Distributed Rate Assignments for Simultaneous Interference and Congestion Control in CDMA-Based Wireless Networks

Jennifer Price and Tara Javidi

Department of Electrical and Computer Engineering

University of California, San Diego

jenn-price@ucsd.edu, tara@ece.ucsd.edu

Abstract

This paper considers the optimal, distributed assignment of rates to elastic users in a wideband multi-cell CDMA network with arbitrary but known layout and variable rate assignments, connected to a traditional wired IP network. We show that by using an optimization framework, it is possible to construct a modular implementation of a distributed rate assignment algorithm at the MAC and transport layers which regulates rates separately based on the congestion and interference constraints. These rates are coordinated through the addition of an intermediate queue at each wireless source (i.e. MAC queue) and represent a cross-layer optimal solution. In addition, we show that this algorithm converges to the optimal rate assignment under static network conditions, and conduct simulation studies comparing the dynamic behavior of this algorithm with that of a “one-shot” approach in which the MAC and transport layer functionalities are merged.

I. INTRODUCTION

In data-only (DO) systems, most traffic is elastic; that is, it can tolerate variable transmission rates and delay [1]. Hence, a successful design is one that responds to fluctuations in demand. Similarly, at the MAC layer of wireless networks we would like to develop resource allocation schemes that can respond to randomly fluctuating channel conditions by adapting user transmission rates at the sources. As a result, in a wireless network each user’s rate must be regulated by both the transport layer (in response to the congestion status of the links/buffers in the network) and the MAC layer (in response to the interference levels and channel quality of the wireless medium). Traditionally, transport and MAC layer protocols are implemented separately, as shown in Figure 1(a). Unfortunately, interactions between transport and MAC layer protocols
can significantly degrade performance in terms of both throughput and delay [2], [3]. Because of this, cross-layer designs have begun to gain attention. Figure 1(b) shows an example of cross-layer design where each wireless user is assigned a single rate which manages both congestion and interference, merging the functionality of the transport and MAC layers (hereafter referred to as a one-shot approach) even though it is desirable to implement transport and MAC layers in separate modules (hereafter referred to as a modular approach).

In this paper, we introduce a modular implementation of a distributed rate assignment protocol which follows the structure of Figure 1(a). We show that it is possible not only for such a modular implementation to achieve a cross-layer optimal resource allocation, but that this can be done with minimal modification of the TCP-type transport layer for data.

The rest of this paper is organized as follows. Section II contains background information including a discussion of related works, network setting, and the interference model used. Section III contains a description of the modular rate assignment protocol. Section IV discusses issues related to practical implementation, and Section V provides simulation results. Finally, Section VI presents our conclusions and areas of future work.

II. BACKGROUND MATERIAL

A. Previous and Related Work

One of the key motivating factors in developing the rate-assignment algorithms described throughout this paper is the idea of having users respond to randomly fluctuating conditions by continuously adapting rates for efficient channel utilization while maintaining fairness. The concept of adaptation to changing network conditions draws heavily from several well-known
works on congestion control in wired networks (see [1], [4], and [5]). In these works, the network is modeled as a collection of individual users and network components who are in constant negotiation with one another. When carefully designed, the interactions/negotiations between users and network components (such as link buffers, etc) result in a distributed network-wide optimal solution. In summary, it is possible to design distributed resource allocation algorithms in the form of feedback-based dynamic response to network congestion.

Recently, this paradigm of adaptive distributed resource allocation has been extended to wireless networks in a number of different contexts and with different focus. At the MAC layer, [6] and [7] develop distributed power control algorithms, while we have developed distributed rate assignment algorithms in [8] - [10]. Both sets of works focus on resource allocation at the MAC layer assuming infinite source (full buffers) at the MAC.

To account for the impact of higher layers, the authors in [11], [12], [13], [14] and [15] address cross-layer design for ad-hoc networks. Specifically, [11], [14] and [15] address design across the transport and MAC layers for a generic wireless network, [13] addresses design across the MAC and physical layers for a TDMA and a joint TDMA/CDMA network, and [12] addresses design across the transport and physical layers for a generic wireless network. Our paper, in spirit, is similar this genre of work. In particular, it is most closely related to [12] (in its construction and decoupling of the optimization problem) and [14] and [15] (in their use of intermediate queues as a mechanism to coordinate different layers). We examine these works more carefully below.

In [12], the authors address joint rate and power control for ad-hoc wireless networks. Although the problem formulations are similar, the main difference between our work and [12] is the treatment of the wireless constraints. The authors in [12] treat wireless connections as links with variable capacity using information theoretic capacity. The structure in this model allows the authors to decompose the problem into separate rate and power assignments using message passing. In realistic systems, however, an information theoretic capacity may not be practical - particularly in the context of delay. Instead, we work directly with a constraint on $\frac{E_b}{N_0}$ (or $SINR$).
This is similar to the idea of delay-limited capacity in [16], and is more practical. As seen later, the constraint on $\frac{E_b}{N_0}$ has a nonlinear form, making the decomposition used in [12] ineffective in our context.

In addition, our work is related to the cross-layer design in [14] and [15]. In these works, the authors address joint scheduling and rate control in wireless multi-hop networks. The idea of introducing queues at the wireless link (which we refer to as MAC layer queues) in these papers is similar to our work. The introduction of these MAC queues allows for the overlay of two sets of rate control problems: 1) at the transport layer, and 2) at the wireless link. The main difference between [14], [15] and our work is in the formulation of the latter problem. In [14] this problem is treated as a throughput-optimal service rate scheduling, and in [15] it is addressed as a time scheduling rate-control problem. Solutions of this kind may hinder the development of distributed solutions at the wireless link, as they assume a central or offline controller at the MAC. In contrast, we restrict our attention to a CDMA-based MAC, where we can use dynamic spreading gain variations and bases’ signaling to provide distributed rate control at the wireless link. The extensions of our work to a packet-based system (similar to [14]) or to an ad-hoc setting (similar to [15]) remain open.

