocation across the sub-bands as in the UNIVERSAL algorithm, benefits from some level of interference avoidance due to fast channel-aware scheduling and proper fast (re)assignment of users to sub-bands. However, our algorithms perform better by allocating the available transmit power across the different sub-bands efficiently. We also show, albeit in an extreme case, that selfish utility maximization by each of the sectors independently may not lead to efficient FFR patterns. This led us to the heuristic SA algorithm proposed in this paper. Finally, as part of MGR approach, we propose and rigorously substantiate an efficient – "virtual scheduling" – algorithm, which allows efficient real-time computation by each sector of the gradient of the system utility function with respect to the current sub-band transmit powers in the sector.

The paper is organized as follows. In Section 1.1 we briefly discuss some related work. In Section 2 we describe the system model under consideration and provide the overview of the proposed algorithms. Section 3 defines the MGR algorithm, with Sections 4 and 5 addressing its key part - the virtual scheduling algorithm for the utility gradient estimation. (The technical development substantiating the gradient calculation is relegated to the Appendix.) In Section 6 we define SA algorithm. The numerous simulation studies, comparing the performance of MGR, SA and UNIVERSAL algorithms in a realistic setting are given in Section 7. The phenomenon of "automatic interference aviodance" due to channel-aware scheduling in the case of fast fading, is illustrated in Section 8. We conclude with a summary and discussion of future work in Section 9.

1.1 Related Work

Numerous papers have been published on scheduling in OFDMA systems. However, most of these papers are focused on single cell scheduling and typically do not consider the effect of out-of-cell interference. Several papers [7], [5], [3] have been published on coordinated scheduling, although not in the context of OFDMA. These papers propose algorithms that are centralized and are not based on simple exchange of messages between sectors as in this paper. Dynamic distributed resource allocation in the context of Gaussian interference channels has been considered in [4] and [1]. Neither of these papers consider the model of this paper with multiple interfering base stations each serving several, differently located users. (As a result, in our model, even within the same cell, different users experience different interference levels in different resource sets.) The concept of FFR for best effort traffic in the context of OFDMA systems has appeared in cellular network standardization fora technical contributions [11], [12] and in [6]. As mentioned earlier, we proposed and studied a self-organizing FFR scheme for constant bit rate traffic such as voice over Internet Protocol (VoIP) in our prior work [10].

2 System model

2.1 OFDMA description and key assumptions

We begin with a very brief description of an OFDMA system from [10]. In an OFDMA system the transmission band is divided into a number of sub-carriers and information is transmitted by modulating each of the sub-carriers. Further, time is divided into slots consisting of a number of OFDM symbols and transmissions are scheduled to users by assigning a set of sub-carriers on specific slots. The frequency resources scheduled are usually logical sub-carriers. The logical sub-carriers are mapped to physical sub-carriers for transmission. The mapping can change from time to time and is referred to as frequency hopping. Frequency hopping is employed to achieve interference averaging.

OFDMA systems supporting FFR for interference mitigation divide frequency and time resources into several *sub-bands*. Frequency hopping of sub-carriers is restricted to be within the sub-band so that users scheduled on a certain sub-band experience interference only from users scheduled in neighboring sectors in the same sub-band. Soft fractional frequency reuse can be achieved by setting the transmit power on each sub-band in each sector in a manner that suppresses inter-sector interference. Note that sub-band is a special case of a *resource set* which could be a combination of a set of sub-carriers in frequency and a set of time-slots. FFR can be implemented using resources sets instead of sub-bands. All of the results presented in this paper can be extended to the notion of resource sets straightforwardly.

Another important aspect of the system, which is assumed in the model described below, is the so called channel quality indicator feedback that is sent by the mobiles back to the base station for the purpose of resource allocation. The feedback is used by the channel aware scheduler to select users for each of the sub-carriers for transmission in each slot, and also to determine the transmission modulation format and channel code rate for the selected users. For this purpose, relatively frequent channel quality feedback is required. In addition to this, relatively infrequent feedback indicating the level of interference experienced in each of the sub-bands is also required for our algorithms. Average signal-to-interference-and-noise ratio (SINR) for each sub-band is assumed to be fed back relatively infrequently (for example, once every 500 slots or 1/2 second in the simulations) for this purpose. Additional feedback that is unique to the MGR algorithm is also fed back infrequently. The required information corresponds to pathloss ratio between the signal and interfering base stations. This is further explained in Section 7.1.5.

2.2 Formal model

We have K cells (sectors) $k \in \mathcal{K} = \{1, \dots, K\}$, and J sub-bands $j \in \mathcal{J} = \{1, \dots, J\}$ in the system. We assume that each sub-band consists of a fixed number c of sub-carriers, and denote by W the bandwidth of one sub-band. The noise spectral density is denoted by N_0.

Time is slotted, so that transmissions within each cell are synchronized, and do not interfere with each other. A transmission in a cell, assigned to a sub-band, causes interference to only those users in other cells, that are assigned to the same sub-band; there is no inter-sub-band interference.

The *utility* $\mathcal{U}$ of the system (or network) is defined as the sum

$$\mathcal{U} = \sum_k U^{(k)}$$

of utilities $U^{(k)}$ of all sectors. In turn, sector k utility $U^{(k)}$ is a smooth concave function of the *average rates* X_i of users i served by the sector k. The precise conditions on a sector utility function, will be specified in Section 4; for example, it can be $U^{(k)} = \sum_i \log X_i$ (with the summation over users i within sector k).

We denote by $P_j^{(k)}$ the power allocated in sub-band j of sector k. The total power within each sector is upper bounded by P^*, so that $\sum_j P_j^{(k)} \leq P^*$.

The system objective is to maximize the total utility $\mathcal{U}$, by setting and adjusting the power levels $P_j^{(k)}$. The exact solution to this problem is very difficult to obtain, even using centralized schemes, as the problem is "highly non-convex." In addition, any practical algorithm should involve very limited real-time information exchange (signaling) among sectors.

The two different algorithms, MGR and SA, that we propose in this paper are such that the power levels $P_j^{(k)}$ are adjusted over time (relatively slowly) with the purpose of improving the system utility, given current set of the users in the system and their current sector assignments. MGR tries to imitate the gradient ascend method, and involves some inter-sector/cell information exchange. Our main contribution in MGR is the *virtual scheduling* algorithm, which constantly estimates the partial derivatives $\partial\mathcal{U}/\partial P_j^{(k)}$ in a very efficient and "distributed" way. Algorithm SA does *not* involve any inter-sector/cell signaling, and is based on reasonable (but not straightforward) heuristics.

3 MGR: Dynamic power allocation algorithm with base station coordination

We now describe the MGR algorithm, according to which sectors dynamically allocate/reallocate the power levels among sub-bands. The algorithm involves sectors (base stations) exchanging information information on how "costly" to their utility is the interference caused by other sectors. (We describe the algorithm as if each sector shares this information with *all* other sectors; in reality, and in our simulations, each sector exchanges information only with a small number of its neighboring sectors.)

The idea of the algorithm is simple. Each sector k constantly adjusts its power allocation to different sub-bands in a way that improves the total utility $\mathcal{U} = \sum_m U^{(m)}$ of the system.

MGR ALGORITHM (SUB-BAND POWER ADJUSTMENT PART):

Each sector $k \in \mathcal{K}$ maintains the *estimate* of the utility $U^{(k)}$ which the sector *could potentially* attain, given its current power allocation among sub-bands, $P_j^{(k)}$, $j \in \mathcal{J}$, $\sum_j P_j^{(k)} \le P^*$, and current interference level from other sectors. Moreover, sector k maintains estimates of partial derivatives $D_j^{(m,k)} = \partial U^{(k)}/\partial P_j^{(m)}$ of its (maximum attainable) utility on the power levels $P_j^{(m)}$ in all sectors m (including self, $m = k$) and all sub-bands j. The key part of the algorithm, and our key contribution, is *how* these estimates are computed; the *virtual scheduling* algorithm which does that is described in detail in Section 5 (which in turn relies on the results of Section 4).

Sector k periodically sends values of $D_j^{(m,k)}$, for all j, to each sector $m \neq k$. Correspondingly, it also periodically *receives* the values of $D_j^{(k,m)}$, for all j, from each sector $m \neq k$. (The frequency of such exchange does *not* have to be high. See Section 7.1.5.)

Sector k maintains the current values of

$$D_j^k = \sum_m D_j^{(k,m)}, \quad \text{for each sub-band } j. \tag{1}$$

Clearly, D_j^k is the estimate of the partial derivative $\partial\mathcal{U}/\partial P_j^{(m)}$.

In each physical time slot (or more generally, every n_p physical slots), sector k does the following. We use fixed parameter $\Delta > 0$, and denote by $P^{(k)} = \sum_j P_j^{(k)}$ the current total power in the sector.

Then, the powers updated, sequentially, as follows:

1. We pick j_* (if such exists) such that $D^k_{j_*}$ is the smallest among those j with $D^k_j < 0$ and $P^{(k)}_j > 0$, and do
$$P^{(k)}_{j_*} \doteq \max\{P^{(k)}_{j_*} - \Delta, 0\}.$$

2. If $P^{(k)} < P^*$, we pick j^* (if such exists) such that $D^k_{j^*}$ is the largest among those j with $D^k_j > 0$, and do
$$P^{(k)}_{j^*} \doteq P^{(k)}_{j^*} + \min\{\Delta, P^* - P^{(k)}\}.$$

3. If $P^{(k)} = P^*$ and $\max_j D^k_j > 0$, we pick a pair (j_*, j^*) (if such exists) such that $D^k_{j^*}$ is the largest, $D^k_{j_*}$ is the smallest among those with $P^{(k)}_j > 0$, and $D^k_{j_*} < D^k_{j^*}$. Then,
$$P^{(k)}_{j_*} \doteq \max\{P^{(k)}_{j_*} - \Delta, 0\},$$
$$P^{(k)}_{j^*} \doteq P^{(k)}_{j^*} + \min\{\Delta, P^{(k)}_{j_*}\}.$$

The initial values are $P^{(k)}_j = P^*/J$. The algorithm runs "continuously", and, therefore, the choice of initial state - at the system start-up or reset - is not crucial.

