Seeking Foschini's Genie: Optimal Rates and Powers in Wireless Networks

Dan ONeill, David Julian, and Stephen Boyd
Department of Electrical Engineering
Stanford University
dconeill,djulian,boyd@stanford.edu

***Abstract*— An adaptive rate and power control approach is presented in this paper. The approach is layered, with a rate optimization algorithm selecting the optimal rates for the system and a power control algorithm subsequently calculating the associated powers. The approach adapts to changes in demands on the system such as the arrival or departure of users, changes in the rate thresholds expected by users, and changes in the mix of traffic presented to the system.**

Basic properties and performance limits are established within the mathematical framework of Perron Frobenius matrix theory. The right and left eigenvectors of a matrix associated with the network are found to describe the optimal performance of each link in the system. Leveraging this analysis, a method is described that seeks optimal rates. A simulation demonstrates the performance gain associated with this approach.

***Index Terms*— wireless networks, power and rate control, utility functions, optimization, adaptation and admission, Perron Frobenius**

I. Introduction

FUTURE wireless networks will need to automatically adjust link rates and transmitter powers in response to changes in user demand, network configuration, or QoS requirements. A wireless Internet terminal may demand a large bursty data rate when compared to that of a typical cell phone. Likewise, two way video will demand data rates substantially different from those of voice only communications. In multi-hop networks the problem is exacerbated by the need to account for multiple data flows across multiple links. The challenge for the network is to find the "best" link rates and associated set of "best" transmitter powers to support the demands on the system.

There are two general ways to solve the problem of joint power and rate optimization. The first is to simultaneously optimize both the powers and rates in one algorithm, and the second is to iteratively optimize the rates and powers separately. The joint optimization problem can be solved globally and efficiently by many methods, such as geometric programming [1], [2], [3]; however, we are not aware of a distributed implementation for the joint optimization.

The second way to optimize is to iteratively optimize the powers and rates. The problem of finding the optimal set of transmitter powers for a known set of link data rates has been extensively investigated. The problem of finding the optimal set of link data rates, however, has received considerably less attention [4], [5], [6], possibly because until recently link rates have generally been fixed. G. Foschini and Z. Miljanic in their paper on distributed power control [7] addressed the issue of finding optimal rates by positing a "Genie" that provided the network optimal target link rates.

This paper presents a method for finding the optimal link rates in a network, and, in this way, functions as the Genie described by G. Foschini and Z. Miljanic. Optimal in this context means that the set of rates maximizes a system performance measure or alternatively a set of user utility functions. The system gain matrix is key to this approach. This matrix of power gains implicitly defines the set of link rates feasible for the system. This set is convex and can be used to find the optimal rates for reasonable performance measures.

Figure 1 illustrates the basic concept. The upper level system corresponds to the rate optimization algorithm and the lower level system to the algorithm for calculating the correct powers for the optimal rates. The powers can be calculated in several different ways. A distributed approach is suggested in [8]. The two levels exchange information in an iterative fashion. The rate optimization algorithm provides rate targets for the power algorithm and receives several types of power information in return. The variables $p, q, \lambda_{\rm pf}$ correspond to the transmitter powers for the network, the transmitter powers for the adjoint network (with $G' = G^T$) which might be used to carry acknowledgements, and the Perron Frobenius eigenvalue associated with the system, which is discussed extensively in this paper. Interestingly, p and q correspond to the right and left eigenvectors associated with the system. Surprisingly, q completely characterizes the interference seen by links in the network. In particular q will be seen

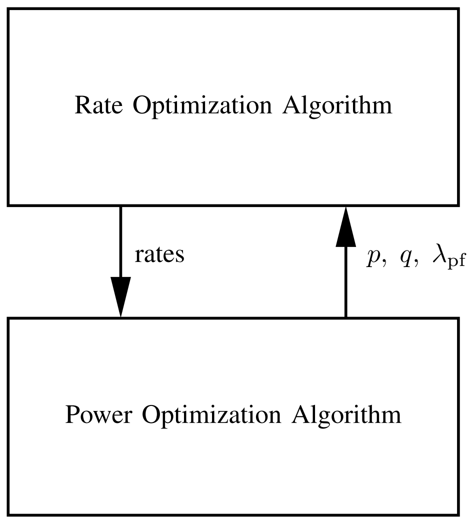

Fig. 1. Rate and Power Control Relationship.

to scale the effect of transmitter powers on link rates.

The rate optimization algorithm is adaptive. As users enter or leave the network, or users change their rate demands, the algorithm automatically searches for rates that best meet the new demands or set of QoS requirements. The rate optimization algorithm searches the set of feasible rates by systematically exploring the associated rate-region. This region is defined by the Perron Frobenius eigenvalue associated with the network, and is also convex. As users change their requirements, the algorithm iteratively seeks to align these new demands with points in the rate-region in such a way that the system performance metric is maximized.

This paper is divided into several sections. Section II describes the wireless network model. Section III mathematically formulates the problem. Section IV analyzes the problem and develops tools for developing a rate optimization algorithm. Section V describes the rate optimization algorithm. Finally, section VI presents simulation results.

II. System Model

A. Wireless Network

This paper considers a network of n links, each link with one transmitter and one receiver. All links utilize a CDMA transmission scheme and share the same bandwidth.

The line of sight propagation model between a transmitter and a receiver is

$$p_r = pK_{\mathrm{d}}\left(\frac{d_0}{d}\right)^{\gamma}, \tag{1}$$

where p_r is the received power, p is the transmitted power, d is the propagation path length, and d_0 is a reference distance for the antenna far-field, usually taken so that the normalization constant K_{d} equals 1. The path loss exponent γ is usually between 2 and 6 for most propagation environments.

B. Signal to Interference Ratio

The SIR for the ith link is defined as

$$\rho_i = \frac{K_{\mathrm{s}} p_i L_{ii} d_i^{-\gamma_i}\alpha_i}{\sum_{j\neq i}^{N} p_j L_{ij} d_j^{-\gamma_j}\alpha_j}. \tag{2}$$

The constant K_{s} is the spreading gain for the link, and the constant L_{ii} captures the effect of fading for the path from the transmitter to the receiver on the ith link. The L_{ij} are similarly defined as the gain associated with fading from the transmitter on the jth link and the receiver on the ith link. Receiver noise is assumed to be dominated by interference power and is neglected in this model. The factors α_j are introduced to accommodate normalization constants and other factors, such as the effect of beam-forming in multi-antenna systems.