**B. Network Setting and CDMA Interference Model**

We focus on a wideband CDMA network with variable transmission rates, similar to the CDMA2000 1xEVDO system, which uses a distributed feedback mechanism for reverse-link MAC rate assignment. Mobiles communicate directly with the base stations, which are connected directly to the wired IP network, as shown in Figure 2. In order to describe the operation of wireless and wired links in greater detail, we use the following notation.

There are a total of $M$ sources transmitting with transport-layer rate $x_i$. Without loss of generality, the set of all sources can be ordered as $\{1, \ldots, N, N + 1, \ldots, M\}$, where the first $N$ are wireless sources. Each wireless source transmits over the air with a MAC-layer rate $\alpha_i$.

There is a set $\mathcal{J} = \{1, 2, \ldots, J\}$ of wired links, each with capacity $C_j$. The set of wired links
used by source $i$ is fixed, and denoted by $l_i$. The routing function is defined as

$$
\psi_{ij} = \begin{cases} 
1 & \text{if } j \in l_i \\
0 & \text{if } j \notin l_i
\end{cases}
$$

There is a set $L$ of $L$ CDMA-based wireless sectors associated with wireless access points (bases), and each sector has an associated antenna gain. The tracking sector for wireless source $i$, denoted $b(i)$, is the sector to which wireless source $i$ is connected. This is also the sector responsible for wireless source $i$’s power control. For simplicity, we assume that each wireless source is tracked by exactly one sector. $P_i$ is the transmitted power for user $i$, and $g_{il}$ is the channel gain (assumed to be fixed) from user $i$ to sector $l$. $W$ is the chip bandwidth, and $N_0$ is the thermal noise density. The spreading gain for mobile $i$ is defined as $s_i = \frac{W}{\alpha_i}$.

Consider wireless source $i$ which is tracked by sector $l = b(i)$. The ratio of transmit energy per chip to interference power of mobile $i$ at sector $l$ can be written as

$$
\frac{E_b}{N_t}(i) = \frac{W^2 P_i g_{il}}{N_0 W + \sum_{k=1,k \neq i}^N P_k g_{kl}}
$$

(1)

The signal-to-noise ratio of mobile $i$’s data signal at sector $l$ can then be written as $SINR^l(i) = \frac{E_b}{N_t}(i)(\frac{\alpha_i^l}{W})$.

Notice that in an interference limited system such as CDMA, the relationship between SINR and information rate given by the Shannon equation can be approximated as a linear one (i.e. $\log(1 + y) \approx y$ when $y \ll 1$) [17]. This means that an increase in MAC rate $\alpha_i$ translates directly into a linear increase in information rate if and only if $\frac{E_b}{N_t}$ is kept the same (e.g. at $\gamma$).
In other words, we assume the condition $\frac{E_b}{N_t} = \gamma$ is a necessary condition for decodability of transmissions of information at a rate proportional to $\alpha_i$ [18].

C. Cross-Layer Optimal Rate Assignments

In order to address rate control as a constrained optimization problem, we must first introduce the notion of feasible rate assignments. In a wired network, a feasible rate assignment is typically one in which sum rate of all users transmitting across a given link is less than the capacity of the link. In a wireless network, however, the definition of what constitutes a feasible rate-power pair is highly dependent upon both the application and design of a particular system. Constraints may include minimum or maximum power constraints, interference limits, minimum rate guarantees, QoS metrics, or maximum delay requirements. Since the focus of this work is a DO network, we choose to focus on a commonly used definition of feasible MAC rates which depends on both a target $\frac{E_b}{N_t}$ (denoted by $\gamma$), and a target interference level (denoted by $K$). A more detailed explanation of these feasibility criteria can be found in the 3GPP2 standards for CDMA2000 [19].

An important (and often neglected) issue in high data rate CDMA networks is the performance degradation due to multi-path interference when low spreading gains are used [20], [21]. While the standard Gaussian approximation used for performance analysis of a matched filter receiver is valid for high spreading gain, it becomes less and less valid as the spreading gain decreases. As such, we restrict our attention to reverse-link transmission rates that satisfy $\alpha_i \leq \frac{W}{4}$. This results in a spreading gain which the authors in [20] have shown to exhibit only moderate performance degradation due to multi-path interference.

With these issues in mind, we say a pair of rate vectors $(x, \alpha) = (x_1, \ldots, x_M), (\alpha_1, \ldots, \alpha_N)$ is feasible if there exists a non-negative power vector $(P_1, \ldots, P_N)$ such that the following conditions are satisfied:

A. $\sum_{i=1}^{M} x_i \psi_{ij} \leq C_j \quad \forall j \in J$

B. $\sum_{i=1}^{N} P_i g_{il} \leq K N_0 W \quad \forall l \in L$
C. \( \frac{E_b}{N_t}(i) \geq \gamma \quad \forall i \leq N \) and \( l = b(i) \)

D. \( x_i = \alpha_i \quad \forall i \leq N \)

E. \( 0 \leq \alpha_i \leq \frac{W}{4} \quad \forall i \leq N \)

where \( \gamma \) and \( K \) are pre-defined constants. The first condition is simply the link capacity constraint for the wired network [4], and depends on the routing matrix \( \psi \). The second and third conditions specify the target \( \frac{E_b}{N_t} \) and interference levels, respectively. Note that these two conditions together guarantee an acceptable BER for wireless transmissions [19]. The fourth condition guarantees the stability of MAC layer queues, and the last condition ensures the validity of the standard Gaussian approximation.