END ALGORITHM

We want to emphasize the fact that the power adjustment algorithm, as well the virtual scheduling algorithm (being its part), works with estimated maximal possible utility a sector can potentially attain (given current power levels), and not the actual current utility. If power allocations in the system converge, and stay approximately constant, then the "virtual utilities," used by the algorithms run in sectors, will be close to actual ones. However, the system is very dynamic, with users arriving, departing, and moving from sector to sector. As a result, the actual sector utilities can "lag behind" the optimal ones for the current power levels. Virtual utilities estimate the optimal utilities, and thus better determine the desired directions of power adjustments.

4 Differentiability of a sector utility function on available transmission rates

In this section we consider a fixed sector k, and study the dependence of its utility U on the rates R_{ij}, where R_{ij} is the rate available to user i (*in this sector*) in subbabnd j, if this user is chosen for transmission in a time slot. (We assume that rates R_{ij} do not change with time.) More specifically, we derive the expression for the partial derivative $(\partial/\partial R_{ij})U$. To simplify the notation, within this Section 4, we suppress sector index k in the variables, including $U^{(k)}$.

The users in the sector are indexed by $i \in \mathcal{I} = \{1, \ldots, N\}$. In each time slot, for each sub-band one user is chosen to transmit data to; $R_{ij} \in [0, B]$, $B < \infty$, is the transmission rate in sub-band j to user i, if this user is chosen. We will denote $R = \{R_{ij},\ i \in \mathcal{I},\ j \in \mathcal{J}\}$. A scheduling algorithm runs over many time slots. Denote by $\phi_{ij} \in [0, 1]$ the fraction of time an algorithm chooses user i for transmission in sub-band j. (A scheduling algorithm does not have to - and typically does not -

allocate those fractions explicitly; typically, they are what they turn out to be under the algorithm.) Naturally,

$$\sum_i \phi_{ij} \le 1, \quad \forall j.$$

Then the average rate user i actually receives is

$$X_i = \sum_j \phi_{ij} R_{ij}, \quad \forall i. \tag{2}$$

Given R, the set of all vectors $X = (X_1, \dots, X_N)$ for all possible $\phi = \{\phi_{ij},\ i \in \mathcal{I},\ j \in \mathcal{J}\}$, is a convex compact set $V = V(R)$ in the positive orthant $\mathbb{R}_+^N$. Clearly, $X_i \in [0, JB]$ for all i, for any $X \in V(R)$ and any R.

The utility function $U(X)$ of the average rate vector X can be one of two types. Let constant $A > JB$ be fixed. Type-1 function $U(X)$ is continuously differentiable, strictly concave, strictly increasing in each argument, and defined for $X \in [0, A]^N$. Type-2 function $U(X)$ has the form

$$U(X) = \sum_i U_i(X_i),$$

where each U_i is continuously differentiable, strictly concave, strictly increasing function in $(0, A]$, and either $U_i'(X_i) \uparrow U_i'(0) < +\infty$ as $X_i \downarrow 0$ (and then necessarily $U_i(X_i) \downarrow U_i(0) > -\infty$), or $U_i(X_i) \downarrow U_i(0) = -\infty$ as $X_i \downarrow 0$ (and then necessarily $U_i'(X_i) \uparrow +\infty$). A standard example of type-2 utility function is $U(X) = \sum_i \log X_i$.

For each R and corresponding region $V(R)$, consider the unique vector

$$X(R) = \arg\max_{X \in V(R)} U(X).$$

The uniqueness follows from convexity of $V(R)$ and strict concavity of U.

The question is: what is the expression for $(\partial/\partial R_{ij}) U(X(R))$? To gain intuition, consider a ϕ corresponding to $X(R)$, i.e. ϕ satisfying (2) with $X(R)$ in place of X. Then, if we formally differentiate $U(X(R))$ on R_{ij}, using (2) and assuming ϕ is constant, we obtain

$$\frac{\partial}{\partial R_{ij}} U(X(R)) = \frac{\partial U}{\partial X_i}(X(R)) \phi_{ij}. \tag{3}$$

This is not a proof, of course, and in fact (3) does not always hold. However, we can prove that, "typically", (3) does hold. The formal result, Theorem 10.4, is presented and proved in the appendix.

5 Sensitivity of a sector utility to power changes

5.1 General expressions

As in Section 4, we consider a fixed sector k, and use the same notations with suppressed index k: the users $i \in \mathcal{I} = \{1, \dots, N\}$ are those in the sector; their average throughputs are X_i; R_{ij} are per-user, per-sub-band rates (nominal, i.e., if user is selected); $U(X)$ is the sector utility function, defined also as in Section 4. However, for the per-sector, per-sub-band powers $\{P_j^{(m)},\ j \in \mathcal{J}, m \in \mathcal{K}\}$ we will retain the sector index m.

Let us denote by $G_i^{(m)}$ the propagation gain from sector m to user i. For the purposes of determining the sensitivity of the sector utility to power changes we assume that the propagation gains are *not* dependent on the sub-band. The values $G_i^{(m)}$ represent the channel gains averaged over the fast fading. This is because the goal of the algorithm is to adapt the transmit power levels to average interference levels and not to track the fast fading. Correspondingly, the instantaneous rates R_{ij} are the rates *as they would be* if the channel gains $G_i^{(m)}$ were constant.

Our goal is to derive expressions for the partial derivatives $\partial/\partial P_j^{(m)}[U(X)]$ for a sub-band j and all sectors $m \in \mathcal{K}$, including $m = k$. We have the following general expression (using (3) and assuming the set R is "typical" in the sense of Theorem 10.4):

$$D_j^{(m,k)} \doteq \frac{\partial}{\partial P_j^{(m)}} U(X) = \sum_i \frac{\partial U}{\partial X_i}(X)\phi_{ij}\frac{\partial R_{ij}}{\partial P_j^{(m)}}. \qquad (4)$$

Thus, we need expressions for $\partial R_{ij}/\partial P_j^{(m)}$. We use Shannon formula for the rate

$$R_{ij} = W \log_2\left(1 + \frac{G_i^{(k)}P_j^{(k)}}{N_0 W + \sum_{m\neq k} G_i^{(m)}P_j^{(m)}}\right) = H(F_{ij}(P)), \qquad (5)$$

where N_0 is noise spectral density and W is the sub-band bandwidth, and

$$H(y) \doteq W\log_2(1+y), \quad F_{ij}(P) \doteq \frac{G_i^{(k)}P_j^{(k)}}{N_0 W + \sum_{m\neq k} G_i^{(m)}P_j^{(m)}},$$

and $G_i^{(m)}$ is the propagation gain from sector m to user i. Thus,

$$\frac{\partial R_{ij}}{\partial P_j^{(m)}} = H'(F_{ij}(P))\frac{\partial F_{ij}(P)}{\partial P_j^{(m)}}. \qquad (6)$$

Finally, given the form of function F_{ij}, we easily obtain

$$\frac{\partial F_{ij}(P)}{\partial P_j^{(k)}} = \frac{F_{ij}(P)}{P_j^{(k)}}, \qquad (7)$$

$$\frac{\partial F_{ij}(P)}{\partial P_j^{(m)}} = -\frac{[F_{ij}(P)]^2}{P_j^{(k)}}\frac{G_i^{(m)}}{G_i^{(k)}}, \quad \text{if } m \neq k. \qquad (8)$$

The important observation about (6)-(8) is that these expressions *can be easily evaluated by the sector k controller*, because the values of $P_j^{(k)}$ and $F_{ij}(P)$ are directly available to it, and the ratios $\frac{G_i^{(m)}}{G_i^{(k)}}$ of propagation gains for each user i can be evaluated by the user (from the pilot power measurements) and reported to the controller (see Section 7.1.5).

5.2 Virtual scheduling to estimate sensitivity to power changes

In Section 5.1 we have shown that the sensitivity of a sector k utility to changes in power levels $P_j^{(m)}$ (in all sectors m and sub-bands j), is "typically" given by (4), where the values of partial derivatives $\partial R_{ij}/\partial P_j^{(m)}$ in the RHS are available to sector k controller. The question remains, how the controller can compute or estimate the optimal values of the fractions ϕ_{ij}, maximizing the sector utility $U(X)$? These fractions are hard to find analytically.

Our approach is as follows. To estimate and update the values of partial derivatives $D_j^{(m,k)}$ in (4), for all m and j "simultaneously," each sector k continuously runs a *virtual scheduling* algorithm which is known to (asymptotically) maximize the sector utility. This is a well-known *gradient scheduling algorithm* (see [8] and references therein). In the special case of $U(X) = \sum_i \log X_i$, it is the *proportional fair* algorithm.

MGR ALGORITHM (VIRTUAL SCHEDULING AND $D_j^{(m,k)}$ ESTIMATION):

The algorithm is run by each sector k independently, over a sequence of "virtual time slots." (The algorithm runs a fixed number n_v of virtual slots within each physical time slot of the system. The greater the n_v the greater the accuracy of the algorithm and its responsiveness to changes in system state; but, the computational burden is greater as well.) The algorithm maintains the current values X_i of average user (virtual) throughputs, and current values of $D_j^{(m,k)}$. It uses small averaging parameters $\beta_1, \beta_2 > 0$, which are chosen in conjunction with n_v. As a general rule, as n_v changes, the product $\beta_j n_v$ has to be kept constant.

In each virtual time slot, we sequentially pick each sub-band j and perform the following steps.

