# Balancing Supply and Demand of Bandwidth in Wireless Cellular Networks: Utility Maximization Over Powers and Rates

Mung Chiang and Jason Bell
Electrical Engineering Department, Princeton University, NJ 08544

***Abstract*— In wireless cellular networks and wireless local area networks, nonlinear network utility maximization need to be conducted over both user rates and transmit powers. For each of the three cases considered in this paper, we present an algorithm that converges to the jointly optimal pair of rate vector and power vector.**

**For the simple case when data rates are not limited by interferences, for example in single-cell downlink transmissions, Algorithm 1 we propose is an iterative bidding mechanism between the base station and mobile users, where knowledge about channel conditions and individual user utility functions is only needed locally at each user but *not* needed at the base station.**

**In the case when data rates are limited by interferences, the utility maximization problem is complicated both by nonlinear coupling between powers and rates, and by interference among powers. Through centralized iterative steps, Algorithm 2 we propose converges to a joint and global optimum over the solution space of rates and powers.**

**We then consider end-to-end transmissions in cellular networks, which traverse both wireless fading channels and many hops of wired links shared by other traffic. There is a tradeoff between attaining air-interface capacity in the wireless hop and controlling congestion in the wired backbone wide area network. We formulate this end-to-end resource allocation problem in such hybrid networks, and present Algorithm 3 to obtain the Pareto optimal tradeoff between attaining wireless multi-access fading channel capacity and maximizing global network utility.**

**Keywords:** Convex optimization, Lagrange duality, Power control, Rate allocation, Transport Control Protocol, Utility maximization, Wireless local area networks, Wireless cellular networks.

## I. Introduction

Communication system performance is sometimes best measured not by a weighted sum of attainable rates, but by some nonlinear utility functions of rates. Each user has a utility function that is assumed to be continuously differentiable, concave, and increasing, and the sum of all users' utility functions is called the network utility. Network utility maximization under linear flow constraints is an important class of problems in wired networks and has been extensively studied.

In wireless networks, rate feasibility is often affected by channel conditions and adaptive resource allocations like power control. Power control mechanisms determine the bandwidth 'supply': how much throughput can be attained on each wireless link, while rate allocation algorithms regulate the bandwidth 'demand': how much throughput should be given to each user. Total network utility must now be maximized over the joint solution space of powers and rates. The nonlinear dependency of rates on channel conditions and powers, as well as possible interference among the transmit powers, are the main challenges of solving utility maximization problems in wireless networks. Utility maximization over powers and rates in wireless *ad hoc* networks with multihop wireless transmissions has been studied in the context of joint congestion control and power control, *e.g.*, in [4]. This paper investigates three different cases in wireless *cellular* networks.

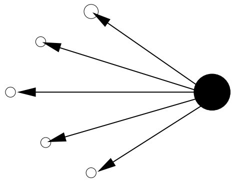

Fig. 1. Single-cell downlink case without interference.

In section II, we first consider the simpler case where rates are not limited by interference, for example in wireless downlinks in a single cell as depicted in Figure 1. Unlike [17] where the rate is assumed to be a linear function of the received power, here we assume that rate is proportional to the logarithm of the received power. We present a pricing algorithm through an iterative bidding mechanism that solves the problem even when the base station has no knowledge about each individual user's channel condition and utility function.

Then in section III, we turn to the general case of uplink/downlink transmission in a multi-cell CDMA system as depicted in Figure 2. Intended transmissions, either downlink or uplink, are shown in solid lines, and some of the unintended interferences are shown in dashed lines. In addition to the nonlinear dependency of rates on transmit powers and channel conditions, due to signal interference, the attainable rate on each link now becomes a global function of all the transmit powers. Foschini and Miljanic [6] propose an iterative power control that finds a set of transmit powers to achieve some *fixed* target rates. Here we propose a complementary rate control

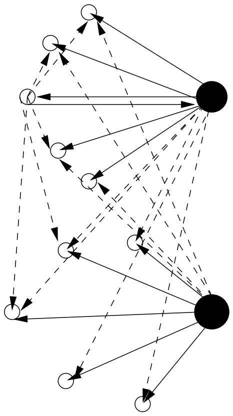

Fig. 2. Multi-cell uplink/downlink case with interference.

algorithm that couples with Foschini and Miljanic power control to maximize network utility over the *joint* solution space of powers and rates.

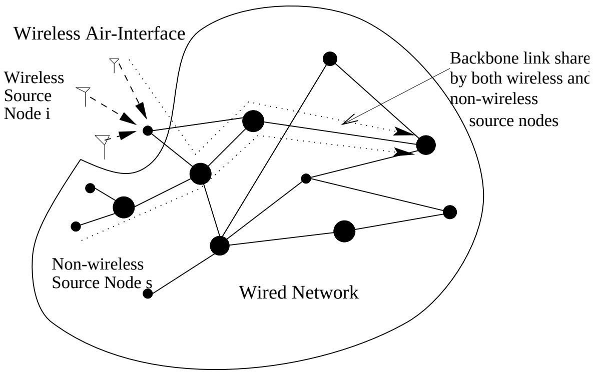

Fig. 3. End-to-end connections in a hybrid network.

Contrary to its name, wireless cellular networks in fact consist primarily of *wired* links. In section IV, we consider a hybrid wireless-wired network depicted in Figure 3, and upload transmission using uplinks in the wireless air-interface. There are two distinct parts in the network: a wireless multiple-access channel (MAC) and a wired mesh backbone. This models a cellular wireless network, where the wireless MAC is from mobile users to a base station, and the wired backbone is from base stations through mobile switching centers and ATM switches to either the PSTN or an IP wide area network. It also models a wireless local area network (LAN), where the wireless MAC is from laptops to an access point, and the wired backbone is from access points through an Ethernet LAN to an IP wide area network. A unique feature of resource allocation in such hybrid networks is that the effect of channel variations at the wireless hop is coupled with the effect of congestion on various wired links in the backbone network. Section IV presents Algorithm 3 as an end-to-end resource allocation for such hybrid networks to trace out the Pareto optimal tradeoff between attaining local MAC capacity and maximizing global network utility.

## II. Single-Cell Downlink Case

### *A. Background*

Consider a general multihop network, where some nodes are sources of transmission, and sequences of connected links form routes. We use $r, s$ and $l$ as the indexing variables for routes, sources and links, respectively. Let $x_s$ be the transmission rate of source $s$, $y_r$ be the total flow along route $r$, and $c_l$ as the capacity in terms of supportable data rate on link $l$. There are two $0-1$ incidence matrices: $\mathbf{H} = \{H_{sr}\}$ and $\mathbf{A} = \{A_{lr}\}$. Entry $H_{sr} = 1$ iff route $r$ serves source $s$, and entry $A_{lr} = 1$ iff link $l$ is on route $r$.

The standard problem of network utility maximization for elastic traffic source (*e.g.*, in [10], [5]) is to maximize the sum of individual sources' utilities represented through differentiable, increasing, and concave functions $U_s(x_s)$, subject to flow conservation constraint $\mathbf{Hy} = \mathbf{x}$ and link capacity constraint $\mathbf{Ay} \preceq \mathbf{c}$:

$$\begin{array}{ll} \text{maximize} & \sum_s U_s(x_s) \\ \text{subject to} & \mathbf{Hy} = \mathbf{x}, \\ & \mathbf{Ay} \preceq \mathbf{c}, \\ & \mathbf{x}, \mathbf{y} \succeq 0 \end{array} \qquad (1)$$

where the variables are $\mathbf{x}$ and $\mathbf{y}$ ($\succeq$ denotes component-wise inequality).

Kelly et al. [10], [11] showed that problem (11) can be decomposed into two sets of problems. First are subproblems *SOURCE*$_s$, one for each source $s$, to be solved locally over $m_s$:

$$\text{maximize}_{m_s \geq 0} \left[ U_s \left( \frac{m_s}{\mu_s} \right) - m_s \right]. \qquad (2)$$

Second is subproblem *NETWORK* to be solved for the entire network over $\mathbf{x}$ and $\mathbf{y}$:

$$\begin{array}{ll} \text{maximize} & \sum_s m_s \log x_s \\ \text{subject to} & \mathbf{Hy} = \mathbf{x}, \\ & \mathbf{Ay} \preceq \mathbf{c}, \\ & \mathbf{x}, \mathbf{y} \succeq 0. \end{array} \qquad (3)$$

To state that utility maximization (1) can be decomposed into the above subproblems is equivalent to the following statement [10], [11]: there exist $\{m_s\}, \{\mu_s\}, \{x_s\}, \{y_r\}$ such that $m_s = \mu_s x_s$, $\forall s$, $\{m_s\}$ solve the *SOURCE*$_s$ problem, and $\{x_s\}, \{y_r\}$ solve the *NETWORK* and utility maximization (1).

One of the important advantages gained through the above decomposition is that the *NETWORK* problem can be distributively solved and does not require the knowledge of each individual user's utility function $U_s$.

The simple linear flow constraints in (1) can be extended for wireless cellular systems to take into account the nonlinear dependencies of link rates on channel conditions and adaptive resource allocations.

- In [17], the flow constraints are made to depend on local channel conditions and resources (time or power) linearly, which is an appropriate model for TDMA systems or CDMA systems in the wide-band regime. It is shown that as channel quality varies across the users, the base station should charge different users different prices based in part on their channel qualities. The optimal pricing requires the knowledge about each user's utility function at the base station. A suboptimal scheme that does not require this knowledge is shown to be asymptotically optimal.
- In this section, the flow constraints are made to depend on local channel conditions and transmit powers logarithmically, which is an appropriate model for CDMA systems in the high SIR regime. We will show that the optimal algorithm can be interpreted as an iterative bidding mechanism that does not require the knowledge of each user's utility function or channel condition at the base station. It turns out that this is possible in part because of the logarithmic dependency of rates on powers and channel conditions.