In [8] and [9], we have developed a simpler feasibility region with a linear-type structure. Using similar modifications, we formally define our feasibility region.

**Definition 1:** A pair of rate vectors \((x, \alpha) = (x_1, \ldots, x_M), (\alpha_1, \ldots, \alpha_N)\) belongs to the feasible region \( \Delta \) if and only if it satisfies the following conditions:

C1. \( \sum_{i=1}^{M} x_i \psi_{ij} \leq C_j \quad \forall j \in J \)

C2. \( \sum_{i=1}^{N} \frac{\alpha_i W + \gamma \alpha_i g_{il}(i)}{(1+K)} \leq \frac{K}{\gamma(1+K)} \quad \forall l \in L \)

C3. \( x_i = \alpha_i \quad \forall i \leq N \)

C4. \( 0 \leq \alpha_i \leq \frac{W}{4} \quad \forall i \leq N \)

In Theorem 3 in [8], we have shown that if \( \alpha \in \Delta \) then it satisfies Conditions A-D.

Among all feasible rate vectors, we wish to choose a rate vector that is proportional fair [1]. In other words, we choose to optimize the utility function \( \sum_{i=1}^{M} \log(x_i) \), resulting in the following optimization problem:

P. Find the pair of rate vectors \((x, \alpha)\) that is the solution to:

\[
(x^*, \alpha^*) = \arg \max_{(x, \alpha) \in \Delta} \sum_{i=1}^{M} \log(x_i)
\]

In the remainder of this paper, we see how optimization and dual methods can be used to develop a distributed solution to Problem P.
III. DISTRIBUTED ALGORITHM FOR MODULAR RATE ASSIGNMENTS

The goal of this section is to design distributed algorithms that implement transport and MAC layer protocols in separate modules following the structure of Figure 1(a) while still converging to the cross-layer optimal solution to Problem P. The challenge in directly applying dual methods to Problem P is twofold: the objective function in Problem P is not a concave function of all primal variables (i.e. it does not depend on the value of \( \alpha \)), and Problem P is not a convex optimization problem.

In order to remedy the first challenge, we rewrite the utility of wireless user \( i \) as

\[
\log(x_i) = (1 - \sigma) \log(x_i) + \sigma \log(x_i) = (1 - \sigma) \log(x_i) + \sigma \log(\alpha_i)
\]

where \( 0 < \sigma < 1 \) is a constant. To simplify notations, we define

\[
V_i(x_i) = \begin{cases} 
  \log(x_i) & \text{if } i > N \\
  (1 - \sigma) \log(x_i) & \text{if } i \leq N
\end{cases}
\]

In order to remedy the second challenge, we introduce the change of variable \( r_i = \frac{\alpha_i}{\alpha_i \gamma + W} \).

This gives the following problem:

Problem P1

\[
\max_{x, r} \sum_{i=1}^{N} \sigma \log\left( \frac{r_i W}{1 - \gamma r_i} \right) + \sum_{i=1}^{M} V_i(x_i)
\]

s.t.

\[
\sum_{i=1}^{M} x_i \psi_{ij} \leq C_j \quad \forall j \in J
\]

\[
\sum_{i=1}^{N} r_i g_{li} \leq \frac{K}{\gamma(1 + K)} \quad \forall l \in L
\]

\[
x_i = \frac{r_i W}{1 - \gamma r_i} \quad \forall i \leq N
\]

\[
0 \leq r_i \leq \frac{W}{\frac{W}{4} \gamma + W} \quad \forall i \leq N
\]

Note that because of Condition C3, the solution to Problem P1 coincides with the solution to Problem P. Furthermore, we show in Appendix I that there is no duality gap for Problem P1. Thus, we consider the Lagrangian associated with Problem P1:
The dual problem can be formulated as follows:

DP1. Find the Lagrange multipliers \((\lambda_1, \ldots, \lambda_J), (\mu_1, \ldots, \mu_L), (\nu_1^+, \ldots, \nu_N^+)\) and \((\nu_1^-, \ldots, \nu_N^-)\) such that they solve

\[
\min_{\lambda, \mu, \nu \geq 0} \sum_{i=1}^{M} \phi_i(q_i, \nu_i) + \sum_{i=1}^{N} \rho_i(p_i, \nu_i) + \sum_{j=1}^{J} \lambda_j C_j + \frac{K}{\gamma(1+K)} \sum_{l=1}^{L} \mu_l
\]

where

\[
q_i = \sum_{j=1}^{J} \lambda_j \psi_{ij}
\]

\[
p_i = \begin{cases} 
\sum_{l=1}^{L} \frac{g_{il}}{g_{ib(i)}} \mu_l & i = 1, \ldots, N \\
0 & i = N+1, \ldots, M 
\end{cases}
\]

\[
\nu_i = \begin{cases} 
\nu_i^+ - \nu_i^- & \text{if } i \leq N \\
0 & \text{if } i > N 
\end{cases}
\]

and for \(i = 1, \ldots, M:\)

\[
\phi_i(q_i, \nu_i) = \max_x \left( V_i(x) - x(q_i + \nu_i) \right)
\]  

(2)

and for \(i = 1, \ldots, N:\)

\[
\rho_i(p_i, \nu_i) = \max_r \left( (1 - \sigma) \log \left( \frac{r_i W}{1 - \gamma r_i} \right) + \frac{r_i W}{1 - \gamma r_i} \nu_i - r_i p_i \right)
\]  

(3)

Notice that for a given set of Lagrange multipliers, Eqns (2) and (3) are autonomous rules that can be implemented at each source using locally available information. We also note the presence of a “cross-layer coordination signal”, \(\nu_i\), which is used to coordinate each user’s two separate rate adjustments: one at the transport layer \((x_i)\), and one at the MAC layer \((r_i\) or \(\alpha_i)\). This is an extremely attractive property since it allows for distributed computation of the rate assignments: when the multipliers are chosen appropriately, the autonomous rules given by Eqns (2) and (3) result in a globally optimal, proportional fair solution. Furthermore, if we continually update the multipliers, rates can be adjusted to changes in the network via a set of distributed
feedback loops as shown in Figure 3. A gradient projection method is used to generate the Lagrange multipliers.