1. Choose user i^*,

$$i^* \in \arg\max_i \frac{\partial U}{\partial X_i}(X) R_{ij}.$$

2. Update:

$$X_{i^*} = \beta_1 J R_{i^*j} + (1-\beta_1) X_{i^*},$$
$$X_i = (1-\beta_1) X_i, \quad \text{for all } i \neq i^*.$$

3. For each m (including $m = k$, that is the sector itself), we update:

$$D_j^{(m,k)} = \beta_2 \frac{\partial U}{\partial X_{i^*}}(X) \frac{\partial R_{i^*,j}}{\partial P_j^{(m)}} + (1-\beta_2) D_j^{(m,k)}. \tag{9}$$

The initial values of the variables are chosen in some arbitrary, but reasonable way (so that their absolute values are not much larger than "correct" values). For example, $X_i = (1/N)\sum_j R_{ij}$ and all $D_j^{(m,k)} = 0$. The algorithm runs "continuously", and, therefore, the choice of initial state - at the system start-up or reset - is not crucial.)

END ALGORITHM

Remark. In the case when the actual scheduling algorithm (described in Section 7.1.6) has non-zero minimum rate requirement, the terms $\frac{\partial U}{\partial X_i}(X)$ in the above virtual scheduling algorithm are everywhere replaced by $\exp(aT_i)\frac{\partial U}{\partial X_i}(X)$, where the factor $\exp(aT_i)$ is fed from the actual scheduler.

6 Sector autonomous power allocation algorithm

6.1 An illustration of why a version of MGR, but without coordination, does not work

Suppose that for some reason (standards constraints, performance constraints, etc.) inter-cell coordination which is a part of MGR is impossible or undesirable. Then, a natural question is: What if we run a version ("special case") of MGR, but exclude inter-cell coordination? Namely, suppose each sector k estimates only the values of $D_j^{(k,k)}$ (see (9)), that is, sensitivities of its utility to its "own" powers $P_j^{(k)}$; and it uses $D_j^k = D_j^{(k,k)}$ instead of (1). One might hope that such an algorithm, let us call it Single-cell Gradient (SGR), will still result in substantial performance improvement over UNIVERSAL (even if its performance is worse than that of MGR). Unfortunately, our simulation experiments showed that this is not the case: SGR typically does not produce a good fractional frequency reuse pattern, and instead has the tendency to equalize powers across sub-bands in most sectors; thus, it typically reverts to UNIVERSAL. This phenomenon can be illustrated by the following "toy" example.

Consider a system consisting of two sectors $k = 1, 2$, two sub-bands $j = 1, 2$, and the number of users being two, with user 1 served by sector 1, and user 2 served by sector 2. Assume the propogation gains are constant and "symmetric" as follows:

$$G_1^{(1)} = a > G_2^{(1)} = b > 0 \text{ and } G_2^{(2)} = a > G_1^{(2)} = b,$$

i.e. the gain a from a sector to the user it serves is strictly greater than the gain b to the user it does not serve.

It is easy to observe that at least for some values of the parameters, in particular when the ratio b/a is close to 1 and noise density N_0 is small, the UNIVERSAL scheme (allocating equal powers to sub-bands in each sector) is very sub-optimal, while the optimal allocation is for the sectors to completely "avoid each other" by using all their powers in different sub-bands. We will now show that, in the scenario we consider, such optimal behavior is *impossible under the SGR scheme.*

Namely, we will show that, under SGR, a symmetric power allocation

$$P_1^{(1)} = x > P_2^{(1)} = y \geq 0, \quad P_2^{(2)} = x > P_1^{(2)} = y,$$

where $x + y = P^*$, cannot be stable in that each sector will try to reduce the difference between powers allocated to different sub-bands to increase the rate delivered to its user. Indeed, consider sector 1 and its user 1. The total rate user 1 receives in both sub-bands is

$$\frac{W}{\log 2}\left[\log\left(1 + \frac{a(x-t)}{N_0 W + by}\right) + \log\left(1 + \frac{a(y+t)}{N_0 W + bx}\right)\right]_{t=0}.$$

The derivative on t (at $t = 0$) of the expression in the square brackets above is

$$\frac{a}{N_0 W + ay + bx} - \frac{a}{N_0 W + ax + by} > 0.$$

This means that, for the power allocation as described above, the SGR algorithm run by sector 1, will increase power allocated to sub-band 2 at the expense of power in sub-band 1; sector 2 will do the opposite. Therefore, SGR will "drive" the power allocation in the direction of the reduction of power imbalances between sub-bands in both sectors, i.e. towards the equal power allocation.

The discussion in this section thus shows that for the purposes of the rate-based utility maximization of the best-effort traffic in the system, a "selfish" behavior, ignoring the impact of a sector power changes on the neighboring sectors, is insufficient.

6.2 SA: A different algorithm without coordination

Still, the idea of having a completely distributed (with no inter-sector communication) algorithm, producing good FFR patterns and outperforming UNIVERSAL, is very attractive. We will now propose such an algorithm, and call it Sector Autonomous (SA). Although this algorithm does not explicitly maximize the sector utility itself, we believe that it is based on reasonable heuristics. We will show by simulations that its performance is good, (although, as expected, not as good as that of MGR); this algorithm may be an attractive option for applications where extra inter-cell communication is undesirable or infeasible.

The idea of SA is this. We will make each sector to selfishly solve a somewhat different, "artificial" optimization problem, which is however, (a) "highly correlated" with the original one and (b) inherently "encourages" an uneven power allocation to sub-bands (when such is beneficial).

Namely, let us "pretend" that a sector operates in the following way. (We are talking about a single sector, and will suppress sector index k.) Suppose a parameter $\bar{P}$, $P^*/J \leq \bar{P} \leq P^*$, is fixed. In each (virtual) time slot, in each sub-band j, sector either serves (transmits to) exactly one of the users i at power level $\bar{P}$ (and then the transmission rate is R_{ij}, depending on the *actually measured* SNR of user i), or does not serve any user at all (in which case the power used is 0). Now, given this setting, suppose that we employ a scheduling strategy which, over time, solves the following problem: Maximize $\sum_i U_i(X_i)$, where X_i are users' average throughputs, subject to the constraint on the total average power

$$\sum_i \bar{P}_j \leq P^*,$$

where $\bar{P}_j$ is the *average* power (per virtual slot) allocated in sub-band j. This problem is efficiently solved by a virtual scheduling algorithm described below, which runs continuosly. (The algorithm is a special case of Greedy Primal-Dual algorithm [9].) Then, the *actual* per-sub-band power levels P_j are set and adjusted to be equal to the average powers $\bar{P}_j$ (continuously produced and adjusted by the virtual scheduling).

SA ALGORITHM: VIRTUAL SCHEDULING FOR $\bar{P}_j$ CALCULATION:

The algorithm is run by each sector k independently, over a sequence of "virtual time slots." (The algorithm runs a fixed number $n_v \geq 1$ of virtual slots within each physical time slot of the system. The greater the n_v the greater the accuracy of the algorithm and its responsiveness of to changes in system state; but, the computational burden is greater as well.) The algorithm maintains the current values X_i of average user (virtual) throughputs, the current values of $\bar{P}_j$, and a variable Z. It uses a small (averaging) parameter $\beta > 0$, which is chosen in conjunction with n_v. (As a general rule, as n_v changes, the product βn_v has to be kept constant.)

In each virtual time slot, we sequentially pick each sub-band j and do the following.

IF $\max_i \frac{\partial U}{\partial X_i}(X) J R_{ij} - \beta Z \bar{P} \geq 0$,

1a. Choose user i^*,

$$i^* \in \arg\max_i \frac{\partial U}{\partial X_i}(X) R_{ij}.$$

2a. Update:

$$X_{i^*} = \beta J R_{i^* j} + (1-\beta) X_{i^*},$$
$$X_i = (1-\beta) X_i, \quad \text{for all } i \neq i^*,$$
$$\bar{P}_j = \beta \bar{P} + (1-\beta) \bar{P}_j,$$
$$Z = Z + \bar{P}.$$

ELSE

2b. Update:

$$X_i = (1-\beta) X_i, \quad \text{for all } i,$$
$$\bar{P}_j = (1-\beta) \bar{P}_j.$$

END

3. Update:

$$Z = \max\{Z - P^*/J, 0\}.$$

The initial values of the variables are, for example, as follows: $X_i = (1/N)\sum_j R_{ij}$, $\bar{P}_j = P^*/J$, $Z = 0$. (The algorithm runs "continuously", and, therefore, the choice of initial values - at the system start-up or reset - is not crucial.)

END ALGORITHM

Remark. In the case when the actual scheduling algorithm (described in Section 7.1.6) has non-zero minimum rate requirement, the terms $\frac{\partial U}{\partial X_i}(X)$ in the above virtual scheduling algorithm are everywhere replaced by $\exp(aT_i)\frac{\partial U}{\partial X_i}(X)$, where the factor $\exp(aT_i)$ is fed from the actual scheduler.

7 Simulations

7.1 System model for simulations and MGR and SA algorthms' implementation aspects.

We consider a hexagaonal grid of 19 base stations each with three sectors. The sector antennas are assumed to be oriented in a clover-leaf pattern so that the adjacent cell sectors are not facing each other directly. A wrap-around model for interference where the hexagonal arrangement is replicated by translation to create the same number of interfering cells around every one of the 19 cells is adopted.

Standard propagation parameters as listed in Table 7.1 are used to determine the received signal power level for a given transmit power level. For these parameters, with the site-to-site distance set at 2.5 Km, the cell edge SNR (signal to thermal noise ratio, when there is no interference from

Parameter	**Assumption**
Cell Layout	Hexagonal 57 sector
Inter-site distance	2.5 Km
Path Loss Model	$L = 133.6 + 35\log_{10}(d)$
Shadowing	Log Normal with 8.9 dB Std. Dev.
Penetration Loss	10 dB
Noise Bandwidth	1.25 Mhz
BS Power	40 dBm
BS Antenna Gain	15 dB
Rx Antenna Gain	0 dB
Rx Noise Figure	7 dB
Channel Model	No fading, Frequency-selective fading

Table 1: Propagation parameter values used in the simulation results

surrounding cells, assuming total available power is distributed uniformly over the entire bandwidth) turns out to be 20 dB. To demonstrate the effect of no fast fading we run the simulations with and without fast fading. When fast fading is used in the simulations, the model is representative of frequency-selective Rayleigh fading with temporal characteristics captured through Jakes fading model with vehicle speed of 20 Km/hr and carrier frequency of 2 Ghz. The frequency-selectivity is modeled by simulating independent fading across sets of coherence bands. In our simulations we consider 6 sub-bands that are divided into three sets of coherence bands each with two sub-bands.