To simplify notation the various constants in this equation can be collected and SIR rewritten as

$$\rho_i = \frac{G_{ii}p_i}{\sum_{j\neq i} G_{ij}p_j}. \tag{3}$$

G_{ii} represents the effective gain between the transmitter and receiver on link i and includes the multiplicative spreading gain K_{s}, antenna gain, coding gain, and other gain factors. Likewise G_{ij} represents the effective gain from the interfering transmitter on link j to the receiver on link i. Note that this allows for more general models than that given in equations 1 and 2. The gain matrix G with elements G_{ij} is assumed to be irreducible in what follows.

Because noise is neglected in this model, the set of powers can be arbitrarily scaled without effecting ρ_i; That is, SIR is homogeneous of order zero in the transmitter powers. By choosing to scale the sum to one, $\mathbf{1}^T p = 1$, the p_i can be interpreted as representing the relative powers of the transmitters or equivalently the percent of total power transmitted by the system.

In most systems $\rho_i \gg 1$ since it represents the effective SIR after spreading gain, antenna gain, and coding gain

C. Link Rate Transfer Rate

The link rate R_i specifies the capacity of the link in units of bits or fixed length packets per second for a given power vector p. It models the physical capacity of the link and is the maximum rate at which data can actually be transferred over the link. The transfer rate r_i is the rate at which the system chooses to actually send data over the link. For a given p the transfer rate r_i is necessarily less than the link rate R_i, $r_i \leq R_i$. For obvious physical reasons both rates are non-negative, $r_i, R_i \geq 0$.

An empirically based link rate model for M-QAM and M-PSK modulation is [9]

$$R_i = \alpha \log(1 + K\rho_i), \tag{4}$$

where $K = (-1.5)/(\ln(5\text{BER}))$ and BER is the target average bit error rate. The constant α is a scaling constant and represents the base of the logarithm used and several other system constants.

This is similar in form to the information theoretic capacity model,

$$C_i = W \log(1 + \rho_i). \tag{5}$$

For reasons that will become clear later, the link rate model used in this paper is a simplified version of equation 4:

$$R_i = \log(\rho_i). \tag{6}$$

The constant K is absorbed into G_{ii}, and α is taken as equal to 1, since this constant has the effect of scaling all rates equally. The assumption $\rho_i \geq 1$ is necessary to prevent negative rates, and $\rho_i \gg 1$ makes equation 6 a good approximation of equation 4. Since ρ_i is the effective SIR accounting for the various gains in the system, $\rho_i \gg 1$ will be the case for most real systems in normal operation.

D. Rate-Region

The rate-region is the set of feasible transfer rates $r \in \mathbf{R}_+^n$ for the system. A transfer rate $r \in \mathbf{R}_+^n$ is feasible if it is possible for the system to simultaneously transfer data over the network at the specified rates for some power vector p. By considering all power allocations the set of feasible transfer rates can be found. Analytically the rate-region can be described as

$$\mathcal{R} = \{r \in \mathbf{R}_+^n | r \leq R(p) \text{ for some } p\}, \tag{7}$$

where $r \leq R$ for two vectors means component-wise inequality, i.e. $r_i \leq R_i$ for all i.

The rate-region $\mathcal{R}$ is convex. This is shown by first defining the set of feasible transfer rate and power pairs (r, p), demonstrating its convexity, and projecting it onto the rate transfer space. The set of feasible transfer rate and power pairs, $\mathcal{M}$, is the set of (r, p) such that $r_i \leq \log(\rho_i)$ for all links i. Analytically,

$$\begin{aligned} \mathcal{M} &= \{(r,p) \in \mathbf{R}_+^{2n} | r_i \leq \log(\rho_i), \forall i\} \\ &= \textstyle\bigcap_i \{(r,p) \in \mathbf{R}_+^{2n} | r_i \leq \log(\rho_i)\} \\ &= \textstyle\bigcap_i \mathcal{M}_i. \end{aligned} \tag{8}$$

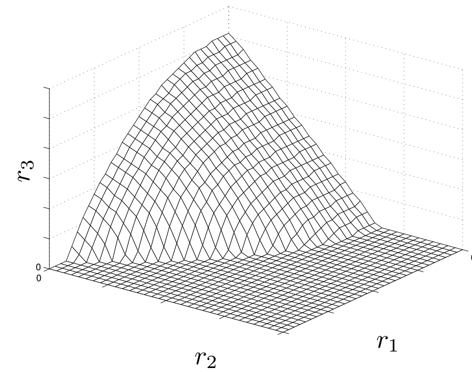

Fig. 2. Rate Region

The $\mathcal{M}_i = \{(r,p) \in \mathbf{R}_+^{2n} | r_i \leq \log(\rho_i)\}$ are convex. This can be seen by the change of variables $x_i = \log p_i$ and rewriting the set qualifier as follows:

$$\begin{aligned} r_i \leq \log(\rho_i) \quad &\Leftrightarrow \quad e^{-r_i} \geq \rho_i^{-1} \\ &\Leftrightarrow \quad e^{-r_i} \geq \textstyle\sum_{j \neq i} G_{ij} e^{x_j} G_{ii}^{-1} e^{-x_i} \\ &\Leftrightarrow \quad 1 \geq \textstyle\sum_{j \neq i} G_{ij} e^{x_j} G_{ii}^{-1} e^{-x_i} e^{r_i} \\ &\Leftrightarrow \quad 0 \geq \log(\textstyle\sum_{j \neq i} G_{ij} e^{x_j} G_{ii}^{-1} e^{-x_i} e^{r_i}). \end{aligned} \tag{9}$$

It is known [1] that the function $\log(\sum \alpha_i e^{y_i})$, for $\alpha_i \in \mathbf{R}_+$ and $y_i \in \mathbf{R}$, is convex in y. Sub-level sets of convex functions always define convex sets, so equation 9 defines a convex set in the variables $\log p_i$ and r_i. Since the intersection of convex sets is convex $\mathcal{M}$ must also be convex.