Note that, although not treated in this paper, the network utility function in general does not have to be a function of user rates, or a concave function (*e.g.*, [7], [9], [25]), or separable into each individual user's utility function.

### B. Problem formulation

Consider the single cell downlink case in Figure 1 with $M$ logical users, and assume CDMA transmission with orthogonal codes. The base station has a total transmit power of $\bar{P}$ that is divided into $P_i \geq 0$ for transmitting to user $i$ such that $\sum_{i=1}^{M} P_i \leq \bar{P}$. The channel gain is denoted as $G_i$ for channel $i$, and the received power is $G_i P_i$. The attainable rate is modelled as $L \log(1 + \mathrm{SNR})$ where $L$ is a constant. Assuming high SNR, user rate is upper bounded by $R_i \leq L \log\left(\frac{G_i P_i}{N_i}\right)$ where $N_i$ is the noise. Without loss of generality, normalize over $L$ and let $g_i = \frac{G_i}{N_i}$, we have the nonlinear constraint $R_i \leq \log(g_i P_i)$.

Therefore, we need to solve the following problem of network utility maximization over both transmit powers $\mathbf{P}$ and user rates $\mathbf{R}$:

$$\begin{array}{ll} \text{maximize} & \sum_i U_i(R_i) \\ \text{subject to} & R_i \leq \log(g_i P_i), \;\; \forall i, \\ & \sum_i P_i \leq \bar{P}, \\ & \mathbf{P} \succeq 0 \end{array} \tag{4}$$

where the variables are $\mathbf{P}$ and $\mathbf{R}$, and $\mathbf{g} \succeq 0$ is a constant vector.

Note that the underlying model for (4) is not the information-theoretic optimal one for multi-user fading channels [13], [22], [23], which will be discussed in section IV. The focus of this section is to show how to maximize the nonlinear objective under nonlinear constraints as in (4), through an iterative pricing algorithm that does not require the knowledge of $\{g_i\}$ and $\{U_i\}$ at the base station.

### C. Algorithm

We will show that the following Algorithm 1 solves (4). The algorithm can be interpreted as an iterative pricing mechanism. Based on only local information: channel condition $g_i$ and its own utility function $U_i$, each user $i$ in turn calculates a 'bid' $\lambda_i$ to submit to the base station. The base station simply updates the sum of the 'bids' without knowing $g_i$ or $U_i$. After the (guaranteed) convergence of the iterative bidding process, base station allocates power $P_i$ proportional to the (normalized) equilibrium bid $\lambda_i^*$.

**Algorithm 1**

```
Given accuracy tolerance $\epsilon > 0$. Counter $k = 0$.
Base station initiate a vector $\boldsymbol{\lambda}^0$.
repeat
Base station passes $I = \frac{\sum_j \lambda_j^{(k)}}{\bar{P}}$ to user 1.
for $i = 1 : M$
User $i$ computes $\lambda_i^{(k+1)}$ such that $\frac{g_i \lambda_i^{(k+1)}}{\exp(U_i'^{-1}(\lambda_i^{(k+1)}))} = I$ and passes $\lambda_i^{(k+1)}$ to base station.
Base station passes $I = \frac{\sum_{j=1}^{i} \lambda_j^{(k+1)} + \sum_{j=i+1}^{M} \lambda_j^{(k)}}{\bar{P}}$ to user $i+1$.
end
$k = k + 1$.
until $|\lambda_i^{(k+1)} - \lambda_i^{(k)}| \leq \epsilon, \;\; \forall i$.
$P_i^* = \frac{\lambda_i^*}{\sum_j \lambda_j^*} \bar{P}$.
```

It will be shown that the above power control leads to the following rate allocation: $R_i^* = U_i'^{-1}(\lambda_i^*), \;\; \forall i$.

As an example of Algorithm 1, if the utility functions are weighted log: $U_i(R_i) = \beta_i \log R_i, \; \beta_i \geq 0, \; \forall i$, the equation to be solved for $\lambda_i$ by user $i$ is $\frac{g_i \lambda_i}{\exp(\frac{\beta_i}{\lambda_i})} = I$.

Different pricing mechanisms have been used for wireless power control, *e.g.*, in [17], [20], [23]. The novelty of Algorithm 1 is in using a (provably convergent) iteration of bidding process to maximize *nonlinear* utility under *nonlinear* constraints *without* the knowledge of $\{g_i\}$ and $\{U_i\}$ at the base station. This extends the results by Kelly et al. [10], [11] for utility maximization in wired networks without global knowledge about individual utility functions $\{U_i\}$.

Several propositions can be proved on the properties of Algorithm 1. We focus on the most important one in this paper:

*Theorem 1:* Algorithm 1 converges to a globally optimal $(\mathbf{P}^*, \mathbf{R}^*)$ of utility maximization (4).

### D. Proof of Theorem 1

Since $U_i$ are increasing functions, it is obvious that at optimum, the first constraint in (4) must be tight. Since log is an increasing function, the second constraint in (4) must also be tight at optimum. Therefore, without loss of generality, we can replace these inequality constraints in (4) with equalities, and write the Lagrangian as

$$L(\mathbf{P}, \mathbf{R}, \boldsymbol{\lambda}, \gamma, \boldsymbol{\sigma}) = \sum_i U_i(R_i) + \sum_i \lambda_i (\log(g_i P_i) - R_i)$$

$$+\gamma(\bar{P} - \textstyle\sum_i P_i) + \sum_i \sigma_i P_i$$

where $\boldsymbol{\sigma} \succeq 0, \boldsymbol{\lambda}$, and $\gamma$ are the Lagrange multipliers associated with the three constraints.

Maximizing $L$ over $\mathbf{R}$, we obtain:

$$\frac{\partial L}{\partial R_i} = U_i'(R_i) - \lambda_i = 0, \ \ \forall i$$

which implies the following optimality condition:

$$R_i^* = V_i(\lambda_i), \ \ \forall i \tag{5}$$

where $V_i = U_i'^{-1}$ is defined as the inverse of the derivative of utility function.

Maximizing $L$ over $\mathbf{P}$, we obtain:

$$\frac{\partial L}{\partial P_i} = \frac{\lambda_i}{P_i} - \gamma + \sigma_i, \ \ \forall i,$$

which implies the following optimality condition:

$$P_i = \frac{\lambda_i}{\gamma - \sigma_i}, \ \ \forall i. \tag{6}$$

Substituting (5) and (6) into $L$, we obtain the following Lagrange dual function:

$$g(\boldsymbol{\lambda}, \gamma, \boldsymbol{\sigma}) = \textstyle\sum_i U_i(V_i(\lambda_i)) + \sum_i \lambda_i \left(\log\left(\frac{g_i\lambda_i}{\gamma-\sigma_i}\right) - V_i(\lambda_i)\right)$$
$$+\gamma\bar{P} - \gamma \textstyle\sum_i \frac{\lambda_i}{\gamma-\sigma_i} + \sum_i \frac{\sigma_i\lambda_i}{\gamma-\sigma_i}.$$

In (4), the objective is maximizing a concave function, the first constraint $\log(g_iP_i) - R_i \geq 0$ is concave in $(\mathbf{P}, \mathbf{R})$, the second and third constraints are affine, and there obviously exists an interior point in the feasible set. Therefore, duality gap is zero and solving (4) is equivalent to solving its Lagrange dual problem: minimizing the Lagrange dual function over the Lagrange multipliers:

$$\begin{array}{ll} \text{minimize} & g(\boldsymbol{\lambda}, \gamma, \boldsymbol{\sigma}) \\ \text{subject to} & \boldsymbol{\sigma} \succeq 0. \end{array} \tag{7}$$

Notice that because $\boldsymbol{\lambda}$ and $\gamma$ correspond to equality constraints in the primal problem (4), they are unconstrained in the Lagrange dual problem (7).

Obviously, the last two terms of $g(\boldsymbol{\lambda}, \gamma, \boldsymbol{\sigma})$ add up to become $-\sum_i \lambda_i$. This leaves $\log\left(\frac{g_i\lambda_i}{\gamma-\sigma_i}\right)$ as the only terms in $g(\boldsymbol{\lambda}, \gamma, \boldsymbol{\sigma})$ involving $\boldsymbol{\sigma}$. Since each of such terms is an increasing function of $\sigma_i$, to minimize $g(\boldsymbol{\lambda}, \gamma, \boldsymbol{\sigma})$ over $\boldsymbol{\sigma} \succeq 0$, we should simply let $\boldsymbol{\sigma} = 0$. Thus the Lagrange dual problem becomes an unstrained optimization of minimizing

$$g(\boldsymbol{\lambda}, \gamma) = \textstyle\sum_i U_i(V_i(\lambda_i)) + \sum_i \lambda_i \left(\log\left(\frac{g_i\lambda_i}{\gamma}\right) - V_i(\lambda_i)\right)$$
$$+\gamma\bar{P} - \textstyle\sum_i \lambda_i.$$

over $\boldsymbol{\lambda}$ and $\gamma$. We first minimize $g(\boldsymbol{\lambda}, \gamma)$ over $\gamma$:

$$\frac{\partial g}{\partial \gamma} = -\frac{\sum_j \lambda_j}{\gamma} + \bar{P} = 0,$$

which implies that at optimum,

$$\gamma^* = \frac{\sum_j \lambda_j}{\bar{P}}. \tag{8}$$

Substituting (8) into $g(\boldsymbol{\lambda}, \gamma)$ and simplifying the expression, we have:

$$g(\boldsymbol{\lambda}) = \textstyle\sum_i U_i(V_i(\lambda_i)) + \sum_i \lambda_i \left(\log(g_i\lambda_i) - V_i(\lambda_i)\right)$$
$$+(\log \bar{P})(\textstyle\sum_i \lambda_i) - (\sum_i \lambda_i)\log(\sum_i \lambda_i),$$

which we must minimize over $\boldsymbol{\lambda}$. Taking derivative and simplifying the expression, we obtain:

$$\tfrac{\partial g}{\partial \lambda_i} = U_i'(V_i(\lambda_i))V_i'(\lambda_i) - \lambda_i V_i'(\lambda_i) + \log(g_i\lambda_i)$$
$$-V_i(\lambda_i) + \log\bar{P} - \log(\textstyle\sum_j \lambda_j).$$