7.1.1 Traffic model

The full buffer traffic model is used for most of the simulation results. In this case, the assumption is that all users have an infinite amount of back-logged traffic. To test the robustness of the proposed algorithms, we also consider a bursty traffic model that is representative of web browsing, with clustered packet arrivals and reading times. In our model, a large burst of data arrives instantaneously at the base station that transitions the user state to active. The user then remains in active state until all of the data is transmitted to the user. When the buffer becomes empty the user state is transitioned to inactive state. An exponentially distributed random time later, the state is transitioned to active state and another burst of data arrives. In the simulations, the size of the data burst in 9000 bits and the sojourn time in the inactive state is exponentially distributed with mean of 400 slots. The simulations are run for 10000 slots for this traffic model. Simulations using this bursty traffic model indicate that results are similar to that for the full buffer traffic model.

7.1.2 Frequency hopping

In our simulation, we consider an OFDMA system with 48 sub-carriers divided into 6 sub-bands with the same number of sub-carriers in each sub-band. Random frequency hopping is implemented from slot to slot by permuting the sub-carrier indices (within a sub-band) independently across the different sub-bands and sectors.

7.1.3 Transmit power allocation

The transmit power of each sub-band is determined by the algorithm in Section 3 for the MGR algorithm and by the algorithm in Section 6 for the SA algorithm.

In the case of MGR, the virtual scheduling algorithm described in Section 5.2 is run every slot. The number of virtual slots is set at 30. The various parameter values used in the virtual scheduling algorithm are $\beta_1 = 0.005, \beta_2 = 0.01$. The values of the rates R_{ij} and of the gain ratios $\frac{G_i^{(m)}}{G_i^{(k)}}$ used by the virtual scheduling are computed as specified in Section 7.1.5.

In the case of SA the virtual scheduling algorithm described in Section 6 is also run every slot with 30 virtual slots. The various parameter values are $\beta = 0.01, J = 6, \bar{P} = (2/3)P^*$.

7.1.4 SINR feedback

A key requirement for the algorithm is the feedback of signal-to-interference-and-noise (SINR) ratio. The mobile feeds back instantaneous SINRs achieved by received pilot signal transmitted by the sector with which it is associated for each of the sub-bands once every few slots. Pilot signals are transmitted in a (same size) sub-set of the sub-carriers in each sub-band. Pilot transmission power per sub-carrier is held fixed at nominal value P_p independent of the data power, which is adjusted for each sub-band over time (based on virtual scheduling). However, the locations of the pilot sub-carriers are randomized within each sub-band across the different sectors and hence data signal of surrounding sectors interferes with the received pilot signal. SINR is estimated by a user for each sub-band separately in every slot. Clearly, such instantaneous SINR depends on the powers allocated in other sectors in the same sub-band, as well as on the instantaneous propagation gains. More precisely, denote by $\tilde{\Gamma}_{ij}^{(k)}$ the SINR estimated from the pilot signal from sector k in sub-band j by user i. It is given by

$$\tilde{\Gamma}_{ij}^{(k)} = \frac{P_p \tilde{G}_{ij}^{(k)}}{(W/c)N_0 + (1/c)\sum_{m \neq k} P_j^{(m)} \tilde{G}_{ij}^{(m)}} \tag{10}$$

where $P_j^{(m)}$ is the current transmit power in sub-band j of sector m and $\tilde{G}_{ij}^{(m)}$ is the *instantaneous* sub-band j channel gain from sector m base station to user i.

Since there is a delay in feeding back the estimates to the transmitter and the feedback may only be sent once every few slots, it is more appropriate for the receiver to feedback a short term average SINR $\hat{\Gamma}_{ij}^{(k)}$. A reasonable way to do such averaging is to actually average the rates corresponding (by Shannon formula) to the instantaneous SINR, and then convert the average rate back to the SINR form. Namely, we do it as follows (with t being time slot):

$$\hat{r}_{ij}^{(k)}(t) = \frac{4}{5}\hat{r}_{ij}^{(k)}(t-1) + \frac{1}{5}\frac{W}{c}\log_2\left(1 + \tilde{\Gamma}_{ij}^{(k)}\right) \tag{11}$$

and then determine the $\hat{\Gamma}_{ij}^{(k)}$ that is fed back, from the equation

$$\hat{r}_{ij}^{(k)}(t) = \frac{W}{c} \log_2 \left(1 + \hat{\Gamma}_{ij}^{(k)}\right). \tag{12}$$

The short-term average SINR $\hat{\Gamma}$ is reported by a user to its base station for the purposes of channel-aware, opportunistic scheduling (as described in Section 7.1.6). Before we proceed describing how it is used by the base station, let us make two assumptions, without loss of generality. (They are just to simplify formulas - without them more general, but more cumbersome, formulas can be easily derived; these assumptions are used throughout the rest of Section 7.)
(Simp1) Power-per-subcarrier of the pilot transmission is equal to total sector power P^* divided by total number of sub-carriers, i.e. $P_p = P^*/(Jc)$.
(Simp2) We assume that additional power and sub-carriers are available for pilot signal transmission. Thus the entire power $P_j^{(k)}$ allocated by sector k in sub-band j and all c sub-carriers of the band are assumed to be used for data transmission.
Then, the estimate (via Shannon formula) of the instantaneous rate $\hat{R}_{ij}$ which *would be* available to user i in sector k if it *would be* scheduled (in a slot) in sub-band j with the power level $P_j^{(k)}$ is

$$\hat{R}_{ij} = W \log_2 \left(1 + \hat{\Gamma}_{ij}^{(k)} \frac{P_j^{(k)}}{P^*/J}\right). \tag{13}$$

7.1.5 Feedback for virtual scheduling algorithms (MGR and SA) and for the gradient computation (in MGR)

The virtual scheduling components of both MGR and SA require the estimation of the rates R_{ij}, which are the (longer-term) rates available to user i in sub-band j of sector k, given the current power-to-sub-band allocation in the system, and if the user *were to be scheduled.* The rates R_{ij} are estimated completely analogously to $\hat{R}_{ij}$, except the averaging is over much longer time interval. Namely, the average pilot SINRs $\Gamma_{ij}^{(k)}$ are computed the same way as $\hat{\Gamma}_{ij}^{(k)}$, except in (11) the coefficients are 499/500 and 1/500 instead of 4/5 and 1/5, respectively. The values of $\Gamma_{ij}^{(k)}$ are reported, *much less frequently than* $\hat{\Gamma}_{ij}^{(k)}$, by each user to its base station, which computes

$$R_{ij} = W \log_2 \left(1 + \Gamma_{ij}^{(k)} \frac{P_j^{(k)}}{P^*/J}\right). \tag{14}$$

In addition to feeding back the pilot SINRs, for the MGR operation, the user also has to feedback the ratios of the path gains from the serving sector to interfering sectors for a set of strongest interfering sectors. (These ratios are required for the gradient computation (4)-(8), carried out by the virtual scheduling algorithm in Section 5.2.) These are estimated by a user i in sector k for all sectors m as follows:

$$\frac{G_i^{(m)}}{G_i^{(k)}} = \frac{1 + \frac{1}{J}\sum_j \frac{1}{\Gamma_{ij}^{(k)}}}{1 + \frac{1}{J}\sum_j \frac{1}{\Gamma_{ij}^{(m)}}}. \tag{15}$$

The estimate (15) is justified by the fact that it is *exact* in the case when propagation gains are indeed constant in time and equal to $G_i^{(m)}$, and, additionally, we assume that the total power transmitted by each sector is equal to the maximum P^*. (The latter assumption is typically true - sectors tend to use all the available power under both MGR and SA.) Indeed, in this case it is verified directly that, for $l = k, m$,

$$1 + \frac{1}{J} \sum_j \frac{1}{\Gamma_{ij}^{(l)}} = \frac{1}{J} \sum_j \left[1 + \frac{1}{\Gamma_{ij}^{(l)}} \right] = \frac{N_0 W + \sum_{m'} G_i^{(m')} P^*}{G_i^{(l)} P^*},$$

which yields (15).

For the case when fast fading is present, expression (15) is a good approximation.

The path gain ratio can be fed back to the base station relatively infrequently since, as discussed earlier, the values of $\Gamma_{ij}^{(k)}$ change relatively slowly.

7.1.6 Actual scheduling of transmissions

With the powers $P_j^{(k)}$ for each sub-band in sector k dynamically determined by the appropriate algorithm (either MGR or SA), actual scheduling is implemented independently by each sector k. The scheduling algorithm is such that it maximizes the utility $U^{(k)}$ of the sector, *given the current power-to-sub-band allocation in the system.* (This obviously means that the total system utility under the current power allocation is maximized as well.) We use the utility function $U^{(k)} = \sum_i \log(\bar{X}_i)$, where the summation is over users i served by sector k, and $\bar{X}_i$ are users' *actual* average throughputs. (These are generally *not* the average throughputs X_i used in the virtual scheduling algorithms.) This utility function results in the well known *proportional fair* scheduling (cf. [2]). In each time slot, the potential instantaneous rates, $\hat{R}_{ij}$, are determined as explained in Section 7.1.4 by the serving sector for all its users in all sub-bands. Then, in this slot, in each sub-band j a user with the maximum value of the metric $\hat{R}_{ij}/\bar{X}_i$ is scheduled (and assigned the entire sub-band). The average rates $\bar{X}_i$ are updated only upon successful transmission of the packets of corresponding users.

In our simulations we, in fact, use a generalization of the proportional fair scheduling algorithm (see [2]) which allows us to introduce minimum rate requirements of the form $\bar{X}_i \geq b$ for some constant $b \geq 0$. The generalized algorithm maintains a *token counter* variable T_i for each user i, and uses a more general scheduling metric of the form $\exp(aT_i)\hat{R}_{ij}/\bar{X}_i$, where $a > 0$ is a parameter. The factor $\exp(aT_i)$, maintained by the actual scheduler, is also fed to and used by the virtual scheduling algorithms (see remarks in Sections 5.2 and 6.2.)