Substituting $\alpha_i$ back in, we get the following algorithm consisting of five parts:

**Base Algorithm**

Each base station produces a regulatory signal (Lagrangian multiplier $\mu_l$) that indicates the level of interference at that sector. This signal evolves according to the following difference equation:

$$
\Delta \mu_l = \begin{cases}
\beta \left( \sum_{i=1}^{N} \frac{q_i}{g_i} \frac{g_i}{W+\gamma\alpha_i} - \frac{K}{\gamma(1+K)} \right) & \text{if } \mu_l(t) > 0 \\
\beta \left[ \sum_{i=1}^{N} \frac{q_i}{g_i} \frac{g_i}{W+\gamma\alpha_i} - \frac{K}{\gamma(1+K)} \right] & \text{if } \mu_l(t) = 0
\end{cases}
$$

where $\beta$ is a constant and $\sum_{i=1}^{N} \frac{q_i}{g_i} \frac{g_i}{W+\gamma\alpha_i}$ is a measure of interference at each sector. These signals are then used to generate the aggregate signals $p$.

**Link Algorithm**

Each link produces a regulatory signal (Lagrangian multiplier $\lambda_j$) that indicates the level of congestion at that link. This signal evolves according to the following difference equation:

$$
\Delta \lambda_j = \begin{cases}
\xi \left( \sum_{i=1}^{M} x_i \psi_{ij} - C_j \right) & \text{if } \lambda_j(t) > 0 \\
\xi \left[ \sum_{i=1}^{M} x_i \psi_{ij} - C_j \right] & \text{if } \lambda_j(t) = 0
\end{cases}
$$

where $\xi$ is a constant and $\sum_{i=1}^{M} x_i \psi_{ij}$ is the total traffic on link $j$. These signals are then used to generate the aggregate signals $q$.

**Wireless Algorithm**
Each wireless source produces two internal coordination signals (Lagrangian multipliers $\nu^+_i$ and $\nu^-_i$) that indicate the difference between transport and MAC layer rates at that source. These signals evolve according to the following difference equations:

$$\Delta \nu^+_i = \begin{cases} 
\zeta_1(x_i - \alpha_i) & \text{if } \nu^+_i(t) > 0 \\
\zeta_1[x_i - \alpha_i]^+ & \text{if } \nu^+_i(t) = 0
\end{cases} \quad (6)$$

and

$$\Delta \nu^-_i = \begin{cases} 
\zeta_2(\alpha_i - x_i) & \text{if } \nu^-_i(t) > 0 \\
\zeta_2[\alpha_i - x_i]^+ & \text{if } \nu^-_i(t) = 0
\end{cases} \quad (7)$$

where $\zeta_1$ and $\zeta_2$ are scalars. Note that these signals are generated only at the wireless sources, and are then used to generate the aggregate signals $\nu$.

**Transport Layer Source Algorithm**

Each source reacts to the levels of congestion (indicated by the link coordination signals) and the mismatch between transport and MAC layer rates (indicated by the cross-layer coordination signals) by adjusting its transport-layer rate such that

$$x_i = \arg \max_x (V_i(x) - x(q_i + \nu_i)) \quad (8)$$

This algorithm is run at both wired and wireless sources.

**MAC Layer Source Algorithm**

Each wireless source reacts to the interference levels at its neighboring sectors (indicated by the base coordination signals) and the mismatch between transport and MAC layer rates (indicated by the cross-layer coordination signals) by adjusting its MAC-layer rate such that

$$\alpha_i = \arg \max_\alpha (\sigma \log(\alpha) + \alpha \nu_i - \frac{\alpha}{W + \gamma \alpha} p_i) \quad (9)$$

This algorithm is run only at the output of wireless sources.

Now, we introduce the following convergence theorem, whose proof can be found in Appendix II.
**Theorem 1:** There exist values $\beta_0$, $\xi_0^0$, $\zeta_1^0$ and $\zeta_2^0$ such that for all $\beta < \beta_0$ and for all $\xi < \xi_0^0$ and for all $\zeta_1 < \zeta_1^0$ and for all $\zeta_2 < \zeta_2^0$, the modular distributed algorithm described by Eqns (4)-(9) converges to the solution to Problem P.

IV. **Practical Distributed Implementation**

In this section, we address the practical implementation of our proposed algorithm. In other words, we distinguish between the above described parallel implementations, and truly distributed rate control in which locally available observations are used. Here, we note that the Lagrange multipliers need not be locally available, even though they can always be computed in parallel via (in general complicated) message passing schemes. Throughout this section, we substantiate our claim that our proposed algorithms can be implemented in a distributed manner with reasonable overhead.

**A. Signaling Mechanisms**

The computation and communication of regulating signals is the basis of the distributed algorithm described in the previous section. In particular, we are interested in addressing 1) the computation of the regulating signals $\mu$ and the availability of the corresponding aggregate signals $p$, 2) the computation of the regulating signals $\lambda$ and the availability of the corresponding aggregate signals $q$, and 3) the computation of the regulating signals $\nu^+$ and $\nu^-$.  