Since $U_i'$ and $V_i$ are inverse functions, $U_i'(V_i(\lambda_i)) = \lambda_i$ and the first two terms cancel. The optimality condition $\frac{\partial g}{\partial \lambda_i} = 0$ is reduced to:

$$\log(g_i\lambda_i) - V_i(\lambda_i) + \log\bar{P} - \log\left(\sum_j \lambda_j\right) = 0.$$

An equivalent and more illuminating expression is:

$$\frac{g_i\lambda_i}{\exp(V_i(\lambda_i))} = \frac{\sum_j \lambda_j}{\bar{P}}, \ \ \forall i. \tag{9}$$

Notice that as desired, the right hand side does *not* depend on the local channel condition $g_i$ or utility function $U_i$ at each user $i$, and the left hand side contains *only* information $(g_i, U_i)$ and variable $\lambda_i$ local to user $i$. Furthermore, substituting (8) into (6) and using $\boldsymbol{\sigma} = 0$, we have

$$P_i^* = \frac{\lambda_i^*}{\sum_j \lambda_j^*}\bar{P}.$$

The rest of proof follows readily from the convexity properties and known results on the successive realization of the nonlinear Gauss-Siedel algorithm [2].

### E. Numerical example

Results from an illustrative numerical example is summarized in this subsection. We consider a cellular system with one base station and six downlink users, each with a different channel gain and a weighted logarithmic utility function with a different weight. Total transmit power at the base station is 6 units and all bids are initialized to be 1. Figures 4, 5 and 6 show the bids from the six users through the first eight iterations according to Algorithm 1. It can be seen that the bids quickly converge in about 4 rounds, and Figure 7 shows the resulted power and rate allocation after convergence for the six users.

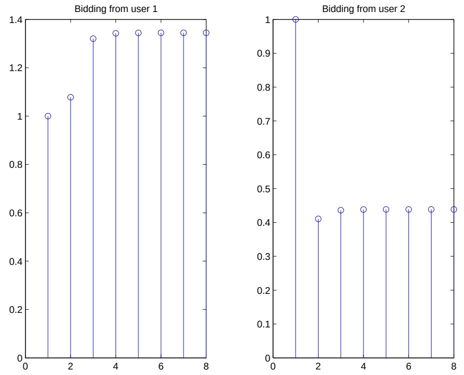

Fig. 4. Algorithm 1 example: Bidding from users 1 and 2.

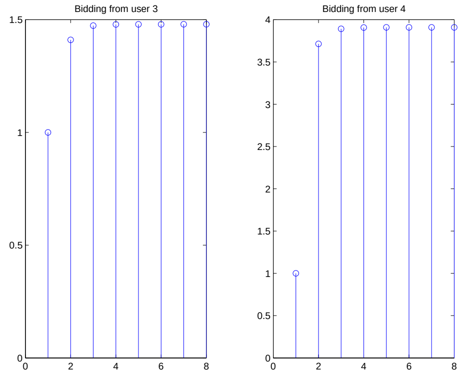

Fig. 5. Algorithm 1 example: Bidding from users 3 and 4.

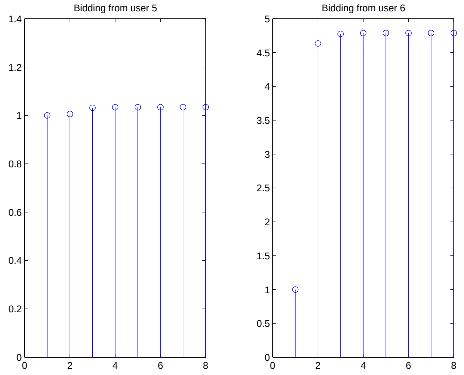

Fig. 6. Algorithm 1 example: Bidding from users 5 and 6.

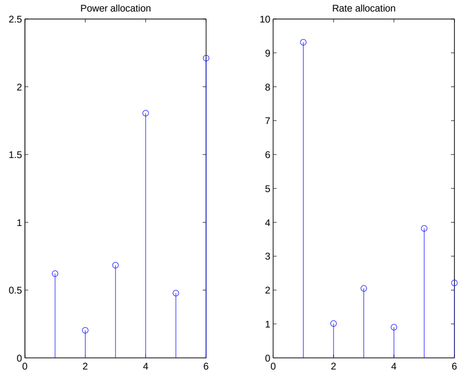

Fig. 7. Algorithm 1 example: Power and rate allocation for six users.

## III. Multi-Cell General Case

### A. Problem formulation

In this section, we consider the general case of multi-cell up or downlink transmissions with interference in Figure 2. Signal to Interference Ratio (SIR) for the $i^{th}$ logical link is defined as

$$\text{SIR}_i(\mathbf{P}) = \frac{P_i G_{ii}}{\sum_{j \neq i}^{N} P_j G_{ij} + N_i} \tag{10}$$

where $G_{ij}$ is the path loss from the transmitter on logical link $j$ to the receiver on logical link $i$, taking into account propagation loss and normalization factors, and $G_{ii}$ is the path gain for the intended transmission on logical link $i$, taking into account propagation loss and other factors such as spreading gain and the effect of beamforming. For a large class of modulations, the attainable data rate can be written as

$$R_i = \frac{1}{T} \log(1 + K\text{SIR}_i)$$

where $T$ is the symbol time and $K$ is a constant depending on modulation type and desired bit error probability. Due to high spreading gain, $K\text{SIR}_i$ is usually much larger than 1 for medium to high SIR environments, and we make an approximation to write $R_i = \frac{1}{T}\log(K\text{SIR}_i)$. Throughput of most CDMA cellular systems are interference limited, where noise $N_i$ is much smaller than the total interference and SIR is approximated as $\text{SIR}_i = \frac{P_i G_{ii}}{\sum_{j \neq i}^{N} P_j G_{ij}}$. For notational simplicity and without loss of generality, we take $T$ to be 1 time unit and absorb $K$ into the SIR formula (10). We now have the functional dependence of link data rate on SIR, which is in turn a nonlinear, global function of transmit power vector $\mathbf{P}$: $R_i(\mathbf{P}) = \log \text{SIR}_i(\mathbf{P})$.

A user demanding a certain data rate is requesting that the SIR of her link be high enough to sustain the desired rate. This request, however, is limited by competing rate demands from other users. This interference-limited nature of wireless CDMA system is captured by defining the feasible rate-power region $\mathcal{RP}$ as the set of pairs of power vectors and the associated feasible rate vectors:

$$\mathcal{RP} = \{(\mathbf{R}, \mathbf{P}) \in \mathbf{R}_+^{2n} | \mathbf{R} \preceq \log \text{SIR}(\mathbf{P})\}.$$

There can be an infinite number of $(\mathbf{R}, \mathbf{P})$ pairs in the feasible rate-power region. As a general design problem, we would like to pick the one that maximizes the network utility $U(\mathbf{R}) = \sum_i U_i(R_i)$. Therefore, we need to solve the following network utility maximization problem constrained in the feasible rate-power region:

$$\begin{array}{ll} \text{maximize} & \sum_i U_i(R_i) \\ \text{subject to} & (\mathbf{R}, \mathbf{P}) \in \mathcal{RP}, \end{array} \tag{11}$$

where the optimization variables are both rate allocation vector $\mathbf{R}$ and power allocation vector $\mathbf{P}$. Nonlinear relationship between $\mathbf{R}$ and $\mathbf{P}$, as well as interference among $\mathbf{P}$, make this problem difficult to solve.

### *B. Background*

Given a rate vector $\mathbf{R}$, we define a diagonal matrix $\mathbf{D}$ as a function of $\mathbf{R}$: $\mathbf{D}(\mathbf{R}) = \textbf{diag}\left(\frac{e^{R_i}}{G_{ii}}\right)$. Given a gain matrix $\mathbf{G}$, we construct a matrix $\tilde{\mathbf{G}}$: $\tilde{G}_{ij} = G_{ij}, i \neq j$ and $\tilde{G}_{ii} = 0$.

Foschini and Miljanic [6] proposed a simple and distributive power control algorithm to achieve a set of rate requirements. The given rate demands $\mathbf{R}^{\text{target}}$ is equivalent to a set of SIR requirements $\text{SIR}^{\text{target}}$, and at each iteration, each transmitter $i$ adjusts its power so that the resulted SIR would equal $\text{SIR}_i^{\text{target}}$ if all other competing users kept their transmit powers constant. This power update can be written as [6]:

$$\mathbf{P}^{k+1} = \mathbf{D}\tilde{\mathbf{G}}\mathbf{P}^{k}. \tag{12}$$

By Perron Frobenius theory of positive matrix [8], it is known [6] that, whenever the SIR requirement is feasible (*i.e.*, required rate vector inside $\mathcal{RP}$), the iterative power update in (12) will converge to a Pareto optimal power vector that achieves the desired SIR.