7.1.7 Rate computation for actually scheduled transmissions

For all scheduled users in a given slot, the number of successfully received bits during the slot needs to be computed. Assume, for example, that a user i being served by base station k is assigned two sub-bands j_1, j_2. Then the actual transmission rate for this user is computed as

$$\hat{R}_i = \hat{R}_{i,j_1} + \hat{R}_{i,j_2} \tag{16}$$

where $\hat{R}_{i,j}$ are the instantaneous rates defined earlier. From this transmission rate and the number of OFDM symbols per slot, 8 in the simulations, the actual number of bits achievable is computed.

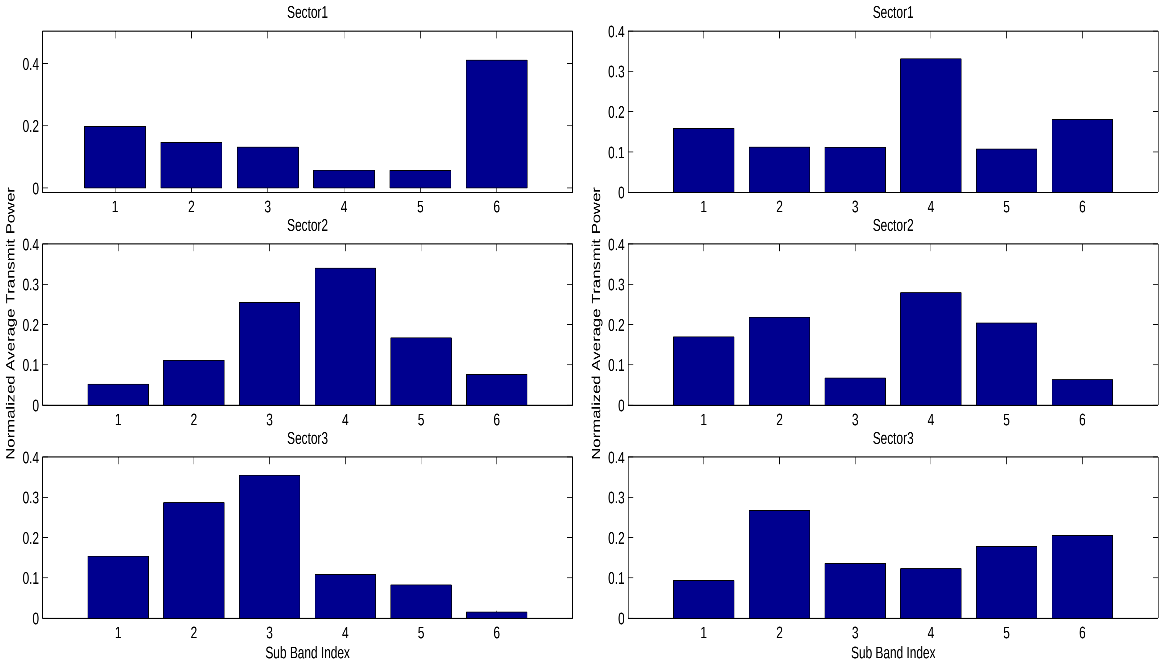

Figure 1: Normalized average power with full buffer traffic model for MGR (left) and SA (right)

7.1.8 Incremental redundancy

The simulation also captures incremental redundancy, commonly used in cellular systems. Incremental redundancy is required because the transmitter cannot exactly predict the achievable transmission rate. The size of the transmitted information packet is determined by the transmission rate corresponding to the SINR fed back. When transmitted, this packet can take more than one slot to be successfully received. In our simulations the bandwidth resources assigned to an user selected by the scheduler for transmission are also assigned to the same user in subsequent slots of the interlace until the packet is successfully transmitted. Packets are deemed to be successfully received when the total channel capacity in bits over all the transmissions exceeds the information packet size.

7.2 Results and discussion

To illustrate that both the MGR and SA algorithms create soft fractional frequency reuse patterns automatically, we show the average transmit power in each of the six sub-bands in Figure 1 for both the algorithms. The average powers are obtained by averaging the transmit powers set by the algorithms over the entire duration of the simulation. The powers are shown for three out of the 57 sectors that are roughly facing each other. Average powers normalized by the total sector power are plotted in these figures. The results are for the case of the fading channel, uniform user distribution, and full buffer traffic model. The corresponding results for the bursty traffic model are shown in Figure 2. As can be seen from the figures there is a clear separation of powers with both algorithms for both traffic models. It is also clear that the reuse pattern achieved is a soft reuse in the sense that all sub-bands are used in all sectors but with different power levels.

The slot by slot transmit power levels set by the algorithms are shown in Figures 3 and 4 for the

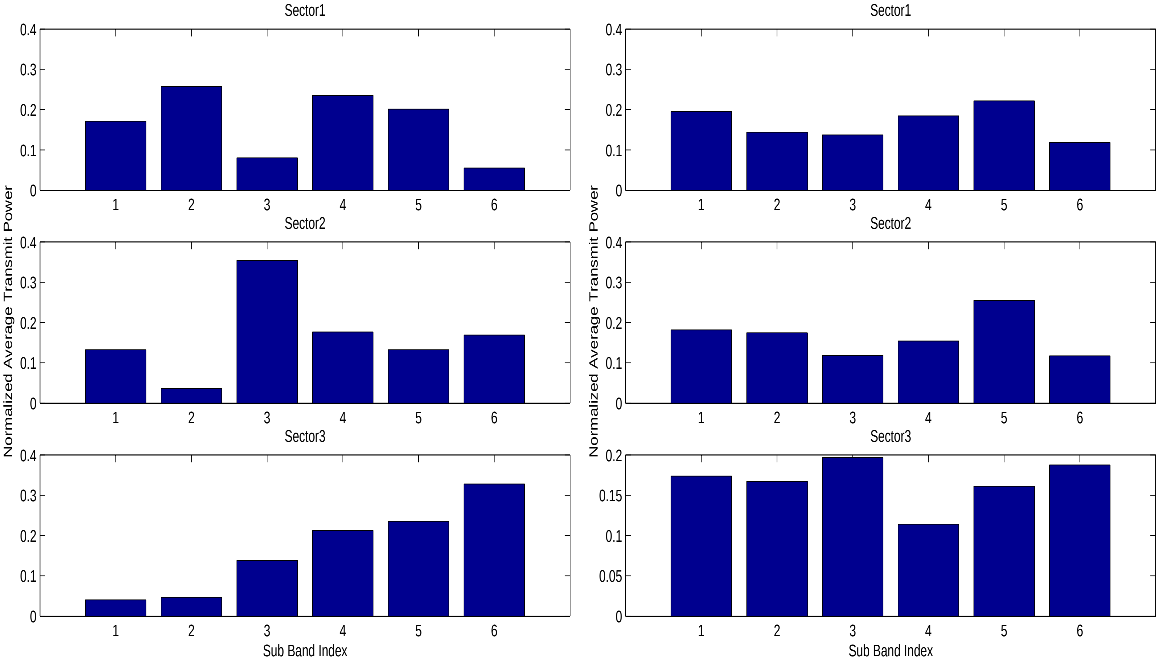

Figure 2: Normalized average power with bursty traffic model for MGR (left) and SA (right)

full buffer and bursty traffic models, respectively. In contrast to the full buffer case, in the bursty case the transmit power levels are being adjusted significantly in a continuous fashion to the time varying active user distributions. In the bursty traffic model, only a small number of users are active (i.e. non empty buffers) in each cell at any given time and thus the optimal fractional frequency reuse pattern will change as users switch between active and in-active states. The figure clearly shows that the algorithms are trying to adapt the reuse patterns by adjusting the power levels.

Simulation results comparing the the performance of the 3 different algorithms, namely MGR, SA, and UNIVERSAL, are presented in the form of geometric average of user throughputs or average sector throughput versus the 5-percentile throughput. We use the geometric average as one of the performance metrics, because maximizing it is the algorithms' objective (recall that the utility function is the sum of log-throughputs), and it is easier to "relate to" than the sum of log-throughputs metric. In particular, percentage improvements are much more meaningful in the geometric average metric than the sum of log-throughputs metric. The 5-percentile throughput is a measure of the cell edge throughput. Different points on the tradeoff curve between sector throughput and edge throughput are obtained using the same scheduling algorithm but with different values of the minimum rate parameter of the scheduling algorithm. Two different scenarios, one where the users are distributed uniformly in each of the 57 sectors and another where the user distribution within each sector is non-uniform were simulated. Non-uniform user distribution is simulated as follows. User distribution for each sector is randomly chosen to be "center" or "edge" distribution. In the case of center distribution, the users are uniformly distributed in a region close to the base station and are guaranteed to have geometry (average SINR without fast fading) of greater than 6 dB. In the edge distribution, users are distributed uniformly in sector edges and have geometry below 0 dB.