Recall the base algorithm from Eqn (4). This equation requires each base to know information about the load at all other bases. In order to facilitate distributed computation, we introduce the following alternative which *approximates* the original solution:

$$
\Delta \mu_l \simeq \begin{cases} 
\beta \left( \sum_{i=1}^{N} \frac{P_i g_{il}}{N_0 W} - K \right) & \text{if } \mu_l(t) > 0 \\
\beta \left[ \sum_{i=1}^{N} \frac{P_i g_{il}}{N_0 W} - K \right]^+ & \text{if } \mu_l(t) = 0
\end{cases}
$$

The quantity $\sum_{i=1}^{N} \frac{P_i g_{il}}{N_0 W}$ can be measured at base station $l$ [19], and represents the sector’s overall interference. We refer to this quantity as *Rise Over Thermal* (ROT). Note that while employing this base algorithm may result in MAC-layer rates that fall outside the feasibility region $\Delta$, these rates will satisfy the original (non-linear) feasibility region from Section II.
It is important to note that neither the original nor the modified base algorithms take into account the fact that sources might not see full buffers at the MAC layer. In other words, the base algorithms require sources to transmit over the air with rate $\alpha_i$, even if the MAC layer queue is not long enough to sustain such a rate. Although wasteful of bandwidth during transient periods, this can be accomplished by transmitting “dummy” packets when necessary and will not impact the optimality of the equilibrium point. This issue becomes more important when dealing with packet-based systems (e.g. [14]), and will be an important area for future work.

Once the regulating signals $\mu_l$ are computed at each base, they are used to generate aggregate signals for each mobile. Recall the definition of each user’s aggregate wireless signal $p_i = \sum_{l=1}^{L} \frac{g_{il}}{g_{d(i)}} \mu_l$. At first glance it seems that in order to calculate $p_i$, each mobile requires full knowledge of the channel. In [8] and [9] we have shown that there exists a practical solution to this problem using the CDMA pilot signal, PS, and a pricing pilot signal, PPS. This pilot symbol is transmitted with a power level proportional to the base signal, $\mu_l$. Hence $p_i$ can be calculated as $p_i = \sum_{l=1}^{L} \frac{g_{il}}{g_{d(i)}} \mu_l \simeq \frac{E_{PPS}}{E_{P}(b(i))}$ where $E_{PPS}$ and $E_{P}(b(i))$ are quantities which can easily be measured locally by mobile $i$ (see [8] and [9] for more details).

The practical scheme to compute the wired link signals $\lambda$ and the corresponding aggregate signals $q$ is well understood since Eqn (5) has a well-known interpretation in terms of queue delay at each link [12], [4], [5]. When dealing with a discrete-time system, however, the usual differential equation for queueing delay $\dot{\lambda}_j = \frac{1}{C_j}(\sum_{i=1}^{M} x_i \psi_{ij} - C_j)$ becomes $\Delta \lambda_j = \frac{\Delta t}{C_j}(\sum_{i=1}^{M} x_i \psi_{ij} - C_j)$, where $\Delta t$ is the time between successive updates. In other words, if we take $\xi = \frac{\Delta t}{C_j}$, then $\lambda_j$ is the queueing delay at link $j$, and $q_i$ is user $i$’s end-to-end queueing delay in the wired IP network. Regarding wireless users running the modular algorithm in a real system, however, it is $\alpha_i$, and not $x_i$, which determines the queuing delay at the intermediate wired links. As such, we propose the approximation $\Delta \lambda_j \approx \frac{\Delta t}{C_j}(\sum_{i=1}^{N} \alpha_i \psi_{ij} + \sum_{i=N+1}^{M} x_i \psi_{ij} - C_j)$.

It is left to address the practical implementation of the cross-layer coordination signals $\nu_i^+$ and $\nu_i^-$ in the context of the modular algorithm. It is interesting to note that the the queueing delay
interpretation described above also applies here. When \( \zeta_1^{(i)} = \frac{\Delta t}{\alpha_i} \) and \( \zeta_2^{(i)} = \frac{\Delta t}{x_i} \), Eqns (6) and (7) are similar to the queueing delay equations if we think of two imaginary queues (see Figure 4): one with input traffic rate \( x_i \) and capacity \( \alpha_i \) (Queue 1), and one with input traffic rate \( \alpha_i \) and capacity \( x_i \) (Queue 2). Thus we choose \( \zeta_1^{(i)} = \frac{\Delta t}{\alpha_i} \) and \( \zeta_2^{(i)} = \frac{\Delta t}{x_i} \) to ensure that the quantities \( \nu_i^+ \) and \( \nu_i^- \) can be interpreted as the delays associated with Queues 1 and 2, respectively. This configuration is shown in Figure 4, and is similar to the concept of token buckets (see [22], [23], and references therein). Every time the transport layer sends traffic to Queue 1, it empties the same amount of traffic from Queue 2. Similarly, every time the MAC layer removes traffic from Queue 1 for transmission, it adds the same amount of traffic to Queue 2. Queue 1 is now our actual link, and Queue 2 is our token bucket. The difference between this and the traditional token bucket is that we do not use the token bucket to regulate service rate, but instead use it to keep track of the mismatch between the rates \( x_i \) and \( \alpha_i \).