### *C. Algorithm*

In this section, we present an algorithm to solve the utility maximization problem (11). Note that, unlike Algorithm 1, knowledge of $\{U_i\}$ is needed in the following centralized computation.

**Algorithm 2**

**Input**: Gain matrix $\mathbf{G}$ of the cellular network and utility functions $\{U_i\}$.

**Output**: Optimal pair of rate-power vectors $(\mathbf{R}^*, \mathbf{P}^*)$.

**Algorithm**:

```
Given an initial rate vector $\mathbf{R}^0$, accuracy tolerance $\epsilon$, and step size $\alpha > 0$. Counter $k = 0$.
Compute the largest modulus eigenvalue $\lambda$ and the associated eigenvectors $\mathbf{p}, \mathbf{q}$ of $\mathbf{D}(\mathbf{R}^0)\tilde{\mathbf{G}}$.
Compute $\delta\mathbf{R}$, where $\delta R_i = \left(\frac{1}{p_i q_i}\right)\left(\frac{U_i'}{\sum_j U_j'}\right) - 1, i = 1, 2, \ldots, N$.
while $\|\delta\mathbf{R}\| > \epsilon$
Compute $\tilde{\mathbf{R}}^k = \mathbf{R}^k + \alpha\delta\mathbf{R}$.
Compute $\mathbf{R}^{k+1} = \tilde{\mathbf{R}}^k - (\log\lambda(\mathbf{D}(\tilde{\mathbf{R}}^k)\tilde{\mathbf{G}}))\mathbf{1}^T$.
Use power control in [6] to update power: $\mathbf{P}^{k+1} = \mathbf{D}(\mathbf{R}^{k+1})\tilde{\mathbf{G}}\mathbf{P}^k$.
Compute the largest modulus eigenvalue $\lambda$ and the associated eigenvectors $\mathbf{p}, \mathbf{q}$ of $\mathbf{D}(\mathbf{R}^{k+1})\tilde{\mathbf{G}}$.
Compute $\delta\mathbf{R}$, where $\delta R_i = \left(\frac{1}{p_i q_i}\right)\left(\frac{U_i'}{\sum_j U_j'}\right) - 1$.
$k = k + 1$.
end
```

We prove the following main theorem for the above algorithm:

*Theorem 2:* Algorithm 2 converges to the a globally optimal $(\mathbf{R}^*, \mathbf{P}^*)$ of utility maximization (11).

### *D. Proof of Theorem 2*

Using a change of variables $\tilde{P}_i = \log P_i$, we have

$$\begin{aligned} R_i \le \log \text{SIR}_i \quad &\Leftrightarrow \quad \frac{1}{\text{SIR}_i} \le e^{-R_i} \\ &\Leftrightarrow \quad G_{ii}^{-1} e^{-\tilde{P}_i} \sum_{j \neq i} G_{ij} e^{\tilde{P}_j} \le e^{-R_i} \\ &\Leftrightarrow \quad \log \sum_{j \neq i} e^{\tilde{P}_j - \tilde{P}_i + R_i + \log G_{ij} - \log G_{ii}} \le 0. \end{aligned}$$

By second derivative test, it can be verified that $\log\sum e^{f(x)}$ is convex in $x$ for all affine $f$. Therefore, $\log\sum_{j\neq i} e^{\tilde{P}_j - \tilde{P}_i + R_i + \log G_{ij} - \log G_{ii}}$ is convex in $(\tilde{\mathbf{P}}, \mathbf{R})$, and its sublevel set $\mathcal{RP}_i = \{(\mathbf{P}, \mathbf{R}) \in \mathbf{R}_+^{2n} | R_i \le \log \text{SIR}_i(\mathbf{P})\}$ is a convex set. Since the intersection of convex sets is also convex, $\mathcal{RP}$ is a convex set in $(\tilde{\mathbf{P}}, \mathbf{R})$. Since logarithmic mapping is injective, we can recover $\mathbf{P}$ from $\tilde{\mathbf{P}}$. Because the objective function in (11) is concave in $\mathbf{R}$ and the constraint set can be turned into convex in $(\mathbf{R}, \mathbf{P})$, it is a convex optimization problem in $(\mathbf{R}, \mathbf{P})$, and a local maximum is also a global maximum.

Since $R_i \le \log \text{SIR}_i(\mathbf{P})$ is equivalent to $P_i \ge \frac{e^{R_i}}{G_{ii}} \sum_{j \neq i} G_{ij} P_j$, we can rewrite (11) as

$$\begin{array}{ll} \text{maximize} & \sum_i U_i(R_i) \\ \text{subject to} & \mathbf{D}(\mathbf{R})\tilde{\mathbf{G}}\mathbf{P} \preceq \mathbf{P}, \\ & \mathbf{R} \succeq 0, \quad \mathbf{P} \succeq 0. \end{array} \tag{13}$$

Consider the joint rate-power control problem (13). If the variables $\mathbf{R}$ are fixed, the problem reduces to a feasibility problem of finding a power allocation $\mathbf{P}$ such that the constraints $D(\mathbf{R})\tilde{\mathbf{G}}\mathbf{P} \le \mathbf{P}$ are satisfied. This feasibility problem may not have a solution, but if solutions exist, one can be found by the iterative power control algorithm (12).

We now decompose (13) into two parts: a power control part that uses the algorithm in [6] to update power $\mathbf{P}^{k+1} = \mathbf{D}(\mathbf{R})\tilde{\mathbf{G}}\mathbf{P}^k$, and a rate allocation part to be solved in the rest of this subsection. We will show how to update the target rate vector toward an optimum, which also drives the power vector toward a joint optimum.

First recall [21], [8] that the Perron Frobenius eigenvalue $\lambda(\mathbf{A})$ is the largest modulus eigenvalue of an element-wise positive matrix $\mathbf{A}$, and the associated right eigenvector $\mathbf{p}(\mathbf{A})$ and left eigenvector $\mathbf{q}(\mathbf{A})$ are called Perron Frobenius eigenvectors. It is a standard fact from Perron Frobenius theory [21], [8] that for a positive matrix $\mathbf{A}$ with Perron Frobenius eigenvalue $\lambda$, there is an $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} \preceq \mathbf{x}$ if and only if $\lambda \le 1$. Therefore, the rate allocation subproblem is now reformulated as the following optimization over $\mathbf{R}$:

$$\begin{array}{ll} \text{maximize} & \sum_i U_i(R_i) \\ \text{subject to} & \lambda(\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}}) \le 1, \\ & \mathbf{R} \succeq 0. \end{array} \tag{14}$$

The inequality constraints $\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}}\mathbf{P} \preceq \mathbf{P}$ in (13) will be met with equality at optimality, because otherwise $\mathbf{R}$ can be increased without violating the constraints and, by monotonicity of $U(\mathbf{R})$, produce a larger objective value for (13). In order for $\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}}\mathbf{P} = \mathbf{P}$ to hold, there must be an eigenvalue of $\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}}$ that is at least 1. Since Perron Frobenius eigenvalue $\lambda$ is the largest modulus eigenvalue of $(\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}})$, we must have $\lambda \ge 1$. Earlier arguments also show $\lambda \le 1$. Thus the constraint in (14) can be written as $\lambda = 1$:

$$\begin{array}{ll} \text{maximize} & \sum_i U_i(R_i) \\ \text{subject to} & \lambda(\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}}) = 1, \\ & \mathbf{R} \succeq 0. \end{array} \tag{15}$$

We denote by $\mathcal{P}$ (P for Pareto) the set of rates $\mathbf{R}$ such that $\lambda(\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}}) = 1$. Geometrically, $\mathcal{P}$ represents Pareto optimal surface of the feasible rate region $\mathcal{R}$ under a given power allocation $\mathbf{P}$. It is the boundary of the feasible rate region because any rate vector $\mathbf{R}$ outside this surface is obviously not achievable, and any $\mathbf{R}$ inside or on it can be achieved by some power allocation $\mathbf{P}$. It is Pareto optimal because any two points $\mathbf{R}_1, \mathbf{R}_2$ on $\mathcal{P}$ cannot dominate each other, if $R_{1,i} > R_{2,i}$ for some $i$, there must be an $j$ such that $R_{2,j} > R_{1,j}$. The global maximizer $\mathbf{R}^*$ of (11) must be a point on $\mathcal{P}$.