Figure 5 shows the results for the case of uniform distribution of users. As can be seen from the figure, when the normalized average sector utility is maintained at 12.8, the 5-percentile throughput

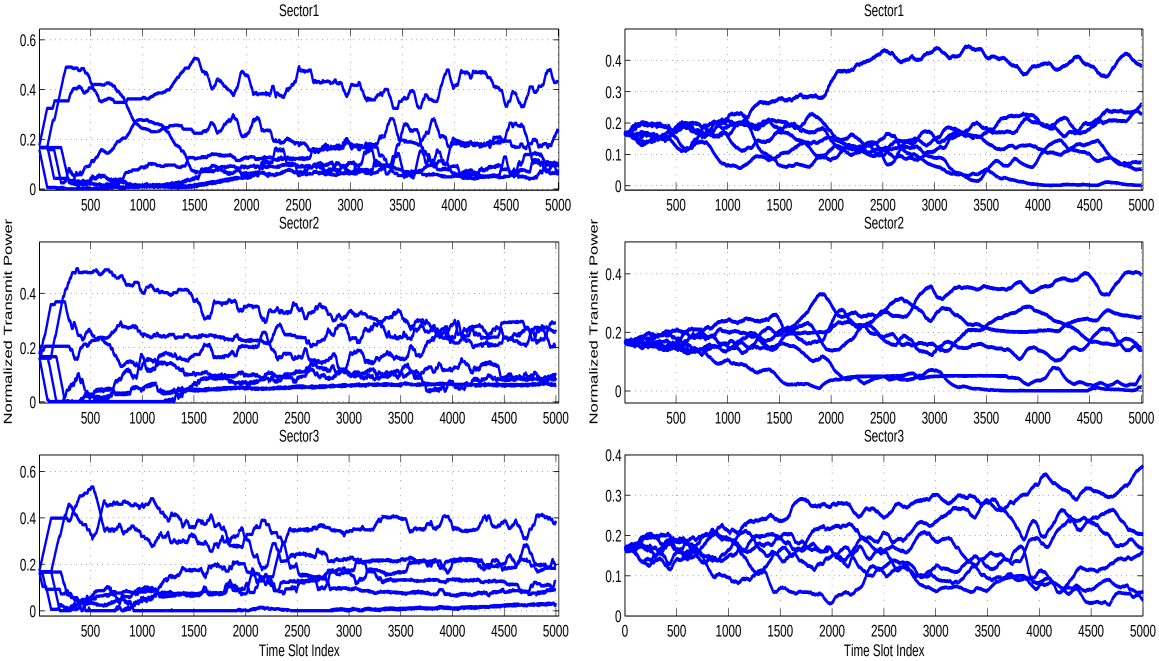

Figure 3: Time series of normalized transmit powers on the different sub-bands with full buffer traffic model for MGR (left) and SA (right)

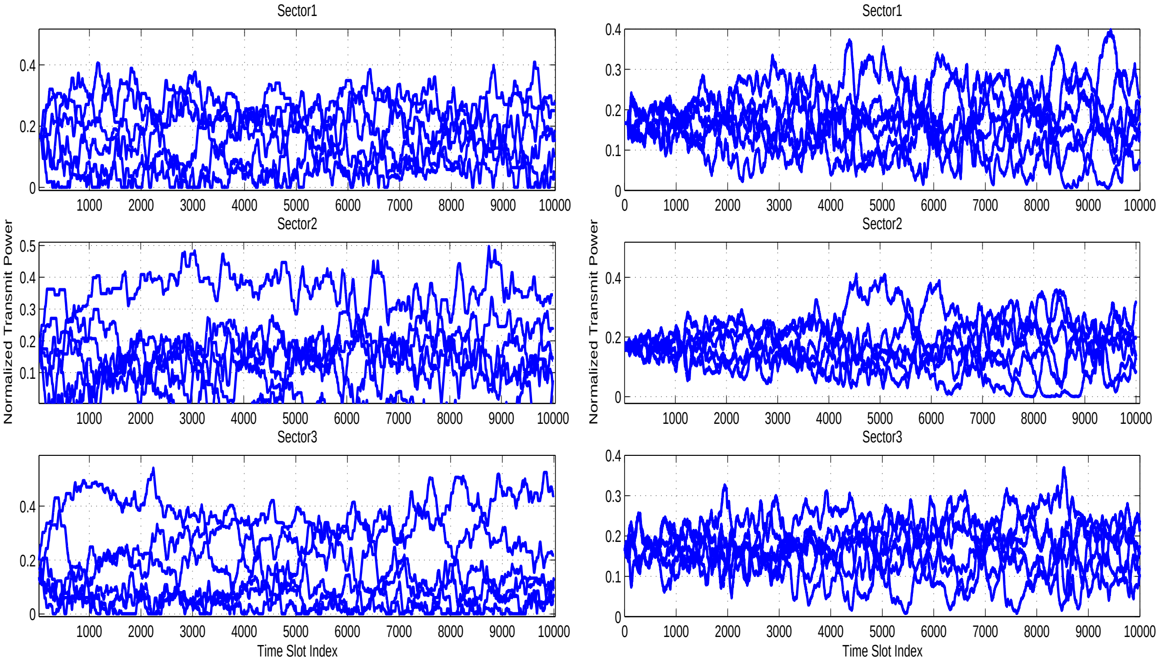

Figure 4: Time series of normalized transmit powers on the different sub-bands with bursty traffic model for MGR (left) and SA (right)

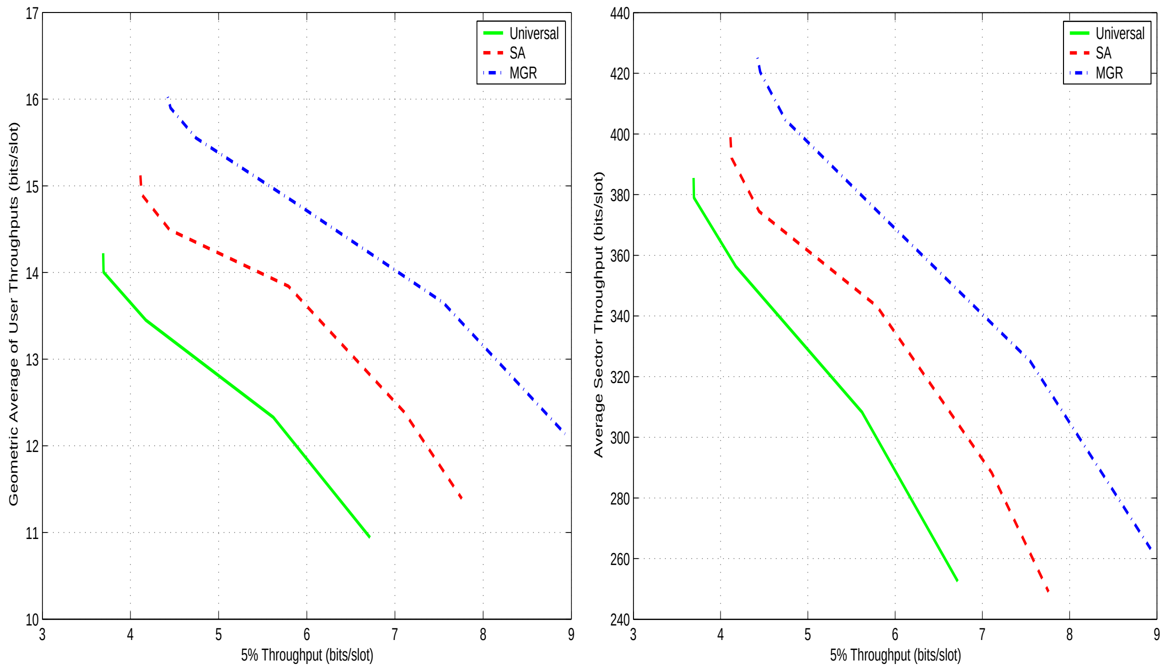

Figure 5: Geometric average of user throughputs Vs. 5-% edge throughput (left) and Average sector throughput Vs. 5-% edge throughput (right)

can be increased by 34% using the SA algorithm and by 66% using the MGR algorithm relative to UNIVERSAL. Also observe that, as expected, the gain in sector utility and sector throughput increase with increasing edge throughput. With the 5-percentile throughput at 5 bits/slot, the gain in sector throughput using MGR is about 21%. For 5-percentile throughput of 6 bits/slot the corresponding gain is about 27%.

Figure 6 shows the results for the case of non-uniform distribution of users. The choice between the "center" and "edge" user distributions for each sector is kept the same across all algorithms and for all points along the curves. As can be seen from the figure, when the sector utility is maintained at 17.3, the 5-percentile throughput can be increased by 25% using the SA algorithm and by 55% using the MGR algorithm relative to the UNIVERSAL. It should be noted that using the proportional fair with minimum rate scheduling algorithm, increasing the minimum rate parameter further does not result in an increase in the edge throughput for the UNIVERSAL. Thus it is possible to achieve a much higher cell edge throughput using the MGR algorithm compared to UNIVERSAL.

Figure 7 shows the results for the case of uniform distribution of users, but with no channel fast fading. For average sector throughput at 280 bits/slot, the 5-percentile throughput is increased by 150% and 200% for the SA and MGR algorithms, respectively. Comparing these figures to those in Figure 5 shows that the gains of both algorithms are larger in the case without fast fading than with fading. As explained in Section 8, this is because fast fading introduces asynchronous fluctuations in SINR of different users that are exploited automatically by the basic channel aware proportional fair with minimum rate scheduling algorithm, which achieves a certain degree of “automatic” interference avoidance resulting in a smaller gap between the SA, MGR and the UNIVERSAL. As seen from the figure the maximum cell edge throughput achievable by the UNIVERSAL is substantially smaller compared to those of the SA and MGR algorithms. For the same 5-percentile throughput of about 3