Thus far, we have explained how the addition of a MAC-layer buffer and token bucket facilitates an online and practical computation of \( \nu_i^+ \) and \( \nu_i^- \). In reality, the addition of a MAC-layer buffer and token bucket has a three-fold impact: 1) it eliminates the need for explicit computation of the cross-layer coordination signals, 2) it creates a natural distributed and adaptive priority scheme based on the MAC buffer and token bucket length (as \( \nu_i = \nu_i^+ - \nu_i^- \) increases user \( i \)’s local rule becomes more aggressive), and 3) it allows the transport layer protocol to unconsciously take interference levels into account without any major modification of current protocols. In order to understand the third impact, recall that TCP Vegas uses end-to-end queueing delay as its feedback mechanism [4]. This quantity is obtained by measuring the round-trip-time and subtracting the propagation delay. Looking at Eqn (8), we notice that the quantity \( q_i + \nu_i \) is nothing more than the end-to-end queueing delay \( (q_i + \nu_i^+) \) minus the token bucket delay \( (\nu_i^-) \). In other words, the only necessary modification to current transport layer protocols is to subtract the token bucket delay (in addition to the propagation delay) from the round-trip time!
B. Convergence Result

From Theorem 1, we know that the convergence of the modular algorithm is dependent upon the step-size being “small enough” (see [24], pages 212-215 for further discussion). Recall, however, that our interpretation of the Lagrangian multipliers as queueing delays was based on choosing the step size as $\frac{\Delta}{C}$, where C is the link capacity or service rate at a queue. As a result, to guarantee convergence we must simply run the algorithm “fast enough.” This means that the convergence of the modular algorithm is dependent upon the time-scale of the distributed feedback loops shown in Figure 3.

With these issues in mind, we present the following corollary regarding the convergence of the modular algorithm.

**Corollary 1:** There exist values $\beta_0$ and $\delta_0$ such that for all $\beta < \beta_0$ and for all $\delta < \delta_0$, if $\xi_j = \delta \frac{C_j}{C}$, $\zeta_1(i) = \delta \frac{\alpha_i}{x_i}$, and $\zeta_2(i) = \delta \frac{x_i}{x_i}$ then the modular algorithm described by Eqns (4)-(9) converges to the solution to Problem P.

**Proof:** See Appendix III.

V. Simulation Results

In order to more carefully examine the behavior of the proposed algorithm, we now present simulations comparing it to a “one-shot” algorithm in which the MAC and transport layer functionalities are merged. In other words, this one-shot algorithm chooses a single rate $x_i$ which manages both the interference and congestion in the network as in Figure 1(b). This algorithm consists of the same Base and Link Algorithms described in Section III, but uses the following Source Algorithm:

$$x_i = \arg \max_x \left( \log(x) - x q_i - \frac{x}{W + \gamma x} p_i \right)$$ (11)

Since this algorithm merges the MAC and transport layer functionalities, there are no cross-layer coordination signals or MAC queues. It has been shown in [25] that this algorithm converges to the solution to Problem P.
Simulations were run using 20 wireless sources and 20 wired sources. The routing matrix was randomly generated over 6 links, each with a capacity of 5 Mbps. The mobile infrastructure consists of four base stations with antenna gains of 2, each 2500 m apart. Mobile positions were randomly generated. The simulations use a Cost-231 propagation model at 1.9 GHz between the mobiles and bases [19]. The values for $\gamma$ and $K$ are 4 dB and 6 dB, and the chip bandwidth $W$ is 1.2 MHz [26]. For the modular algorithm, we choose $\sigma = 0.75$. The PPS signals are broadcast every 0.1 ms (synchronized). The modular MAC Layer Source Algorithm is run every 0.1 ms, while the modular Transport Layer Source Algorithm and the one-shot Source Algorithm are run every 3.0 ms. When used, the parameters for power control, mobility, and shadowing are as follows. The power control loop was run every 0.025 ms using a $\Delta P$ of 0.25 dBm. Mobiles were allowed to move with a model of constant speed (100 km/hr) and random direction. Although this model is known to suffer from decay (where the initial average speed is higher than the long-term average [27]), we believe that it is sufficient for a cursory examination of the impact of mobility on algorithm performance. Finally, we model shadowing using the correlation model from [28] with variance 4.3 dB and correlation of 0.3 at 10 m, as suggested for urban environments.

Note that when they exist, the equilibrium points of the two distributed algorithms are the same. The difference between the two algorithms is in their dynamical behaviors, which we examine through the simulations shown in Figures 5 - 8. Figures 5 and 6 show the ROT (in dB) and delay (in ms) for the one-shot and modular algorithms when a new user is added to the system at time $t = 7.5 ms$ and another user leaves the system at time $t = 17.5 ms$. These simulations compare the algorithm performance under quasi-static conditions (i.e. stationary nodes, no shadowing, and perfect power control). Figures 7 and 8 also show the ROT and delay for the one-shot and modular algorithms. In these figures the number of users remains constant, but mobiles are allowed to move, and shadowing and imperfect power control are simulated. Based on these figures, we make the following observations regarding algorithm performance.

**Impact of Short-Term Changes in Network Topology**
From Figure 5, we see that the one-shot and modular algorithms perform similarly in terms of convergence speed when a user is added to or removed from the system. The difference is that the one-shot algorithm exhibits a larger spike in ROT than the modular algorithm when a new user arrives. This spike in ROT corresponds to a higher-than-acceptable interference level in the system, hence a higher-than acceptable BER. The nature of the modular algorithm seems to be buying something in terms of robustness to increases in interference.