By KKT optimality condition of equality constrained optimization, solving (15) is equivalent to optimizing the Lagrangian $U(\mathbf{R}) - \rho\lambda(\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}})$ where $\rho$ is the Lagrange multiplier, *i.e.*, solving the following nonlinear system of equations for $\mathbf{R}$:

$$\nabla U(\mathbf{R}) = \rho\nabla\lambda(\mathbf{R}).$$

Given the network utility function $U(\mathbf{R}) = \sum_i U_i(R_i)$, its gradient $\nabla U(\mathbf{R})$ can be readily computed. We also need to find $\nabla\lambda(\mathbf{R})$. We start by writing the definitions of right and left Perron Frobenious eigenvectors $\mathbf{p}, \mathbf{q}$ of $\mathbf{D}\tilde{\mathbf{G}}$, normalized to have inner product 1:

$$\begin{aligned} \mathbf{D}\tilde{\mathbf{G}}\mathbf{p} &= \lambda\mathbf{p} \\ \mathbf{q}^T\mathbf{D}\tilde{\mathbf{G}} &= \lambda\mathbf{q}^T \\ \mathbf{q}^T\mathbf{p} &= 1. \end{aligned}$$

Now differentiating both sides of the right eigenvector equation $\mathbf{D}\tilde{\mathbf{G}}\mathbf{p} = \lambda\mathbf{p}$ with respect to $\mathbf{R}$, we obtain

$$(\nabla\mathbf{D}\tilde{\mathbf{G}})\mathbf{p} + \mathbf{D}\tilde{\mathbf{G}}\nabla\mathbf{p} = \lambda\nabla\mathbf{p} + (\nabla\lambda)\mathbf{p}.$$

Multiplying both sides by $\mathbf{q}^T$, and using the left eigenvector equation and the normalization equation, we have

$$\begin{aligned} \mathbf{q}^T(\nabla\mathbf{D}\tilde{\mathbf{G}})\mathbf{p} + \mathbf{q}^T\mathbf{D}\tilde{\mathbf{G}}\nabla\mathbf{p} &= \mathbf{q}^T\lambda\nabla\mathbf{p} + \mathbf{q}^T\nabla\lambda\mathbf{p} \\ \mathbf{q}^T(\nabla\mathbf{D}\tilde{\mathbf{G}})\mathbf{p} + \mathbf{q}^T\lambda\nabla\mathbf{p} &= \mathbf{q}^T\lambda\nabla\mathbf{p} + \mathbf{q}^T\nabla\lambda\mathbf{p} \\ \mathbf{q}^T(\nabla\mathbf{D}\tilde{\mathbf{G}})\mathbf{p} &= \nabla\lambda. \end{aligned}$$

Continuing with the calculation of each component in the gradient vector $\nabla\lambda$, and using $\log\mathrm{SIR}_i = R_i$ on Pareto optimal surface $\mathcal{P}$, we have

$$\begin{aligned} \nabla_i\lambda(\mathbf{R}) &= \mathbf{q}^T(\mathbf{R})\nabla_i(\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}})\mathbf{p}(\mathbf{R}) \\ &= \mathbf{q}^T\nabla_i\left(\mathbf{diag}\left(\frac{e^{R_j}}{G_{jj}}\right)\tilde{\mathbf{G}}\right)\mathbf{p} \\ &= \sum_{j\neq i} q_i e^{R_i} p_j \frac{G_{ij}}{G_{ii}} \\ &= q_i\frac{\mathrm{SIR}_i}{G_{ii}}\sum_{j\neq i} G_{ij}p_j \\ &= q_i p_i, \end{aligned}$$

where the last equality comes from realizing that the power vector $\mathbf{P}$ is the same as the right Perron-Frobenious eigenvector $\mathbf{p}$ [6], [3]. Therefore, the normal to $\mathcal{P}$ is

$$\nabla\lambda(\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}}) = [q_1p_1, q_2p_2, \ldots, q_np_n]^T, \tag{16}$$

where $\mathbf{p}, \mathbf{q}$ are the right and left eigenvectors of $\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}}$ respectively, normalized such that $\mathbf{1}^T\nabla\lambda = 1$. We now find $\rho$ in the equality $\nabla U = \rho\nabla\lambda$: $\rho = \rho\mathbf{1}^T\nabla\lambda = \mathbf{1}^T\nabla U$. The optimality condition becomes:

$$\frac{U_i'}{\sum_j U_j'} = q_ip_i, \ \ \forall i. \tag{17}$$

Now consider a point $\mathbf{R}^k$ on $\mathcal{P}$. We would like to move along $\mathcal{P}$ to a point where the resulted $U(\mathbf{R})$ is larger. The tangent to $\mathcal{P}$ at $\mathbf{R}^k$ is a good local approximation to $\mathcal{P}$. So we move a small step $\alpha > 0$ along the tangent $\{\mathbf{R}|(\mathbf{R} - \mathbf{R}^k)^T\nabla\lambda(\mathbf{R}) = 0\}$ to $\tilde{\mathbf{R}}^k$ to increase $U$, *i.e.*,

$$\tilde{\mathbf{R}}^k = \mathbf{R}^k + \alpha\delta\mathbf{R}^k,$$

where $(\delta\mathbf{R}^k)^T\nabla\lambda(\mathbf{R}) = 0$ is orthogonal to the normal, such that $U(\tilde{\mathbf{R}}^k) > U(\mathbf{R}^k)$. In the following, we simplify the notation by suppressing index $k$ for $\delta\mathbf{R}$.

Due to concavity of $U(\mathbf{R})$, a positive $\delta\mathbf{R}_i$ decreases $\nabla_iU(\mathbf{R})$ and aligns the vectors $\nabla U$ and $\nabla\lambda$. Therefore, moving along the direction of $\frac{U_i'}{\sum_j U_j'} - \nabla\lambda$ increases $U(\mathbf{R})$. We diagonally scale it by $\frac{1}{p_iq_i}$:

$$\delta\mathbf{R} = \mathbf{diag}\left(\frac{1}{p_iq_i}\right)\left(\frac{U_i'}{\sum_j U_j'} - \nabla\lambda\right),$$

so that the resulted point is on the tangent:

$$\begin{aligned} \nabla\lambda^T\delta\mathbf{R} &= \sum_i\left[\frac{U_i'}{\sum_j U_j'} - p_iq_i\right] \\ &= 1 - \sum_i p_iq_i \\ &= 0, \end{aligned}$$

since right and left Perron Frobenious eigenvectors $\mathbf{p}, \mathbf{q}$ of $D\tilde{G}$ are normalized: $\mathbf{q}^T\mathbf{p} = 1$.

The point $\tilde{\mathbf{R}}^k$ in general is not on $\mathcal{P}$ and may be infeasible. We now project $\tilde{\mathbf{R}}^k$ on $\mathcal{P}$ to obtain $\mathbf{R}^{k+1}$ as the next rate allocation vector. We subtract a constant term $\log \lambda(\mathbf{D}(\tilde{\mathbf{R}}^k)\tilde{\mathbf{G}})$ in each component of $\tilde{\mathbf{R}}^k$. This scales the $\mathbf{D} = \textbf{diag}(e^{R_i})$ matrix by $\frac{1}{\lambda(\mathbf{D}(\tilde{\mathbf{R}}^k)\tilde{\mathbf{G}})}$. Therefore, the new rate vector:

$$\mathbf{R}^{k+1} = \tilde{\mathbf{R}}^k - (\log \lambda(\mathbf{D}(\tilde{\mathbf{R}}^k)\tilde{\mathbf{G}}))\mathbf{1}^T$$

is on Pareto optimal surface $\mathcal{P}$, as verified below:

$$\begin{aligned} \mathbf{D}(\mathbf{R}^{k+1})\tilde{\mathbf{G}}\mathbf{P} &= \frac{1}{\lambda(\mathbf{D}(\tilde{\mathbf{R}}^k)\tilde{\mathbf{G}})}\mathbf{D}(\tilde{\mathbf{R}}^k)\tilde{\mathbf{G}}\mathbf{P} \\ &= \frac{1}{\lambda(\mathbf{D}(\tilde{\mathbf{R}}^k)\tilde{\mathbf{G}})}\lambda(\mathbf{D}(\tilde{\mathbf{R}}^k)\tilde{\mathbf{G}})\mathbf{P} \\ &= \mathbf{P} \\ \Rightarrow \lambda(\mathbf{D}(\mathbf{R}^{k+1})\tilde{\mathbf{G}}) &= 1. \end{aligned}$$

The sequence of rate allocation adaptation $\mathbf{R}^k \rightarrow \tilde{\mathbf{R}}^k \rightarrow \mathbf{R}^{k+1}$ through movement along the tangent and projection to $\mathcal{P}$ produces an iteration of feasible rate vectors, which by the optimality condition of (15) converges to an optimal rate allocation $\mathbf{R}^*$ for any possible power control. As rate vector $\mathbf{R}$ adapts, power vector $\mathbf{P}$ changes according to the Perron Frobenius update in [6], which converges since $\{\mathbf{R}^k\}$ are feasible. In particular, because $\mathbf{R}^*$ is on the boundary of the feasible rate region, there is a corresponding power control $\mathbf{P}^*$ that produces $\mathbf{R}^*$. By convexity properties shown, this pair of $(\mathbf{R}^*, \mathbf{P}^*)$ is indeed a globally and jointly optimal rate and power vectors for utility maximization (11).

Note that each component of the gradient $\nabla\lambda$:

$$\frac{\partial\lambda(\mathbf{D}(\mathbf{R})\tilde{\mathbf{G}})}{\partial R_i} = p_i q_i,$$

is only a function of right Perron Frobenius eigenvector $\mathbf{p}$ and left Perron Frobenius eigenvector $\mathbf{q}$, but not of individual interferers' powers. It is known [6] that $\mathbf{p}$ is equivalent to the transmit power vector $\mathbf{P}$. Left Perron Frobenius eigenvector also has an intuitive interpretation: it is a 'summary' of the effect of all the global interferences on utility maximization. Either a higher power $p_i$ or a higher 'summary' $q_i$ of the interference effect implies a higher 'price' for utility maximization.

### E. *Numerical example*

An illustrative example of Algorithm 2 is shown through a simulation summarized below. We simulate a cellular system with five users connecting to a base station. The path loss is based on the randomly generated distances and specified coding gain. We then randomly generate an initial set of rate requirements, *i.e.*, $\mathbf{R}^0$ in Algorithm 2. The objective is to maximize the total sustainable data rate of the network. As the algorithm proceeds, transmit power and allocated rate for each user is shown in Figures 8 and 9. It is observed in Figure 10 that the sum rate of the system increases as the algorithm converges.