The rate-region $\mathcal{R}$ is a projection of $\mathcal{M}$ onto the transfer rate space. Since linear projection conserves convexity, the rate-region $\mathcal{R}$ must also be convex. An example of the rate region is shown in Figure 2.

E. Performance Metrics

System performance is modeled by a function U that represents the value of a rate vector r to the system. This function can represent the utility a user derives from using the system at a particular rate or can be implied by data protocols, service rate agreements, or other system level metrics [10], [11], [12]. By assumption, a higher data rate is valued at least as much as a lower data rate, so U is a non-decreasing function of r. Also, by assumption, there is a diminishing return to additional data rate, so U is a concave function of r. For convenience U is termed the system utility function, although it may be composed of functions other than user utility functions.

One example of the system utility is $U(r) = \sum_i U_i(r_i)$ in a single cell site where each user uses a single link and the figure of merit is the sum of the user's utility functions U_i.

A single voice link might have a protocol that requires a minimum r_i but is indifferent to rates above this rate. An appropriate function is the following:

$$U_i(r_i) = \begin{cases} -\infty, & r_i < R_{\min_i} \\ c, & r_i \geq R_{\min_i}. \end{cases} \quad (10)$$

A wireless Internet user might benefit from an increased r_j and be willing to pay more for this service. A possible link metric is then $U_j(r_j) = \alpha r_j$, where implicitly the user pays more for more bandwidth.

QoS constraints can be added to this formulation by embedding them in the U_i. A video user might have a minimum required rate to ensure a minimum level of picture quality, and subjectively value rates above this threshold according to his personal utility function. Assuming a logarithmic utility function

$$U_i(r_i) = \begin{cases} -\infty, & r_i < R_{\min_i} \\ \alpha \log(r_i) + b, & r_i \geq R_{\min_i}. \end{cases} \quad (11)$$

System performance metrics can also be formulated for multi-hop networks.

In what follows U is assumed to have continuous second derivatives.

III. Problem Formulation

The goal is to find the best set of rates r such that the system figure of merit U is maximized. Formally this is expressed as:

$$\begin{array}{ll} \text{maximize}_r & U(r) \\ \text{subject to} & r \in \mathcal{R}. \end{array} \quad (12)$$

This is a convex optimization problem in variable r since the constraint set is convex and the objective maximized is concave. Each $r \in \mathcal{R}$ is associated with one or more power vectors p.

An alternative expression is more revealing and useful in what follows. The rate constraints may be rewritten as follows:

$$\begin{array}{lll} r_i \leq \log(\rho_i) & \Leftrightarrow & p_i \geq \frac{e^{r_i}}{G_{ii}} \sum_{j \neq i} G_{ij} p_j \\ & \Leftrightarrow & p \geq D\tilde{G}p, \end{array} \quad (13)$$

where the inequality is taken to mean element wise, $D \triangleq \text{diag}\left(\frac{e^{r_i}}{G_{ii}}\right)$ and

$$\tilde{G}_{ij} = \begin{cases} G_{ij}, & i \neq j \\ 0, & i = j. \end{cases} \quad (14)$$

Thus the rate-region $\mathcal{R}$ can be rewritten as $\mathcal{R} = \{r \in \mathbf{R}_+^n | p \geq D\tilde{G}p,\ p > 0\}$, and the problem can be expressed as,

$$\begin{array}{ll} \text{maximize}_{p,r} & U(r) \\ \text{subject to} & p \geq D\tilde{G}p \\ & p > 0, \end{array} \quad (15)$$

This formulation can be solved numerically in several different ways, such as interior point methods. In this paper a new approach based on *Perron Frobenius* theory of positive matrices is developed.

IV. Analysis

A. Perron-Frobenius

For a square matrix A the notation $A > 0$ means A is an element-wise positive matrix. The eigenvalue of A with greatest magnitude is the Perron-Frobenius eigenvalue $\lambda_{\text{pf}}(A)$. If the matrix $A > 0$ is regular, meaning that $(A^k)_{ij} > 0$ for all i, j and some positive integer k, then λ_{pf} is strictly positive and unique, and the associated right and left eigenvectors $p > 0$ and $q > 0$ are strictly positive. If $\lambda_{\text{pf}}(A)$ is the Perron-Frobenius eigenvalue for regular $A > 0$, then the inequality $\beta p \geq Ap$ has a feasible $p > 0$ if and only if $\lambda_{\text{pf}}(A) \leq \beta$. Finally, for any positive matrix, the monotone property states if $A_{ij} \leq B_{ij}$ for all i, j with strict inequality for at least one i, j then $\lambda_{\text{pf}}(A) < \lambda_{\text{pf}}(B)$.

Since D is a function of r, in what follows the Perron Frobenius eigenvalue associated with the network is written as $\lambda_{\text{pf}}(D(r)\tilde{G})$.

The Perron Frobenius eigenvalue associated with the network is convex in r. This can be seen by applying the monotone property of positive matrices and the definition of convexity for $0 \leq \alpha \leq 1$,

$$\begin{aligned} D(\alpha r_1 + (1-\alpha) r_2)\tilde{G} &= \text{diag}\left(\frac{e^{\alpha r_1 + (1-\alpha) r_2}}{G_{ii}}\right)\tilde{G} \\ &\leq \text{diag}\left(\frac{\alpha e^{r_1} + (1-\alpha) e^{r_2}}{G_{ii}}\right)\tilde{G} \\ &= \alpha D(r_1)\tilde{G} + (1-\alpha) D(r_2)\tilde{G}, \end{aligned}$$

where the inequality follows from the convexity of e^x. By the monotone property of positive matrices it follows that

$$\begin{aligned} &\lambda_{\text{pf}}(D(\alpha r_1 + (1-\alpha) r_2)\tilde{G}) \\ &\quad \leq \alpha \lambda_{\text{pf}}(D(r_1)\tilde{G}) + (1-\alpha)\lambda_{\text{pf}}(D(r_2)\tilde{G}). \end{aligned} \quad (16)$$

The Perron Frobenius eigenvalue associated with $D(r)\tilde{G}$ is also a differentiable function of r. By assumption the matrix $D\tilde{G}$ is irreducible, so $\lambda_{\text{pf}}(D(r)\tilde{G})$ is a unique eigenvalue. Since it is a unique root of the characteristic polynomial it is continuous in the elements of $D(r)\tilde{G}$. Further, by an application of *Gerschgorin's* Theorem [13] $\lambda_{\text{pf}}(D(r)\tilde{G})$ is differentiable in the elements of $D(r)\tilde{G}$.