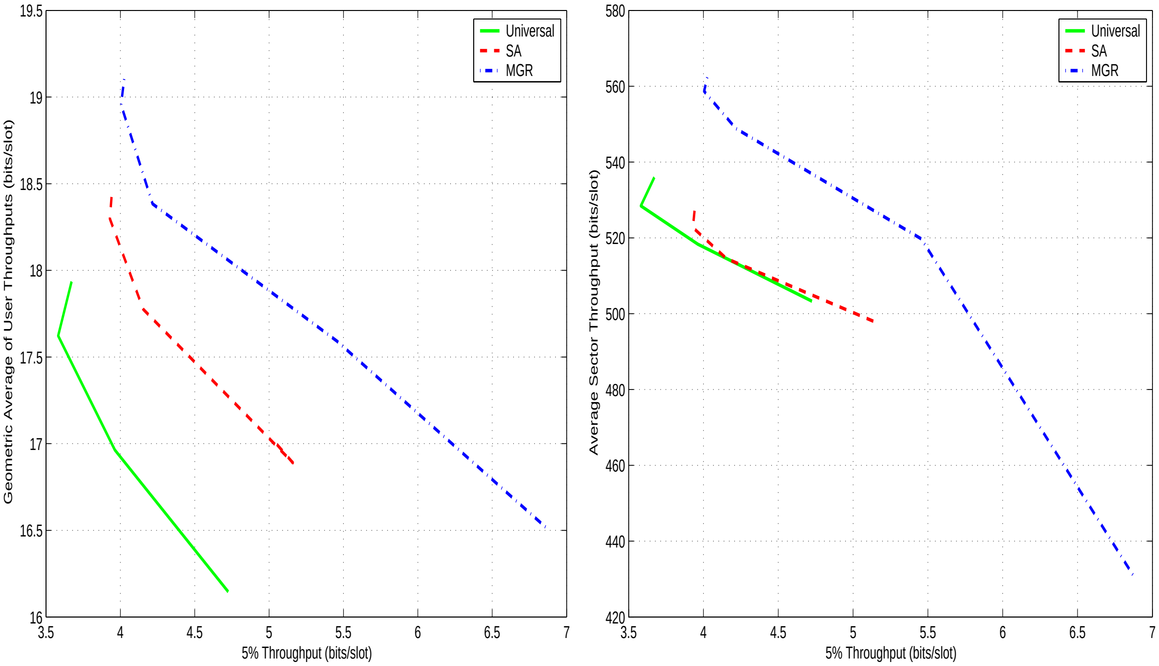

Figure 6: Geometric average of user throughputs Vs. 5-% edge throughput (left) and Average sector throughput Vs. 5-% edge throughput (right)

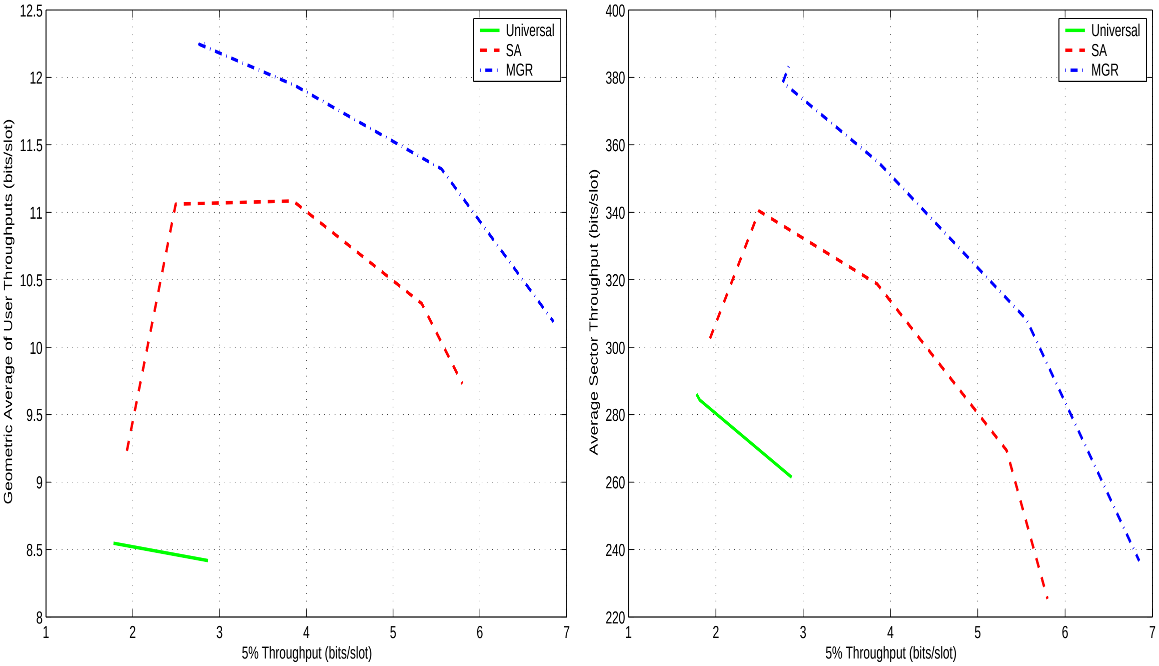

Figure 7: Geometric average of user throughputs Vs. 5-% edge throughput (left) and Average sector throughput Vs. 5-% edge throughput (right)

bits/slot, the sector utility of MGR is 49% better than that of UNIVERSAL while SA is 35% better.

8 Fast fading case: degree of "automatic" interference avoidance

In this section we demonstrate using a simple example that, when fast fading is present and sectors employ opportunistic, channel-aware scheduling, some degree of inter-sector interference avoidance is present "automatically", even if sectors use the straightforward UNIVERSAL approach, i.e. simply divide power equally among the sub-bands and do not coordinate with each other. Therefore, the benefit of FFR in the case of fast fading is reduced or - in extreme cases - may be even non-existent.

Despite the fact that our example is very extreme (or, perhaps, – owning to that), it nicely illustrates a real phenomenon we observed in simulations. Suppose, the system consists of two sectors $k = 1, 2$, the number of users served by each sector is "large", and there are two sub-bands $j = 1, 2$. Assume that the instantaneous propagation gain $\tilde{G}_{ij}^{(k)}$, for any sector k, any sub-band j and any user i (which may or may not be served by k) is random and can take only two values, $g > 0$ and 0 with positive probabilities (say, 1/2 and 1/2); the gains $G_{ij}^{(k)}$ for different (k, i, j) are independent. The utility of each sector is $U^{(k)} = \sum_i \log \bar{X}_i$ (with summation over users i served by k); and there is no minimum user rate constraints.

Obviously, an upper bound on the system utility is obtained if we assume that each user in each sector and each sub-band, at all times has the propagation gain g from its serving sector and gain 0 from the non-serving sector. In this case, the optimal starategy for each sector is to always use half the total power in each sub-band, $P_1^{(k)} = P_2^{(k)} = P^*/2$, because this maximizes the total rate achieved in a time slot. (Which users are served in a slot is not even important, as long as on average users are scheduled "equally frequently".)

Now, let us get back to our scenario with random propagation gains and assume that sectors do not coordinate, each sector k always divides power equally among sub-bands, that is $P_1^{(k)} = P_2^{(k)} = P^*/2$, and it employs the opportunistic scheduling algorithm specified in Section 7.1.6. Since the number of users per-sector is large, in every time slot, there will "always" (with very high probability) be users in each sub-band with gain g from k and gain 0 from the non-serving sector: only such users can and will be scheduled in the slot, and thus the total rate in each slot will be the maximum possible. By symmetry and by the form of gradient scheduling algorithm, all average user throughputs $\bar{X}_i$ in sector k will be approximately equal. This means that the system utility "attains" the upper bound described earlier.

In words, the above phenomenon can be described as follows. If fast fading is present and opportunistic scheduling is employed, just by virtue of picking users with better instantaneous channel condition, each sector tends to serve users that "at the moment" experience lower interference from neighboring sectors. Hence, a degree of interference avoidance is achieved "automatically." We believe this is an important phenomenon. Its impact can be small or large, depending on a real system deployment scenario: in systems with predominantly static users (like femto-cells) the impact is small, and FFR techniques are most beneficial. We also note that the automatic interference avoidance can give a large benefit only if fast and accurate channel state feedback is available.

9 Conclusions and future work

Two algorithms for automatically creating efficient, soft, fractional frequency reuse patterns that dynamically adapt to traffic distribution were presented. Both algorithms adjust the transmit power levels of each sub-band to maximize utility with available information. MGR relies on exchanging of partial gradient information between neighboring sectors that provide information regarding sensitivity to transmit power changes in each of the sub-bands while SA does not require any exchange of information between sectors. For the computation of the required gradient in the MGR algorithm, a virtual algorithm was proposed and rigourously justified, one of the key fundamental contributions of the paper that could be interesting in its own right.

Detailed simulation results were presented to demonstrate the automatic formation of FFR patterns and the performance benefits that could be achieved in typical macro-cellular settings. We observed that these algorithms improve cell edge user throughput (quantified by the 5-percentile throughput) substantially while maintaining the sector throughput at the same level as the traditional algorithm. Modest sector throughput improvements are achieved while comparing the performances for the same cell edge throughput.

One of the interesting observations from the simulation results was that the performance gains of the proposed algorithms were superior when there was no small scale channel fading simulated. This observation is in line with the intuition that a channel aware opportunistic scheduler already exploits fluctuations in the interference levels introduced by the channel. We also provided a simple example with two cells where we could analytically demonstrate the benefits of fading for interference avoidance.

Several avenues for future work are possible. We have focused on the forward link (base station to the users) in this paper. It is of interest to derive algorithms for the reverse link as well. Because of inherent asymmetries in the interference patterns forward link solutions may not carry over as is for the reverse link. We focused on the best effort traffic in this paper, and the latency sensitive traffic was treated in our earlier work [10]. An overall scheme that combines these separate algorithms into a complete solution is another area for research. Finally, we do not have a handle on the fundamental limits on the performance gains from fractional frequency reuse. Obtaining useful bounds to benchmark the performance of the proposed algorithms is of great interest.

10 Appendix

We are within the setting and notation of Section 4.

Theorem 10.1 *Suppose, $R > 0$ (meaning all $R_{ij} > 0$) is such that ϕ corresponding to $X(R)$ (i.e. satisfying (2) with X replaced by $X(R)$) is unique. Then, for any (i, j), (3) holds.*

Before with proceed with the proof of Theorem 10.1, we need to state some basic facts, that are either well known (see [8] and reference therein) or easy to observe. (We always assume $R > 0$, in which case $V(R)$ contains at least one vector with all positive components.) The dependence $X(R)$ is continuous. For either type of utility function U, for any $R > 0$, $[\nabla U](X(R)) =$

$(\partial U/\partial X_1, \ldots, \partial U/\partial X_1)(X(R))$ has all finite (and then strictly positive) components. (This is because even if U is type-2, we must have $X_i(R) > 0$ for those i for which $U_i(0) = -\infty$.) The gradient vector $[\nabla U](X(R))$ is an outer normal vector to region $V(R)$ at point $X(R)$; consequently,

$$X(R) \in \arg\max_{X \in V(R)} [\nabla U](X(R)) \cdot X; \tag{17}$$

consequently, any ϕ corresponding to $X(R)$ is such that

$$\sum_i \phi_{ij} = 1, \quad \forall j, \tag{18}$$

and

$$\phi_{ij} > 0 \ \text{ implies } \ i \in \arg\max_k \frac{\partial U}{\partial X_k}(X(R)) R_{kj}. \tag{19}$$

For a given j, we will call pair (i, j) a *basic activity* if the condition on the right in (19) is satisfied. Then, the facts, collected below in Lemma 10.1 for future reference, easily follow.