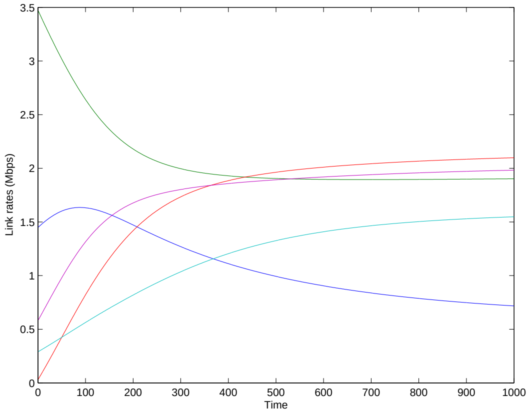

Fig. 8. Algorithm 2 example: rate adaptation.

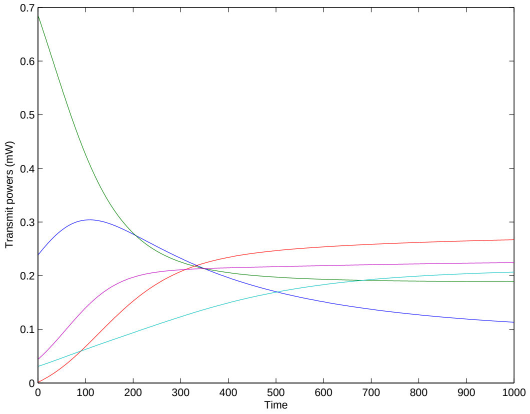

Fig. 9. Algorithm 2 example: power adaptation.

## IV. End-to-end Hybrid Network Case

### A. *Introduction*

Each end-to-end path in a wireless cellular network consists of a wireless air-interface and a wired backbone network. We have considered only the air-interface part thus far, using physical layer models that assume simple, sub-optimal coding and modulation schemes. In this section, we consider the end-to-end problem across both the air-interface and the backbone. We will also use the information-theoretic fading channel capacity region for the air-interface model. We only discuss the case for data upload from the wireless users, *i.e.*, the air-interface is a multiple access channel (MAC). For the case of data download to the wireless users, *i.e.*, the air-interface is a broadcast channel (BC), the results developed in this section can be easily extended.

The wireless MAC is often modelled as time-varying fading channels, and the main issue is how to make the most efficient use of the available bandwidth and power. In particular, assuming that both the transmitters and receiver have channel state information, power control at the transmitters can increase the achievable data rates on the wireless hop. Tse and Hanly

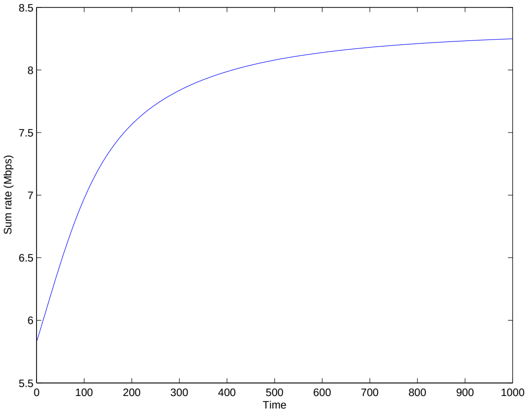

Fig. 10. Algorithm 2 example: sum rate increases.

[23] showed a greedy algorithm for optimal rate and power allocation, which attains the boundary of the channel capacity region of a multiaccess fading channel. The algorithm makes use of a rate reward vector and a power price vector [23].

In the wired backbone network where the links provide transmission 'pipes' of fixed 'sizes', the main issue is to avoid overloading the links. End-to-end congestion control mechanisms, such as those in TCP, are usually used to regulate the allowed rate from each source node. The goal is to prevent any source node from 'pumping' so much data into the network that the total flow on any link exceeds its available capacity. Kelly et. al. [11] showed that distributed rate allocation can be viewed as primal-dual algorithms implicitly maximizing a network utility under link capacity constraints. Recent papers (*e.g.*, [15], [16]) further established the equivalence between TCP congestion control algorithms and different network utility maximization problems, where congestion signals, such as queuing delays, act as pricing variables. These congestion prices are updated at routers and fed back to the sources.

In our hybrid network model in Figure 3, which accurately depicts the end-to-end connections in wireless cellular networks, there is an interesting tradeoff between rate-power allocation local to the wireless MAC and congestion control in wired links that regulate both wireless source nodes and other source nodes in the network. Indeed, other source nodes, *e.g.*, DSL connected servers, may share a backbone link with a source node $A$ connected through a wireless LAN. Suppose at a particular time, the congestion control mechanism informs the wireless source node $A$ to increase its transmission rate, possibly because other sources sharing a bottleneck link with $A$ are transmitting less and the congestion price becomes favorable to $A$. However, due to a particular fading state, the local wireless MAC power price may dictate that $A$ should not be allowed to increase its rate, for otherwise a boundary point on the multiaccess fading capacity region cannot be obtained. Wireless source node $A$ must resolve this conflict between maximizing 'global utility' for end-to-end transmissions and achieving 'local capacity' at the wireless MAC. This conflict is most apparent when laptops in a wireless LAN upload files using TCP as the transport layer protocol, because then the bandwidth requirements from wireless source nodes represent significant portions at some wired backbone links. This section presents an algorithm to trace the Pareto optimal tradeoff curve between these two competing objectives.

## B. Background

We index by $s$ the source nodes connected by wired links to the backbone and denote their allowed transmission rates by $\{x_s\}$. We index by $i$ the source nodes connected by the wireless hop and denote their allowed rates by $\{R_i\}$ and their transmit powers by $\{P_i\}$. We consider end-to-end transmission from both types of sources, assuming fixed and known routing, where $L(i)$ denotes the set of links $l$ traversed by the connection originating from source $i$. Let $\mathcal{L}$ be the set of links that are shared by traffic from both types of source nodes, *i.e.*, links in the backbone experiencing the coupling effects between wireless-hop channel variations and wired backbone congestion. The wireless uplink is modelled as a standard multiaccess fading Gaussian channel with the fading processes $\{H_i\}$ known at the transmitters and receiver. Wired backbone is assumed to have fixed-capacity links for given coding/modulation schemes (in contrast to the wireless ad hoc networks considered in [4]). For notational simplicity, we assume there is only one wireless air-interface with $M$ source nodes using the backbone network.

If we focus only on the wireless MAC, Tse and Hanly [23] showed that the multiaccess fading Gaussian channel capacity boundary is the closure of all $\mathbf{R}$ that are the optimizers of the following optimization problem:

$$\begin{array}{ll} \text{maximize} & \sum_i \mu_i R_i \\ \text{subject to} & \mathbf{R} \in \mathcal{R}(\mathbf{P}), \\ & \mathbf{R}, \mathbf{P} \succeq 0 \end{array} \tag{18}$$

where the variables are $\mathbf{R}$ and $\mathbf{P}$, and $\boldsymbol{\mu}$ is a given rate reward vector. The set $\mathcal{R}(\mathbf{P})$ in the first constraint contains all $\mathbf{R}$ such that

$$\sum_{i\in\mathcal{S}} R_i \leq \mathbf{E}\left[\frac{1}{2}\log\left(1+\frac{1}{\sigma^2}\sum_{i\in\mathcal{S}} H_i P_i(\mathbf{H})\right)\right], \forall \mathcal{S} \subset \{1,\ldots,M\}.$$

Utilizing the polymatroid structure of the constraint set and introducing a 'power price' vector $\boldsymbol{\lambda}$ to coordinate $\mathbf{R}$ and $\mathbf{P}$, Tse and Hanly [23] presented a greedy rate and power allocation algorithm that solves the above problem.

If we instead assume that the entire network consists of only wired links with fixed capacities $\{c_l\}$, Kelly et al. [11] showed that distributed rate allocation across the network is implicitly solving a network utility maximization problem:

$$\begin{array}{ll} \text{maximize} & \sum_s U_s(x_s) \\ \text{subject to} & \sum_{s:l\in L(s)} x_s \leq c_l, \ \ \forall l, \\ & \mathbf{x} \succeq 0 \end{array} \tag{19}$$

where the variables are $\mathbf{x}$. Low et al. [15] further showed how to obtain the utility maximization implicitly being solved for based on a congestion control protocol, as well as how to design a congestion control mechanism starting from some given utility functions. For example, TCP Vegas [16] is implicitly maximizing a weighted logarithmic utility using queuing delays $\{\gamma_l\}$ as the congestion prices to regulate $\{x_s\}$.

### C. Problem formulation

In our hybrid network, both wireless MAC and wired backbone network are present and they are coupled in two ways. They are coupled in the constraints because those links $l \in \mathcal{L}$ in the backbone are shared by traffic due to both $\{x_s\}$ and $\{R_i\}$. They are also coupled in the objective function, because the wireless first-hop is often the end-to-end performance bottleneck and we would like to attain a point on the capacity region's boundary, yet the global network utility should also be maximized. How should $(\mathbf{R}, \mathbf{P})$ be chosen to balance the two? And how may $\mathbf{x}$ be adapted to induce a favorable congestion condition in the backbone so that it becomes feasible for $\mathbf{R}$ to be varied in order to reach the wireless MAC capacity boundary?