The rate-region $\mathcal{R} = \{r \in \mathbf{R}_+^n | p \geq D\tilde{G}p,\ p > 0\}$ can be rewritten as $\mathcal{R} = \{r \in \mathbf{R}_+^n | \lambda_{\text{pf}}(D(r)\tilde{G}) \leq 1\}$, and the network rate optimization problem can then be

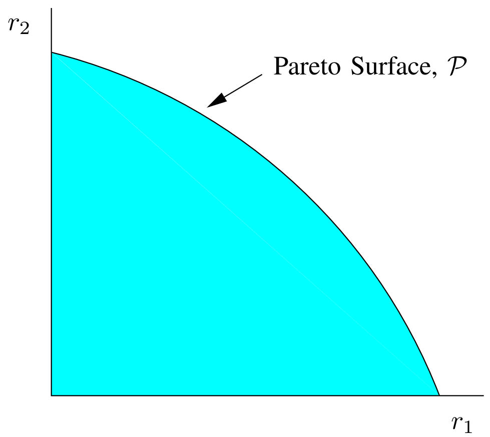

Fig. 3. Rate region, set of achievable rates.

re-expressed as

$$\begin{array}{ll} \text{maximize}_r & U(r) \\ \text{subject to} & \lambda_{\text{pf}}(D(r)\tilde{G}) \leq 1 \\ & r > 0, \end{array} \tag{17}$$

where the objective is concave and the constraints convex.

B. Pareto Surface

A point $r \in \mathcal{R}$ is *Pareto optimal* for the rate-region $\mathcal{R}$ if there does not exist another point $r' \in \mathcal{R}$ that dominates r. A point $r' \in \mathcal{R}$ *dominates* r if $r'_i \geq r_i$ for all i, and $r'_j > r_j$ for some j. The notation $r' \succeq_d r$ means r' dominates r. The Pareto surface is defined as the set of Pareto optimal points,

$$\mathcal{P} = \{r \in \mathcal{R} | \nexists r' \in \mathcal{R} \text{ s.t. } r' \succeq_d r\}. \tag{18}$$

A solution to the utility maximization problem always lies on $\mathcal{P}$ since the system utility function U is non-decreasing in r and concave. To see this, if $r^* \notin \mathcal{P}$ is an optimal set of rates for the utility maximization problem, then there exists a feasible $r' \in \mathcal{P}$ with $r'_i \geq r^*_i \ \ \forall i$. So by the non-decreasing property of the utility function U, $U(r') \geq U(r^*)$. By optimality of r^*, $U(r^*) \geq U(r)$ $\forall$ feasible r, so $U(r') = U(r^*)$ and $r' \in \mathcal{P}$ is also optimal.

The equation $p \geq D\tilde{G}p$ will be met with equality for a rate vector $r^* \in \mathcal{P}$. If the constraint is not tight for a power vector p, then one or more links is operating below the rate its transmitter power could support, and that link rate r_i could be increased without impact on other links. Thus, the inequality must be tight for $r \in \mathcal{P}$ since by definition r is undominated.

Transfer rates r lying in $\mathcal{P}$ can be expressed as the following:

$$\mathcal{P} = \{r | p = D\tilde{G}p\} = \{r | \lambda_{\text{pf}}(D(r)\tilde{G}) = 1\}, \tag{19}$$

where $\lambda_{\text{pf}}(D(r)\tilde{G}) = 1$ corresponds to the surface of the rate-region. Note, $\lambda_{\text{pf}}(D(r)\tilde{G}) > 1$ corresponds to a point outside of the rate region, i.e. that can not be achieved by the system for any set of powers, and $\lambda_{\text{pf}}(D(r)\tilde{G}) < 1$ corresponds to a point inside the rate region, i.e. that can be achieved.

C. Normal to Pareto surface

The outward normal to the Pareto surface at $r \in \mathcal{P}$ is given by $N(r) = \nabla\lambda_{\text{pf}}(D(r)\tilde{G})$, since $\mathcal{P}$ is a level surface of $\lambda_{\text{pf}}(D(r)\tilde{G})$. The gradient of λ_{pf} can be found from matrix perturbation theory [13]. The key result is that the simple eigenvalues of a matrix are differentiable functions of the elements of the matrix. Formally this means

$$\frac{\partial\lambda(A)}{\partial a_{ij}} = q^T \frac{\partial A}{\partial a_{ij}} p. \tag{20}$$

By assumption $\tilde{G}$ is regular and consequently $\lambda_{\text{pf}}(D(r)\tilde{G})$ simple. Thus, the ith component of $\nabla\lambda_{\text{pf}}(D(r)\tilde{G})$ is:

$$\begin{aligned} \frac{\partial\lambda_{\text{pf}}(D(r)\tilde{G})}{\partial r_i} &= q^T(r)\frac{\partial D\tilde{G}}{\partial r_i}p(r) \\ &= q^T\frac{\partial \text{diag}(e^{r_1}/G_{11}, \ldots, e^{r_N}/G_{NN})\tilde{G}}{\partial r_i}p \\ &= q^T\text{diag}(0, \ldots, 0, e^{r_i}/G_{ii}, 0, \ldots, 0)\tilde{G}p \\ &= \sum_{j \neq i} q_i e^{r_i}\frac{G_{ij}p_j}{G_{ii}}. \end{aligned} \tag{21}$$

Further, if r_i is on the Pareto surface $\mathcal{P}$, then $e^{r_i} = e^{R_i} = \rho_i$, so

$$\frac{\partial\lambda_{\text{pf}}(D(r)\tilde{G})}{\partial r_i} = \sum_{j \neq i} q_i\rho_i\frac{G_{ij}p_j}{G_{ii}p_i}p_i \tag{22}$$

$$= q_i p_i. \tag{23}$$

Thus the outward normal is

$$N(r) = \nabla\lambda_{\text{pf}}(D(r)\tilde{G}) = [q_1p_1, q_2p_2, \ldots, q_np_n]^T, \tag{24}$$

which is automatically normalized such that $\mathbf{1}^T N(r) = 1$.