On the other hand, Figure 6 shows the disadvantage of the modular algorithm: even before the new user is added, the modular algorithm has a higher end-to-end queueing delay than the one shot algorithm. This is due largely to the addition of the MAC-layer buffers. It is also interesting to note that when a new user is added, the modular algorithm suffers from a significant increase in the end-to-end delay for wireless users, while the one-shot algorithm results in a very slight decrease. This is due to the fact that the addition of a new user results in a lower total traffic
rate (due to the nature of the proportional fair solution), hence lower congestion in the system. Since there are no MAC-layer buffers in the one-shot algorithm, the decrease in congestion causes the wired links to become less congested, resulting in a decrease in queueing delay. In the modular case, however, the decrease in total rate is accompanied by a decrease in MAC buffer “capacities” (time-varying). It’s not surprising that the decrease in MAC buffer capacities seems to be more dominant than the decrease in overall network load.

**Impact of Multiple Time Scales**

From Figures 7 and 8, we see another benefit of the modular algorithm in a practical setting in which the transport layer is implemented with slow time constants (3.0 ms versus 0.1 ms for MAC layer). In the presence of fast varying channel conditions and imperfect power control, the modular approach remedies the sluggishness of the transport layer by sufficiently separating the operation of the two rate control mechanisms. Although a rigorous analysis of this fact is beyond the scope of our paper, Figures 7 and 8 illustrate this phenomenon. The one-shot algorithm exhibits large oscillations when it is run every 3 ms, while the modular algorithm performs well by separating the time scales of the MAC and transport layers.

What we can infer from these results is that although the one-shot and modular algorithms converge to the same optimal rate-assignment, they do exhibit different dynamic behaviors. In particular, there seems to be an inherent trade-off between robustness and delay. We see that the one-shot implementation results in lower delays, but the advantage of the modular algorithm is that it exhibits a higher level of robustness. Furthermore, the inherent structure of the modular algorithm makes it possible to run the MAC and transport source layer algorithms at different time scales, which can aid in deploying the proposed solution in realistic wireless (and mobile) networks.

**VI. CONCLUSIONS AND FUTURE WORK**

In this paper, we have developed a modular algorithm that uses intermediate MAC-layer buffers to regulate separate rate assignments at the MAC and transport layers. Practical implementation
schemes for the computation and communication of regulating signals have been addressed. Finally, we have shown that both of these algorithms converge to the same optimal points, but exhibit vastly different dynamical behaviors.

One clear area of future work is the extension of this work to other types of MAC schemes (e.g. the ALOHA scheme proposed in [29]), particularly in ad-hoc networks. Such an extension is by no means trivial, since the fundamental relationship between sources and links changes when moving from a cellular to an ad-hoc network (see [11] for further discussion on this topic).

Perhaps the most interesting avenue of future work in this area is to examine the dynamic behavior of the proposed algorithms in general, and stability issues in particular. The modular algorithm consists of three nested feedback loops operating at different time-scales (i.e. transport layer, MAC layer, and power control). The interactions of these three non-linear control loops is a complicated analytical topic which is not easily understood. We have summarized some of these challenges for MAC and power control loops in [10]. With the addition of the transport layer loop, these issues become even more complicated. In particular, the time-scales of these loops and the choice of $\sigma$ seem to have a significant impact on robustness and delay. For instance, our preliminary simulations indicate that a high value of $\sigma$ can be used to offset slow rate control loops, or to dampen the oscillations in total sector interference. In other words, it is an appropriate choice of $\sigma$ and time-scales that allows the modular algorithm to perform well in Figures 7 and 8 when the one-shot algorithm does not. A more definitive study of the impact of these parameters and their interactions is critical to achieving the best possible performance under realistic conditions.

In addition, it will be important to further examine the relationship between the time-scales of the feedback loops, the use of delay as a regulatory signal, and the convergence of the algorithms. Although we have chosen to obtain a small step-size by running the algorithms quickly, we could also have used a scaled version of delay as our regulatory signal. In other words, we could choose our step-sizes to be $\frac{\Delta t}{CS}$, where $S$ is the scaling factor. The aggregate
signals are now the end-to-end delay divided by $S$, which can still be locally computed by the sources. The drawback is that the actual delay is now $S$ times higher than if we had simply run the algorithm $S$ times faster. This is similar to the concept of tightly and loosely coupled links in [15], where the use of scaling results in loosely coupled links.

**Appendix I: Duality Gap for Problem P1**

The difficulty in showing that Problem P1 does not have a duality gap is that the equality constraint is non-linear. In order to address this, we note that by absorbing the equality into the other conditions, we can write the following Auxiliary Problem P2 whose solution coincides with Problem P1:

**Auxiliary Problem P2**

\[
\max_{x,r} \sum_{i=1}^{N} \log \left( \frac{r_i W}{1 - \gamma r_i} \right) + \sum_{i=N+1}^{M} \log(x_i)
\]

s.t.

\[
\sum_{i=1}^{N} r_i W \psi_{ij} + \sum_{i=N+1}^{M} x_i \psi_{ij} \leq C_j \quad \forall \ j \in J
\]

\[
\sum_{i=1}^{N} r_i g_{ib(i)} \leq \frac{K}{\gamma(1+K)} \quad \forall \ l \in L
\]

\[
0 \leq r_i \leq \frac{W}{W + \gamma + W} \quad \forall \ i \leq N
\]

Note that the solution to this problem is of the form \((r_1, \ldots, r_N, x_{N+1}, \ldots, x_M) \in \mathbb{R}^M\).

Our main goal in this section is to show that the solutions to the dual problems for P1 and P2 are the same (Lemma 1). Since Problem P2 does not have a duality gap (Lemma 2) and the primal solutions for P1 and P2 are the same, Problem P1 does not have a duality gap.