These intuitive questions are formulated in the following problem of end-to-end resource allocation in hybrid networks, essentially a 'weighted sum' of (18) and (19), with the second constraint coupling across the wireless and wired parts:

$$\begin{array}{ll} \text{maximize} & \sum_s U_s(x_s) + \theta \sum_i \mu_i R_i \\ \text{subject to} & \sum_{s:l\in L(s)} x_s \le c_l, \;\; \forall l \notin \mathcal{L}, \\ & \sum_{i:l\in L(i)} R_i + \sum_{s:l\in L(s)} x_s \le c_l, \;\; \forall l \in \mathcal{L}, \\ & \mathbf{R} \in \mathcal{R}(\mathbf{P}), \\ & \mathbf{x}, \mathbf{R}, \mathbf{P}, \succeq 0 \end{array} \tag{20}$$

where the variables are the wired source rates $\mathbf{x}$, wireless source rates $\mathbf{R}$, and wireless transmit powers $\mathbf{P}$. The constant parameters include weighting coefficient $\theta$, rate reward vector $\boldsymbol{\mu}$, and wired link capacities $\mathbf{c}$. Due to concavity of the objective and convexity of the constraints, varying $\theta \in (0, \infty)$ and solving (20) for each $\theta$ traces out the entire Pareto optimal curve between the local and global objectives in end-to-end resource allocation in hybrid networks.

### D. Algorithm

We provide the following distributed algorithm to solve (20).

**Algorithm 3**

First, initiate $\alpha_l \ge 0$ on links $l \in \mathcal{L}$ and feed it back to all sources using link $l$, then repeat in parallel the following iterations until convergence:

1. Apply the congestion control algorithm in [15] but use the following modified utility functions $U'_s(x_s)$:

$$U'_s(x_s) = \begin{cases} U_s(x_s) - \sum_{l\in L(s)} \alpha_l & \text{if } s \text{ shares links with some } i \\ U_s(x_s) & \text{otherwise} \end{cases} \tag{21}$$

2. Apply the greedy rate and power allocation algorithm in [23] but use the following modified rate vector $\boldsymbol{\mu}'$:

$$\mu'_i = \begin{cases} \theta\mu_i - \sum_{l\in L(i)} \alpha_l & \text{if } i \text{ shares links with some } s \\ \theta\mu_i & \text{otherwise} \end{cases} \tag{22}$$

3. Update a 'virtual buffer' $h_l$ on each link $l \in \mathcal{L}$ (note that $\{x_s\}$ and $\{R_i\}$ are resulted from items 1 and 2 above, and they are functions of $\boldsymbol{\alpha}$):

$$h_l(\boldsymbol{\alpha}) = c_l - \left( \sum_{s:l\in L(s)} x_s(\boldsymbol{\alpha}) + \sum_{i:l\in L(i)} R_i(\boldsymbol{\alpha}) \right). \tag{23}$$

4. Update $\alpha_l$ and feed it back to all sources using link $l$ (where $\beta > 0$ is a constant):

$$\alpha_l^{(k+1)} = \left[ \alpha_l^{(k)} - \beta h_l^{(k)} \right]^+ . \tag{24}$$

In essence, $\boldsymbol{\alpha}$ acts as an additional set of 'coupling' prices that affects the existing algorithms in [15] and [23] by modifying utility functions and rate rewards, thus coupling congestion prices $\boldsymbol{\gamma}$ and power prices $\boldsymbol{\lambda}$. There are several propositions we can make about the performance and practical issues of this algorithm. Due to space limit in this summary, we focus on the most important issue in the following theorem that will be proved in the next section:

*Theorem 3:* Algorithm 3 (21,22,23,24) converges to a global optimum $(\mathbf{P}^*, \mathbf{R}^*, \mathbf{x}^*)$ of utility maximization (20).

Numerical examples that trace out the desired Pareto optimal tradeoff curve using the solution above, as well as theoretical corollaries utilizing the result of modified $\{U'_s\}$ and $\{\mu'_i\}$, easily follow from Algorithm 3. For example, [27] showed that for Poisson arrival traffic to the wireless source nodes, letting the rate reward $\{\mu_i\}$ be the queue sizes $\{q_i\}$ and then using the Tse-Hanly algorithm is 'throughput optimal' for multiaccess fading channels. Our result on modified rate reward $\boldsymbol{\mu}'$ implies that, for end-to-end consideration in hybrid networks, the rate reward should now be a weighted sum of local buffer size and all the 'virtual buffer' sizes along the path in the wired backbone: $\mu_i = \frac{1}{\theta}(q_i + \sum_{l\in L(i)} \alpha_l), \;\; \forall i$.

### E. Proof of Theorem 3

We first rewrite problem (20) by introducing dummy variables $t_l$ for $l \in \mathcal{L}$:

$$\begin{array}{ll} \text{maximize} & \sum_s U_s(x_s) + \theta \sum_i \mu_i R_i \\ \text{subject to} & \sum_{s:l\in L(s)} x_s \le c_l, \;\; \forall l \notin \mathcal{L}, \\ & c_l - \sum_{s:l\in L(s)} x_s = t_l, \;\; \forall l \in \mathcal{L}, \\ & t_l \ge \sum_{i:l\in L(i)} R_i, \;\; \forall l \in \mathcal{L}, \\ & \mathbf{R} \in \mathcal{R}(\mathbf{P}), \\ & \mathbf{x}, \mathbf{t}, \mathbf{R}, \mathbf{P} \succeq 0. \end{array} \tag{25}$$

We now introduce Lagrange multipliers $\alpha_l$ for $l \in \mathcal{L}$, and write the partial Lagrangian:

$$L(\mathbf{x}, \mathbf{t}, \mathbf{R}, \mathbf{P}, \boldsymbol{\alpha}) = \sum_s U_s(x_s) + \theta \sum_i \mu_i R_i + \sum_{l\in\mathcal{L}} \alpha_l (t_l - \sum_{i:l\in L(i)} R_i).$$

We can maximize the partial Lagrangian to obtain the Lagrange dual function:

$$g(\boldsymbol{\alpha}) = \sup_{\mathbf{x},\mathbf{t},\mathbf{R},\mathbf{P}} \{L(\mathbf{x},\mathbf{t},\mathbf{R},\mathbf{P},\boldsymbol{\alpha}) | \text{other constraints in (25)}\}.$$

It is easy to see that the Lagrange dual function can be obtained by a partial Lagrangian decomposition:

$$g(\boldsymbol{\alpha}) = g_{net}(\boldsymbol{\alpha}) + g_{mac}(\boldsymbol{\alpha})$$

where $g_{net}(\boldsymbol{\alpha})$ and $g_{mac}(\boldsymbol{\alpha})$ are, respectively, the optimized value of the objective function in the following subproblem of *NET*:

$$\begin{array}{ll} \text{maximize} & \sum_s U_s(x_s) + \sum_{l\in\mathcal{L}} \alpha_l t_l \\ \text{subject to} & \sum_{s:l\in L(s)} x_s \le c_l, \;\; \forall l \notin \mathcal{L}, \\ & t_l + \sum_{s:l\in L(s)} x_s = c_l, \;\; \forall l \in \mathcal{L}, \\ & \mathbf{x}, \mathbf{t} \succeq 0 \end{array} \qquad (26)$$

where the variables are $\mathbf{x}$ and $\mathbf{t}$ only, and the subproblem of *MAC*:

$$\begin{array}{ll} \text{maximize} & \theta \sum_i \mu_i R_i - \sum_{l\in\mathcal{L}} \alpha_l \sum_{i:l\in L(i)} R_i \\ \text{subject to} & \mathbf{R} \in \mathcal{R}(\mathbf{P}), \\ & \mathbf{R}, \mathbf{P} \succeq 0 \end{array} \qquad (27)$$

where the variables are $\mathbf{R}$ and $\mathbf{P}$ only.

Notice that, as desired for decoupling problem (20), in the *NET* subproblem, the wireless transmit powers and rates are not being optimized. Similarly, in the *MAC* subproblem, the source rates from the non-wireless nodes are not being optimized.

Since utility functions are concave, the multiaccess fading constraint is convex [23], and all other terms in the objective and constraint functions are affine, (20) is a convex optimization problem (with strictly feasible solutions). Therefore, solving (20) is equivalent to solving its Lagrange dual problem over $\boldsymbol{\alpha}$:

$$\begin{array}{ll} \text{maximize} & g_{net}(\boldsymbol{\alpha}) + g_{mac}(\boldsymbol{\alpha}) \\ \text{subject to} & \alpha_l \ge 0, \;\; \forall l \in \mathcal{L}. \end{array} \qquad (28)$$

In order to solve (28), we need to first know how to solve (26) and (27) to obtain $g_{net}$ and $g_{mac}$ as functions of $\boldsymbol{\alpha}$, and then how to maximize the sum of $g_{net}$ and $g_{mac}$ over $\boldsymbol{\alpha}$. We show below that (26) can be distributively solved using existing congestion control algorithms over the network by a simple modification of the utility functions, and that (27) can be locally solved using Tse and Hanly's greedy algorithm [23] with a simple modification of the rate reward vector.