The left eigenvector q has several interesting interpretations. Surprisingly, for $r \in \mathcal{P}$, $\frac{\partial\lambda_{\text{pf}}(D(r)\tilde{G})}{\partial r_i}$ is a function only of q_i and p_i and not the interference power received from other transmitters. The effect of interference is therefore captured in the left eigenvector, which summarizes the interference at each receiver as a single number q_i. The components of the left eigenvector scale the associated link transmitter powers, modeling the effect of interference by reducing the link's rate. The product q_ip_i can be interpreted as a normalized equivalent transmitter power.

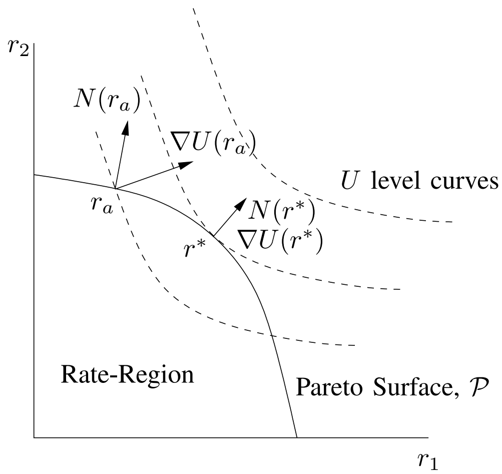

Fig. 4. Optimality condition is $N(r)$ parallel with $\nabla U(r)$.

The $q_i p_i$ can also be interpreted as the marginal prices charged by the system for increases in rate. The constraint $\lambda_{\rm pf} \leq 1$ can be thought of as the "budget constraint" on the network; the network pays for rate increases by spending in $\lambda_{\rm pf}$ dollars. The system is free to spend up-to one unit of $\lambda_{\rm pf}$, but no more. An increase of δr_i results in an increased charge of $q_i p_i \delta r_i$ to the system. This concept can be extended to include factors that reduce the overall capacity of the system and reduce its "budget" from one to a lesser number.

Lastly, the left eigenvector can be interpreted as the power of the adjoint system to the wireless network. In the adjoint system the transmitters and receivers reverse roles, every transmitter becomes a receiver and conversely every receiver becomes a transmitter. So for a system with gain matrix G, the adjoint system has gain matrix G^T. Thus, from $(D\tilde{G})^T q = q\lambda_{\rm pf}$, q is the transmit power vector for the adjoint system.

D. Criteria for Optimality

Because the optimal values of r lie on the Pareto surface $\mathcal{P}$ the network optimization problem can be restated (in its final form) as

$$\begin{array}{ll} \text{maximize}_r & U(r) \\ \text{subject to} & \lambda_{\rm pf}(D(r)\tilde{G}) = 1 \\ & r > 0. \end{array} \tag{25}$$

By *Lagrange's Theorem* $\nabla U = K \nabla \lambda_{\rm pf}$ at optimality, where K is a positive constant of proportionality. In words, the vectors ∇U and $\nabla \lambda_{\rm pf}$ must be parallel at the optimal rates. This is depicted in Figure 4. The gradient ∇U is normal to the level sets of U and $N(r) = \nabla \lambda_{\rm pf}$ is normal to the rate surface. At optimality these normals align and the level set surface for the performance metric is tangent to the rate region.

The constant K is found by noting $\mathbf{1}^T \nabla \lambda_{\rm pf} = \mathbf{1}^T N(r) = \sum_i p_i q_i = 1$, and

$$\begin{array}{rcl} \nabla U &=& K \nabla \lambda_{\rm pf} \\ \mathbf{1}^T \nabla U &=& K \mathbf{1}^T \nabla \lambda_{\rm pf} \\ \mathbf{1}^T \nabla U &=& K, \end{array} \tag{26}$$

thus the constant K of proportionality described in the optimality condition is $K = \mathbf{1}^T \nabla U$. So $r \in \mathcal{P}$ is optimal if and only if $(1/p_i q_i)(\nabla_i U / \mathbf{1}^T \nabla U) = 1$.

At optimum, the marginal performance tradeoffs between rates can also be found from Lagrange's Theorem:

$$\frac{\left(\frac{\partial U(r)}{\partial r_i}\right)}{\left(\frac{\partial U(r)}{\partial r_j}\right)} = \frac{\left(\frac{\partial \lambda_{\rm pf}(D(r)\tilde{G})}{\partial r_i}\right)}{\left(\frac{\partial \lambda_{\rm pf}(D(r)\tilde{G})}{\partial r_j}\right)}. \tag{27}$$

The left-side of the equality is the marginal rate of substitution between rates r_i and r_j for the same overall system performance. The right-side is the ratio of marginal costs for changing these rates. Thus Lagrange's Theorem simply states that at optimality the trade off in performance from a rate change must equal the trade off in costs associated with these new rates.

E. Approximate projection onto the Pareto surface

It will be useful in what follows to formulate methods for finding a rate vector $r_c \in \mathcal{P}$ on the Pareto surface which is close to a given rate vector r_p. A point r_c could be found by projecting r_p onto $\mathcal{P}$; unfortunately, this approach complicates the problem considerably and offers little computational advantage over approximate projections in finding the optimal rates. In this section we describe two simple alternative methods for finding a point $r_c \in \mathcal{P}$ close to a given rate vector r_p by approximate projection.

The two methods are illustrated in Figure 5, and are given in more detail below. The first method moves from a point r_a to the Pareto surface by adding the same fixed offset β_a to each element in the vector. This represents a movement in the $[\ 1 \quad 1 \quad \dots \quad 1\ ]^T$ direction, which for $n = 2$ is simply a 45 degree line . The second method scales each element in the rate vector by the same proportion α_a. This represents a movement on a ray from the origin. Points outside the rate-region can be brought to $\mathcal{P}$ in a similar fashion.

1) Shifting method: The shifting method adds a fixed constant β_p to each element in the rate vector r_p to find a rate vector $r_c \in \mathcal{P}$. Formally $r_c = r_p + \beta_p \mathbf{1}^T$, where $\mathbf{1}^T$ is a vector of ones. Note that for the matrix $D\tilde{G}$ with $D = \text{diag}(e^{r_i} \dots)$, β_p is an additive term in the exponent. Thus β_p scales $D\tilde{G}$ by $\exp(\beta_p)$.