Lemma 10.1 *Let $R > 0$. Then,*
(i) For any ϕ corresponding to $X(R)$ and for a given j, only basic activities can have $\phi_{ij} > 0$, and $\frac{\partial U}{\partial X_i}(X(R))R_{ij}$ has the same, maximum value for all of them; for all non-basic activities, the latter product is strictly less then the maximum.
(ii) For all $\tilde{R}$ sufficiently close to R, the corresponding set of basic activities is a subset of that for R. ("A small change of R cannot turn a non-basic activity into basic.")

Proof of Theorem 10.1. The uniqueness of ϕ for R, and the continuity of $X(R)$, imply the following property. Let $\tilde{\phi}$ be a fixed "set of fractions" corresponding to $\tilde{R}$. (Such $\tilde{\phi}$ exist - we pick any of them in case of non-uniqueness.) Then,

$$\tilde{\phi} \to \phi \text{ as } \tilde{R} \to R. \tag{20}$$

Let us compute the right partial derivative on R_{ij}. Consider $\tilde{R}$ which is equal to R, except $\tilde{R}_{ij} = R_{ij} + \delta$, with small $\delta \geq 0$. First of all,

$$\frac{\partial}{\partial R_{ij}} U(X(R)) = \frac{d}{d\delta} U(X(\tilde{R}_{ij}))|_{\delta=0} \geq \frac{\partial U}{\partial X_i}(X(R))\phi_{ij},$$

because we can always choose a non-optimal value of X instead of $X(\tilde{R})$ by keeping ϕ constant for any δ. Thus, we have the lower bound matching RHS in (3). To prove the upper bound, consider the linearization of U at point $X(R)$, that is the function

$$\bar{U}(\tilde{X}) = U(X(R)) + [\nabla U](X(R)) \cdot (\tilde{X} - X(R));$$

obviously, $\bar{U}(\tilde{X}) \geq U(\tilde{X})$. Then, using notations $\Delta\phi = \tilde{\phi} - \phi$ and $u_i = \frac{\partial U}{\partial X_i}(X(R))$, we can write for all sufficiently small δ:

$$U(X(\tilde{R})) - U(X(R)) \leq \bar{U}(X(\tilde{R})) - U(X(R)) =$$

$$= \sum_{k \neq i} u_k R_{kj} \Delta\phi_{kj} + u_i[(R_{ij} + \delta)(\phi_{ij} + \Delta\phi_{ij}) - R_{ij}\phi_{ij}] =$$

$$= \sum_k u_k R_{kj} \Delta\phi_{kj} + u_i(\phi_{ij} + \Delta\phi_{ij})\delta = u_i(\phi_{ij} + \Delta\phi_{ij})\delta.$$

The last equality is because $\sum_k u_k R_{kj} \Delta\phi_{kj} = 0$, which follows from the facts that $\sum_k \Delta\phi_{kj} = 0$, $\Delta\phi_{kj} = 0$ for all non-basic activities for R (by Lemma 10.1(ii)), and for all basic activities (i,k) the values of $u_k R_{kj}$ are same (Lemma 10.1(i)). Thus, we have

$$U(X(\tilde{R})) - U(X(R)) \leq u_i(\phi_{ij} + \Delta\phi_{ij})\delta.$$

Since $\Delta\phi_{ij} \to 0$ as $\delta \to 0$ (by (20)), we obtain the desired upper bound

$$\frac{\partial}{\partial R_{ij}} U(X(R)) = \frac{d}{d\delta} U(X(\tilde{R}_{ij}))|_{\delta=0} \leq \frac{\partial U}{\partial X_i}(X(R))\phi_{ij}.$$

The proof that the left partial derivative on R_{ij} is also equal to the RHS in (3) is analogous, except we immediately get the *upper* bound, and then use linearization to obtain the *lower* bound. ∎

We now turn to the issue of uniqueness and continuity of ϕ.

Theorem 10.2 *Let $R > 0$. Consider a bi-partite graph with nodes being user indecies $i \in \mathcal{I}$ and sub-band indecies $j \in \mathcal{J}$, and with edges (i,j) corresponding to basic activities. Suppose this "basic activity" graph is a tree (perhaps disconnected). Then,*
(i) ϕ corresponding to $X(R)$ is unique.
(ii) Moreover, for any $\tilde{R}$ sufficiently close to R, the corresponding basic activity graph is still a tree and then $\tilde{\phi}$ corresponding to $\tilde{R}$ is unique as well.
(iii) Moreover, the dependence of $\tilde{\phi}$ on $\tilde{R}$ (via $X(\tilde{R})$) in a neighborhood of R is continuous.
(iv) Consequently, $U(X(R))$ is continuously differentiable in a neighborhood of R, with partial derivatives given by (3).

Proof. Statement (i) is immediate, because all components ϕ_{ij} are uniquely determined by sequentially "eliminating" the leafs of the tree, one at a time. Then, (ii) follows from Lemma 10.1(ii), and (iii) follows from the fact that uniqueness of $\tilde{\phi}$ for $X(\tilde{R})$ implies its continuity on $\tilde{R}$ at that point. Finally, (iv) follows from (ii), (iii) and Theorem 10.1. ∎

We now show that the cases when basic activity graph is *not* a tree are "non-typical."

Theorem 10.3 *Let $R > 0$. Suppose the basic activity graph is* not *a tree, that is there exists a cycle*

$$i(1) \to j(1) \to i(2) \to j(2) \to \ldots \to i(m) \to j(m) \to i(m+1) = i(1),$$

where each $i(\ell)$ is a user index, each $j(\ell)$ is a sub-band index, and all edges are different basic activities. Then,

$$\prod_{\ell=1}^{m} \frac{R_{i(\ell),j(\ell)}}{R_{i(\ell+1),j(\ell)}} = 1. \tag{21}$$

Proof. Let us use notation $u_i = \frac{\partial U}{\partial X_i}(X(R))$. By Lemma 10.1(i), $u_{i(1)} R_{i(1),j(1)} = u_{i(2)} R_{i(2),j(1)}$, and thus

$$\frac{u_{i(2)}}{u_{i(1)}} = \frac{R_{i(1),j(1)}}{R_{i(2),j(1)}}.$$

Writing analogous relations for $u_{i(3)}/u_{i(2)}$ and so on, and multiplying them together, we obtain (21). ∎

Finally, we can state the following Theorem 10.4, which says that "typically," at "almost any" R, the function $U(X(R))$ is continuously differentiable.

Theorem 10.4 *Suppose the value of R is chosen randomly and uniformly from $[0, B]^{NJ}$. Then, with probability 1, this R satisfies conditions of Theorem 10.2. Consequently, $U(X(R))$ is continuously differentiable in a neighborhood of R, with partial derivatives given by (3).*

Proof. With probability 1, $R > 0$ and it does *not* satisfy any of the rational relations (21) for any possible cycle in the bi-partite graph of all activities (i, j). Then, by Theorem 10.3, the basic activity graph must be a tree, and thus R satisfies conditions of Theorem 10.2, whose application completes the proof. ∎

References

[1] E. Altman, K. Avrachenkov, and A. Garnaev, "Closed form solutions for water-filling problem in optimization and game frameworks," in *Proceeding of INFOCOM'2008*, Phoenix, April 14-18, 2008.

[2] M. Andrews, L. Qian, A. L. Stolyar, "Optimal Utility Based Multi-User Throughput Allocation subject to Throughput Constraints," in *Proceeding of INFOCOM'2005*, Miami, March 13-17, 2005.

[3] T. Bonald, S. C. Borst, and A. Proutiere, "Inter-cell scheduling in wireless data networks," in *Proceedings of European Wireless Conference*, 2005.

[4] S. T. Chung, S. J. Kim, J. Lee, and J.M. Cioffi, "A game theoretic approach to power allocation in frequency-selective Gaussian interference channels," in *Proceedings of the IEEE International Symposium on Information Theory*, pp 316-316, July 2003.

[5] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proceedings of INFOCOM*, 2003.

[6] S. Das and H. Viswanathan, "Interference mitigation through intelligent scheduling," in *Proceedings of the Asilomar Conference on Signals and Systems*, Asilomar, CA, November 2006.

[7] A. Gjendemsjo, D. Gesbert, G. E. Oien, and S. G. Kiani, "Optimal power allocation and scheduling for two-cell capacity maximization," in *Proceedings of the IEEE RAWNET (WiOpt)*, April 2006.

[8] A.L. Stolyar, "On the Asymptotic Optimality of the Gradient Scheduling Algorithm for Multi-User Throughput Allocation," *Operations Research*, 2005, Vol. 53, No.1, pp. 12-25.

[9] A. L. Stolyar, "Maximizing Queueing Network Utility subject to Stability: Greedy Primal-Dual Algorithm," *Queueing Systems*, 2005, Vol.50, No.4, pp.401-457.

[10] A. L. Stolyar, H. Viswanathan, “Self-organizing Dynamic Fractional Frequency Reuse in OFDMA Systems,” in *Proceedings of INFOCOM’2008*, Phoenix, April 14-18, 2008.

[11] Third Generation Partnership Project 2, ”Ultra Mobile Broadband Technical Specifications,”, http://www.3gpp2.org, March 2007

[12] Third Generation Partnership Project, Radio Access Network Work Group 1 Contributions, http://www.3gpp.org, September 2005