First, substituting the second constraint in the *NET* subproblem (26), we can rewrite the objective function of (26) as $\sum_s U_s(x_s) + \sum_{l\in\mathcal{L}} \alpha_l (c_l - \sum_{s:l\in L(s)} x_s)$. Since $\sum_{l\in\mathcal{L}} \alpha_l c_l$ is a constant term, the objective function can be equivalently written as

$$\sum_s U_s(x_s) - \sum_{s:L(s)\bigcap\mathcal{L}\neq\phi} \left( \sum_{l\in L(s)} \alpha_l \right) x_s.$$

We thus define modified utility functions $U'_s$ for all sources $\{s : L(s)\bigcap\mathcal{L} \neq \phi\}$ (*i.e.*, those sharing some links in the backbone with a wireless source node): $U'_s(x_s) = U_s(x_s) - \sum_{l\in L(s)} \alpha_l$. The utility functions for other sources remain the same. The *NET* problem (26) now becomes

$$\begin{array}{ll} \text{maximize} & \sum_s U'_s(x_s) \\ \text{subject to} & \sum_{s:l\in L(s)} x_s \le c_l, \;\; \forall l \notin \mathcal{L}, \\ & \mathbf{x} \succeq 0. \end{array} \qquad (29)$$

This is now in exactly the same form as the network utility maximization problem (19) that can be distributively solved by TCP congestion control mechanisms in [11], [15], [16].

Similarly for the *MAC* subproblem (27), we can modify the rate reward $\mu'_i = \theta\mu_i - \sum_{l\in L(i)} \alpha_l$ for all wireless source nodes $i$ sharing some links with non-wireless source nodes, and $\mu'_i = \theta\mu_i$ for all other $i$. Problem (27) can now be expressed as:

$$\begin{array}{ll} \text{maximize} & \sum_i \mu'_i R_i \\ \text{subject to} & \mathbf{R} \in \mathcal{R}(\mathbf{P}), \\ & \mathbf{R}, \mathbf{P} \succeq 0. \end{array} \qquad (30)$$

This is now in exactly the same form as the optimization problem (18) that characterizes the boundary of multiaccess fading channels, which can be solved by the greedy algorithm in [23].

Now that we have reduced both subproblems into the forms that can be solved with existing algorithms, we proceed to solve the 'master' dual problem (28) by optimizing over $\boldsymbol{\alpha}$. Due to non-strict concavity and convexity in the primal problem (20), the objective function in (28) may not be differentiable, and we have to use the subgradient method [1], [24], which extends the gradient descent method to non-differentiable functions. Subgradient method can be made distributive in this case. It can be verified that the following vector $\mathbf{h}$ is a subgradient to $g_{net}(\boldsymbol{\alpha}) + g_{mac}(\boldsymbol{\alpha})$:

$$h_l(\boldsymbol{\alpha}) = c_l - \left( \sum_{s:l\in L(s)} x_s(\boldsymbol{\alpha}) + \sum_{i:l\in L(i)} R_i(\boldsymbol{\alpha}) \right).$$

This subgradient can be interpreted as 'virtual buffers': the difference between bandwidth supply $c_l$ and bandwidth demand from both wireless and non-wireless source nodes, which can be measured locally on each link. Then at time instant $k$, variables $\boldsymbol{\alpha}$ are updated by the subgradient method: $\alpha_l^{(k+1)} = \left[\alpha_l^{(k)} - \beta h_l^{(k)}\right]^+, \quad \forall l \in \mathcal{L}$ where $\beta > 0$ is a constant step size. Using a similar argument as in [4], it can be shown that for a small enough $\beta$, the subgradient iterations converge. Convergence for both subproblem *NET* (26) and subproblem *MAC* (27) can be established after the transformation to (29) and (30), respectively. Because the problems are convex optimization problems, the convergence is toward a global optimum of (28), or equivalently, of (20).

We conclude this section by mentioning that for end-to-end resource allocation for data download to (instead of upload from) wireless users, similar modifications of rate rewards can be made to the algorithm of power and rate allocation to achieve the broadcast fading channel capacity in [13], [22].

## V. Conclusions

Compared to wired networks, utility maximization problems in wireless cellular networks have additional challenges:

- Utility maximization depends on the channel conditions.
- Flow constraints on rates become constraints coupled between powers and rates.
- Linear flow constraints become nonlinear, and may be globally coupled due to interference.
- For end-to-end consideration, there is a conflict between attaining air-interface capacity and maximizing global utility.

This paper presents three algorithms that are provably convergent to the joint and globally optimal pair of rates and powers for three cases of utility maximization in wireless cellular networks:

1) When the rate is a nonlinear and global function of all transmit powers, Algorithm 2 iteratively updates powers and rates, and converges to a globally optimal rate-power pair for utility maximization.
2) When the rate is a nonlinear but local function of the transmit power, the bidding mechanism in Algorithm 1 maximizes network utility without requiring the knowledge of channel conditions and individual utility functions at the base station, thus extending a similar conclusion for wired networks in Kelly et al. [11] to the case of wireless downlink transmissions.
3) For the problem of end-to-end resource allocation for wireless cellular networks formulated in this paper, it is solved distributively by using coupling prices and virtual buffers, and by modifying source utility functions in [11] and rate reward vectors in [23].

Constrained nonlinear optimization theory and Lagrange duality have been the important tools to solve these problems, and convexity properties of the optimization problems are crucial to the proofs. The proof of Theorem 1 also utilizes the special structure of the Lagrange dual problem that leads to a successive Gauss-Siedel algorithm where channel conditions and local utilities can be decoupled from the base station updates. The proof of Theorem 2 relies on Perron Frobenius theory of non-negative matrices to characterize the KKT optimality condition. The proof of Theorem 3 uses partial Lagrangian decomposition and subgradient method for distributed implementation.

## Acknowledgements

We would like to acknowledge helpful discussions with Stephen Boyd, Dan O'Neill, David Julian, and Lin Xiao at Stanford University, Steven Low at Caltech, Edmund Yeh at Yale University, Randy Berry at Northwestern University, and Peter Marbach at University of Toronto.

## References

[1] D. P. Bertsekas, *Nonlinear Programming,* 2nd Ed., Athena Scientific, 1999.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation,* Prentice Hall 1989.

[3] N. Bambos, S. C. Chen, G. J. Pottie, "Radio link admission algorithm for wireless networks with power control and active link quality protection, " *Proc. IEEE Infocom,* April 1995.

[4] M. Chiang, "To layer or not to layer: balancing transport and physical layers in wireless multihop networks," *Proc. IEEE INFOCOM,* March 2004.

[5] C. Douglígeris and R. Mazumdar, "A game theoretic perspective to flow control in telecommunication networks," *J. Franklin Inst.,* vol. 329, pp.383-402, March 1992.

[6] G. J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence," *IEEE Trans. Veh. Tech.,* vol. 42, pp.6441-646, April, 1993.

[7] D. Goodman and N. Mandayam, "Power control for wireless data," *IEEE Pers. Comm.,* vol. 7, pp.48-54, April 2000.

[8] R. A. Horn and C. R. Johnson, *Matrix Analysis,* Cambridge University Press, 1987.

[9] H. Ji and C. Y. Huang, "Non-cooperative uplink power control in cellular radio systems," *Wireless Networks,* vol. 4, pp.233-240, April 1998.

[10] F. P. Kelly, "Charging and rate control for elastic traffic," *European Trans. on Telecommunication,* vol. 8, pp33-37, 1997.

[11] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of Operations Research Society,* vol. 49, no. 3, pp.237-252, March 1998.

[12] T. E. Klein and H. Viswanath, "Centralized power control for multi-hop wireless networks," *Proc. ISIT,* July 2003.

[13] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadbacast channels, part I: ergodic capacity," *IEEE Trans. Inform. Theory,* vol. 47, no. 3, pp. 1083-1102, March 2001.

[14] P. Liu, M. Honig, and S. Jordan, "Forward link CDMA resource allocation based on pricing," *Proc. IEEE Wireless Communications and Networking Conference,* Sept. 2000.

[15] S. H. Low, F. Paganini, and J. C. Doyle, "Internet congestion control," *IEEE Control Systems Magazine,* February 2002.

[16] S. H. Low, L. L. Perterson, and L. Wang, "Understanding Vegas: a duality model," *Journal of the ACM,* vol. 49, no. 2, pp.207-235, March 2002.

[17] P. Marbach and R. Berry, "Downlink resource allocation and pricing for wireless networks," *Proc. IEEE INFOCOM,* June 2002.

[18] D. Mitra, "An asynchronous distributed algorithm for power control in cellular radio systems," *Proc. Winlab Workshop Third Generation Wireless Information Networks,* 1993.

[19] D. O'Neill, "Power control for wireless networks using TCP/RED," *Proc. IEEE WNCN,* 2003.

[20] C. Saraydar, N. Mandayam, and D. Goodman, "Pricing and power control in a multicell wireless data network." *IEEE J. Sel. Areas in Comm.,* 19(10):1883-1892, 2001.

[21] E. Seneta, *Non-negative Matrices,* Springer Verlag, 1981.

[22] D. N. C. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," *Proc. ISIT,* Ulm, Germany, July 1997.

[23] D. N. C. Tse and S. V. Hanly, "Multiaccess fading channels - Part I: polymatroid structure, optimal resource allocation, and throughput capacities," *IEEE Trans. Inform. Theory,* vol.44, no. 7, pp.2796-2815, November 1998.

[24] L. Xiao, M. Johansson, and S. Boyd, "Simultaneous routing and resource allocation for wireless networks," *Proc. IEEE Conference on Decision and Control,* Dec. 2001.

[25] M. Xiao, N. B. Shroff, and E. K. P. Chong, "A utilit-based power-control scheme in wireless cellular systems," *IEEE/ACM Trans. Networking,* vol. 11, no. 2, pp. 210- 221, April 2003.

[26] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Comm.,* vol. 13, pp. 1341-1347, September 1995.

[27] E. Yeh and A. Cohen, "Maximum throughput and minimum delay in fading multiaccess communications," *Proc. International Symposium of Information Theory,* July 2003.