The offset β_p can be found from the Perron Frobenius eigenvalue $\lambda_{\rm pf}(D(r_p)\tilde{G})$ as

$$\beta_p = -\log(\lambda_{\rm pf}(D(r_p)\tilde{G})). \tag{28}$$

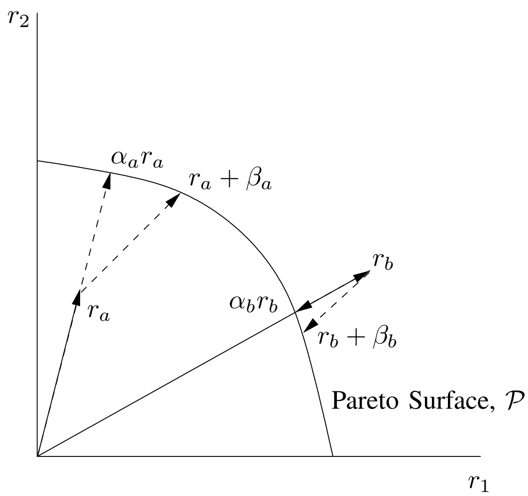

Fig. 5. Moving on to the Pareto surface.

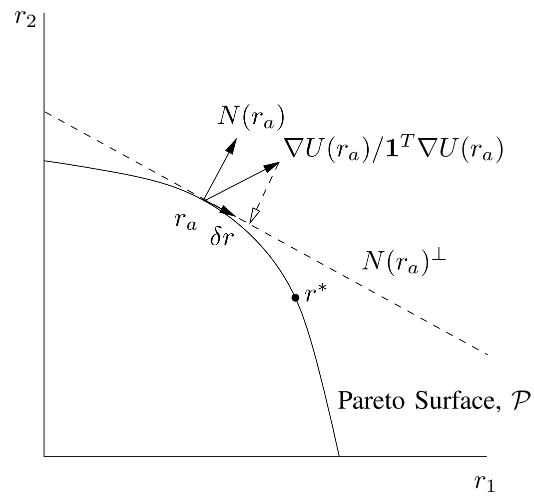

Fig. 6. Finding δr.

At rate r_p, $\lambda_{\text{pf}}(D(r_p)\tilde{G})$ satisfies $D\tilde{G}p = \lambda_{\text{pf}}(D(r_p)\tilde{G})p$, so multiplying this relation by $(\lambda_{\text{pf}}(D(r_p)\tilde{G}))^{-1}$ yields $(\lambda_{\text{pf}}(D(r_p)\tilde{G}))^{-1}D\tilde{G}p = p$.

2) Scaling method: The scaling method multiplies each element in the rate vector r_p by a fixed scalar $r_c = \alpha_p r_p$, $\alpha_p > 0$ to find a rate vector $r_c \in \mathcal{P}$. Increasing α increases all rates $r_c = \alpha r_p$, and in turn increases the elements of $D(r_c)\tilde{G}$. By the monotone property for the Perron Frobenius eigenvalue, $\lambda_{\text{pf}}(D(\alpha r_p)\tilde{G})$ also increases and is monotonic in α. This leads to a bisection algorithm to find α_p.

The bisection algorithm increases α linearly until $\lambda_{\text{pf}}(D(\alpha r_p)\tilde{G}) \geq 1$, so α_p lies between zero and α. Next $\lambda_{\text{pf}}(D(\frac{1}{2}\alpha r_p)\tilde{G})$ is computed and compared with one; if it is greater than one then $\alpha_p \in [0, \alpha/2]$ while if it is less than one then $\alpha_p \in [\alpha/2, \alpha]$. If $\alpha_p \in [\alpha/2, \alpha]$, then $\lambda_{\text{pf}}(D(\frac{3}{4}\alpha)\tilde{G})$ is computed and compared with one to again reduce the range containing α_p by half. The segment that α_p lies in is reduced through repeated bisections until α_p is known to the desired number of decimal points.

F. Feasible ascent direction

It will be useful to have a method for moving toward the optimal rate vector r^* from a given point $r_c \in \mathcal{P}$ on the Pareto surface. Let $N(r)^{\perp} = \left\{r' | (r - r')^T N(r) = 0\right\}$ be the hyper-plane that is tangent to the Pareto surface at r_c. The supporting hyper-plane $N(r)^{\perp}$ is a good approximation of $\mathcal{P}$ for small changes in r. Also, since the rate-region $\mathcal{R}$ is convex, $N(r)^{\perp}$ is a supporting hyper-plane and lies outside of $\mathcal{R}$, except at the point r_c. For this reason a direction δr is defined to be *feasible* if it lies along $N(r)^{\perp}$, or equivalently $\delta r^T N(r) = 0$.

A small change δr is defined as an *ascent* direction if $U(r + \alpha\delta r)$ increases for small $\alpha > 0$. Thus, δr is an ascent direction if and only if $\nabla U(r)^T \delta r > 0$. A point that is both feasible and an ascent direction is termed a *feasible ascent* direction.

Feasible ascent directions δr are not necessarily unique and consequently different approaches can lead to different vectors. Two approaches are presented in this section.

1) Projected Gradient: The projected gradient approach is illustrated in Figure 6. In this approach a normalized version of $\nabla U(r)$ is projected onto $N(r)^{\perp}$. A small step in this direction will approximate a small ascent step on the Pareto surface $\mathcal{P}$ that also brings the two normals, $\nabla U(r)$ and $N(r)$, into closer alignment. More specifically,

$$\begin{aligned} \delta r_{\text{pg}} &= \left(I - \frac{N(r)N(r)^T}{\|N(r)\|^2}\right)\frac{\nabla U}{\mathbf{1}^T\nabla U} \\ &= \left(I - \frac{[p_1q_1,\ldots,p_nq_n][p_1q_1,\ldots,p_nq_n]^T}{\sum_{i=1}^{n}(p_iq_i)^2}\right)\frac{\nabla U}{\mathbf{1}^T\nabla U}, \end{aligned} \quad (29)$$

where p and q are the right and left eigenvectors respectively of $D(r)\tilde{G}$ associated with the Perron Frobenius eigenvalue $\lambda_{\text{pf}}(D(r)\tilde{G})$. The term $\left(I - \frac{N(r)N(r)^T}{\|N(r)\|^2}\right)$ is a projection matrix that projects the scaled normal $\frac{\nabla U}{\mathbf{1}^T\nabla U}$ onto the supporting hyper-plane at the current rate vector r. Lagrange's condition of optimality is met when $\delta r = 0$.

The projected gradient can best be understood by construction. The projected gradient computes a component-wise error metric from the optimality condition and then projects this error onto the supporting hyper-plane. The error estimate is defined as

$$e = \left(N(r) - \frac{\nabla U}{\mathbf{1}^T\nabla U}\right). \quad (30)$$

Because U is concave and λ_{pf} is convex, a rate change δr_i causes the performance metric normal and rate region normal to respond in opposite ways; an increase $\delta r_{\mathrm{pg},i} > 0$ causes the ith component of $\frac{\nabla U}{\mathbf{1}^T \nabla U}$ to decrease and the comparable component of $\nabla \lambda_{\mathrm{pf}}$ to increase. Consequently, the decision to increase the ith component of δr_{pg} can be made by comparing the two normals. If $\frac{\nabla U}{\mathbf{1}^T \nabla U}$ is greater than $\nabla \lambda_{\mathrm{pf}}$ for component i, then the associated rate should be increased.

The change in rates $\delta\theta$ is then initially estimated to be

$$\delta\theta = -\left(\nabla\lambda_{\mathrm{pf}} - \frac{\nabla U}{\mathbf{1}^T \nabla U}\right). \tag{31}$$

The new rate estimate $r = \delta\theta + r_c$ is not necessarily on the rate surface or in the feasible set of rates. This direction $\delta\theta$ is made a feasible direction for δr by projecting it onto the supporting hyper-plane $N(r)^{\perp}$. This gives δr_{pg} shown in equation 29.

2) Direct Step: The direct step approach directly constructs a δr on the supporting hyper-plane $N(r)^{\perp}$, rather than projecting $\frac{\nabla U}{\mathbf{1}^T \nabla U}$ onto this plane as was done in the projected gradient method. As mentioned previously, equation 30 can be used to determine whether a component of r should be decreased or increased to improve the value of $U(r)$ and reduce the angle between $N(r)$ and $\nabla U(r)$. Specifically, δr_{ds} should have the same sign as $-e$, for small rate adjustments. The direct step method uses this information to find a δr_{ds} that is an ascent direction but which is also feasible by construction, that is lies on $N(r)^{\perp}$. Specifically,

$$\delta r_{\mathrm{ds}} = -\,\mathbf{diag}(\frac{1}{q_i p_i}, \ldots)e. \tag{32}$$

Substituting e and simplifying gives

$$\delta r_{\mathrm{ds,i}} = -\left(1 - \left(\frac{1}{p_i q_i}\right)\left(\frac{\nabla_i U}{\mathbf{1}^T \nabla U}\right)\right), \tag{33}$$

where $p_i q_i$ is the ith component of $N(r)$.

That δr_{ds} lies on the supporting hyper-plane can be seen from

$$\begin{aligned} N(r)^T \delta r_{\mathrm{ds}} &= \sum q_i p_i \left(\tfrac{e_i}{q_i p_i}\right) \\ &= \sum e_i \\ &= \mathbf{1}^T e \\ &= \mathbf{1}^T N(r) - \mathbf{1}^T \nabla U / \mathbf{1}^T \nabla U \\ &= 0 \end{aligned} \tag{34}$$

where $\mathbf{1}^T N(r) = 1$.

V. Method of Solution

A general method of solution can be described as the following algorithm. Given $r_c \in \mathcal{P}$ and $\alpha > 0$ iterate the following three steps until $\delta r = \mathbf{0}$:

1) Compute a feasible ascent direction δr at r_c.
2) Compute $r_p = r_c + \alpha\delta r$.
3) Approximately project r_p onto $\mathcal{P}$.

This general method describes several specific schemes. For step 1 either the projected gradient δr_{pg} or direct step δr_{ds} directions may be used for δr. For step 3 either the shifting or scaling method may be used to approximately project r_p onto $\mathcal{P}$. Mixing and matching these methods gives four algorithms.

VI. Simulation

This section describes a simple simulation. The model is of a 5 link single hop network. The performance metric corresponds to the sum of individual user utility functions and is given by

$$U(r) = \sum_{i=1}^{5} a_i \log(r_i) + \Delta \log(r_i - r_{\mathrm{th},i}), \tag{35}$$

where the first term in the sum is the utility associated with a given rate r_i and the second term acts as a barrier limiting this rate to $r_i > r_{\mathrm{th},i}$. For the utility function, the natural log is used. The a_i are scale constants associated with different users, and the constant $\Delta \ll a_i$ $\forall i$. The barrier portion of the individual utility functions is negligible for $r_i > r_{\mathrm{th},i}$ but dominates for $r_i \sim r_{\mathrm{th},i}$, preventing the link rate from dropping below the threshold.

The solution algorithm uses the direct step approach for finding a feasible ascent direction and the scaling method for approximately projecting r_p onto $\mathcal{P}$. The step size parameter is arbitrarily set to a small positive number, $\alpha = 0.001$.

For the simulation, the gain matrix is

$$G = \begin{bmatrix} 144.1 & 0.217 & 0.311 & 0.068 & 0.617 \\ 0.469 & 83.0 & 0.307 & 0.125 & 0.269 \\ 0.537 & 0.053 & 120.5 & 0.166 & 0.221 \\ 0.563 & 0.229 & 0.954 & 144.3 & 0.713 \\ 0.511 & 0.167 & 0.131 & 0.136 & 108.2 \end{bmatrix}, \tag{36}$$

and $\Delta = 0.0001$. Three events are progressively simulated, each 4000 time periods long. At time 0, the initial utility coefficient vector is $a = [\ 0.1 \quad 1 \quad 3 \quad 5 \quad 7\ ]^T$, the initial rate threshold vector is $r_{\mathrm{th}} = [\ 5 \quad 0 \quad 0 \quad 0 \quad 0\ ]^T$, and the initial power vector is $p = [\ 1 \quad 0.01 \quad 0.01 \quad 0.01 \quad 0.01\ ]^T$. During this initial period the network converges to an optimal set of rates subject to a rate constraint on link 1. At time 4000 the second event occurs, link 1's rate-threshold $r_{\mathrm{th},1}$ changes to zero, $r_{\mathrm{th},1} = 0$, removing the rate constraint. All of the other parameters remain unchanged. At time 8000 the third event begins, link 1's utility coefficient a_1 is changed to 7, $a_1 = 7$, while all the other parameters remained the same.

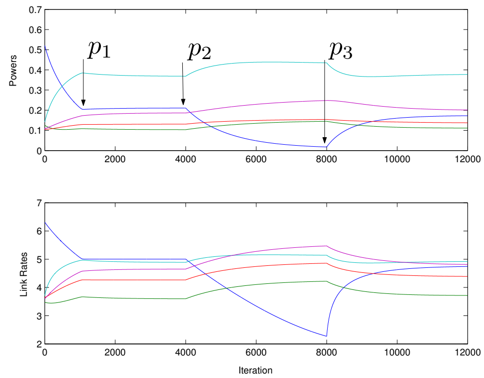

Fig. 7. Power and Link Rates

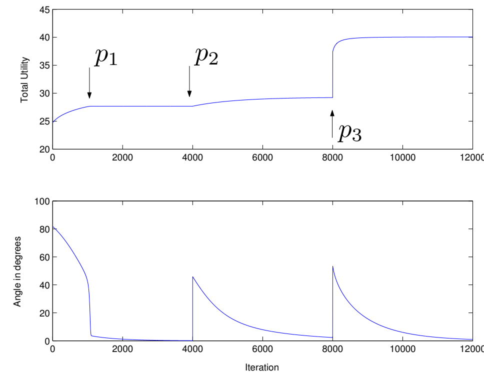

Fig. 8. Total Utility and Error Angle

The simulation results, shown in Figures 7 and 8, show the initial convergence and the two events. In Figure 7 the transmitter powers and link rates for each of the five links is shown. In Figure 8 the associated total utility and optimality measure is depicted. At optimum the angle between the normal to the rate surface and the normal to the iso-utility curves will be zero. Any angle greater than this represents a residual error.

Starting at time zero, the network evolves to a steady state in which the transfer rate of link 1 is constrained to be greater than $r_{\text{th},1} = 5$. As shown in Figure 7, the transmitter powers and associated link rates of the constrained link initially decrease, but upon reaching the point marked p_1 stabilize at values corresponding to $r_1 = 5$. As can also be seen, total utility changes little past point p_1, while the optimality measure quickly converges to zero.

At time 4000 (point p_2), when the rate constraint is removed $r_{\text{th},1} = 0$, the system adapts by further decreasing r_1, and thereby increasing overall utility. The angle between the Pareto surface normal and the utility surface normal jumps to approximately 45 degrees when the constraint is removed, and then exponentially declines to zero. The overall utility changes little, however, since the link 1 utility coefficient $a_1 = 0.1$ is much smaller than the other users, and thus the system was operating near the optimal point before the constraint was removed. At time 8000 (point p_3), when user 1's utility function is changed by increasing $a_1 = 7$, the system responds by finding a new equilibrium point at greater aggregate utility.

VII. Summary

This paper presents a new approach to finding optimal network link rates and transmitter powers. The approach is layered, with a rate optimization algorithm selecting optimal rates from the surface of the rate-region and a power optimization algorithm calculating the associated optimal powers. The approach is adaptive and can respond to changes in the demands on the network.

The analytical rate adaptation approach is based on Perron Frobenius theory and describes the network in terms of λ_{pf}, the transmitter powers p, and the transmitter powers q of the adjoint network. The Perron Frobenius eigenvalue is interpreted as a convex cost function for the network, and the transmitter powers as Perron Frobenius eigenvectors. The left eigenvector q is shown to fully capture the effect of interference on transmitter rates.

References

[1] S. Boyd and L. Vandenberghe. (2003) Convex optimization. [Online]. Available: http://www.stanford.edu/ boyd/cvxbook.html

[2] R. Duffin, E. Peterson, and C. Zener, *Geometric Programming - Theory and Practice*. New York, NY: John Wiley and Sons, 1967.

[3] D. Julian, M. Chiang, and D. ONeill, "Qos and fairness constrained convex optimization of resource allocation for wireless cellular and ad-hoc networks," in *Proc. IEEE Infocom*, June 2002, pp. 477–486.

[4] S. Lim, Z. Rosberg, and J. Zander, "Combined power and transmission rate selection in cellular networks," in *Proceedings of Vehicular Technology Conference*, Nov. 1999, pp. 1653–1656.

[5] C. Sung and W. Wong, "Power control and rate management for wireless multimedia cdma systems," *IEEE Trans. Comm.*, vol. 49(7), pp. 1215–1226, July 2001.

[6] D. ONeill, "Adaptive congestion control for wireless networks using tcp," in *Proc. ICC 2003*, May 2003.

[7] G. Foschini and Z. Miljanic, "A simple distributed autonomour power control algorithm and its convergence," *IEEE Trans. Vehicular Technology*, vol. 42(4), pp. 641–646, Nov. 1993.

[8] N. Bambos, S. Chen, and G. Pottie, "Channel access algorithms with active link protection for wireless communication networks with power control," *IEEE/ACM Trans. on Networking*, vol. 8(5), pp. 583–597, Oct. 2000.

[9] A. Goldsmith, *Wireless Communications*. Stanford, CA: Stanford University EE 359 Course Reader, 2001.

[10] S. Low, L. Peterson, and L. Wang, "Understanding vegas: A duality model," *Journal of ACM*, vol. 49(2), pp. 207–235, Mar. 2002.

[11] F. Kelly, A. Maulloo, and D. Tan, "Rate control in communications networks: Shadow prices, proportional fairness and stability," *Journal of Operations Research Society*, vol. 49, pp. 237–252, Nov. 1998.

[12] S. Low. (2002) Duality model of tcp and queue management algorithms. [Online]. Available: http://netlab.caltech.edu
[13] G. Stewart and J. Sun., *Matrix perturbation theory*. Boston, Mass.: Academic Press, 1990.