**Fact 1:** Let \((\tilde{\lambda}', \mu', \nu'^+, \nu'^-\)) be the solution to the dual of Problem P1. Then

\[
(\lambda', \mu', \nu'^+ , \nu'^-) = \min_{(\tilde{\lambda}, \mu, \nu^+ , \nu^-)} \max_{(x, r)} \mathcal{L}_1(x, r, \tilde{\lambda}, \mu, \nu^+ , \nu^-)
\]
where
\[
\mathcal{L}_1(x, r, \lambda, \mu, \nu^+, \nu^-) = \sum_{i=1}^{N} \left[ \sigma \log\left( \frac{r_i W}{1 - \gamma r_i} \right) + \frac{r_i W}{1 - \gamma r_i} (\nu^+_i - \nu^-_i) - r_i \sum_{l=1}^{L} \mu_l \frac{g_{il}}{g_{ib(i)}} \right] \\
+ \sum_{i=1}^{N} \left[ (1 - \sigma) \log(x_i) - x_i (\nu^+_i - \nu^-_i) - x_i \sum_{j=1}^{J} \lambda_j \psi_{ij} \right] \\
+ \sum_{i=N+1}^{M} \left[ \log(x_i) - x_i \sum_{j=1}^{J} \lambda_j \psi_{ij} \right] + \sum_{j=1}^{J} \lambda_j C_j + \sum_{l=1}^{L} \mu_l \frac{K}{\gamma(1 + K)}
\]

is the Lagrange function associated with P1.

**Fact 2:** Let \((\lambda^*, \mu^*)\) be the solution to the dual of Problem P2. Then
\[
(\lambda^*, \mu^*) = \min_{(\lambda, \mu)} \max_{(x, r)} \mathcal{L}_2(x, r, \lambda, \mu)
\]
where
\[
\mathcal{L}_2(x, r, \lambda, \mu) = \sum_{i=1}^{N} \left[ \log\left( \frac{r_i W}{1 - \gamma r_i} \right) - \frac{r_i W}{1 - \gamma r_i} (\nu^+_i - \nu^-_i) - r_i \sum_{l=1}^{L} \mu_l \frac{g_{il}}{g_{ib(i)}} \right] \\
+ \sum_{i=N+1}^{M} \left[ \log(x_i) - x_i \sum_{j=1}^{J} \lambda_j \psi_{ij} \right] + \sum_{j=1}^{J} \lambda_j C_j + \sum_{l=1}^{L} \mu_l \frac{K}{\gamma(1 + K)}
\]

is the Lagrange function associated with P2.

**Lemma 1:** Let \((x', r') = \arg \max_{(x, r)} \mathcal{L}_1(x, r, \lambda', \mu', \nu'^+, \nu'^-)\)
and
\[
(x^*, r^*) = \arg \max_{(x, r)} \mathcal{L}_2(x, r, \lambda^*, \mu^*)
\]
where \((\lambda', \mu', \nu'^+, \nu'^-)\) and \((\lambda^*, \mu^*)\) are the dual solutions to P1 and P2 as defined in Fact 1 and Fact 2. Then \(x'_i = \frac{r'_i W}{1 - \gamma r'_i} = \frac{r_i W}{1 - \gamma r_i} \forall i \leq N\), and \(x'_i = x^*_i \forall i > N\).

**Proof:**

For any arbitrary set of dual variables \((\lambda, \mu, \nu^+, \nu^-)\) in P1, let \(x^1(\lambda, \mu, \nu^+, \nu^-)\) and \(r^1(\lambda, \mu, \nu^+, \nu^-)\) be the maximizers of \(\mathcal{L}_1(x, r, \lambda, \mu, \nu^+, \nu^-)\). By their definition as maximizers of \(\mathcal{L}_1\), these functions must satisfy the following first-order conditions for \(i \leq N\):
\[
\frac{\partial \mathcal{L}_1}{\partial x_i} \bigg|_{x_i=x^1_i} = \frac{1 - \sigma}{x^1_i} - (\nu^+_i - \nu^-_i) - \sum_{j=1}^{J} \lambda_j \psi_{ij} = 0 \forall i \leq N
\]  \(\text{Eq. (12)}\)
\[
\frac{\partial \mathcal{L}_1}{\partial r_i}\bigg|_{r_i=r_i^1} = \frac{\sigma}{r_i^1(1-\gamma r_i^1)} + \frac{W}{(1-\gamma r_i^1)^2} (\nu_i^+ - \nu_i^-) - \sum_{l=1}^L \mu_l \frac{g_{il}}{g_{ib(l)}} = 0 \quad \forall \, i \leq N
\]

Similarly, let \(\nu^+ (\lambda, \mu)\) and \(\nu^- (\lambda, \mu)\) be the minimizers of \(\mathcal{L}_1 (x^1 (\lambda, \mu, \nu^+, \nu^-), r^1 (\lambda, \mu, \nu^+, \nu^-), \lambda, \mu, \nu^+, \nu^-)\). Again, by definition these functions must satisfy the following first-order conditions:

\[
\frac{\partial \mathcal{L}_1}{\partial \nu_i^+} \bigg|_{\nu_i^+ = \nu_i^{+1}} = \left[ 1 - \frac{\sigma}{x_i^1} - (\nu_i^+ - \nu_i^-) - \sum_{j=1}^J \lambda_j \psi_{ij} \right] \frac{\partial x_i^1}{\partial \nu_i^+} - x_i^1
\]

\[
+ \left[ \frac{\sigma}{r_i^1(1-\gamma r_i^1)} + \frac{W}{(1-\gamma r_i^1)^2} (\nu_i^+ - \nu_i^-) - \sum_{l=1}^L \mu_l \frac{g_{il}}{g_{ib(l)}} \right] \frac{\partial r_i^1}{\partial \nu_i^+} + \frac{r_i^1 W}{1-\gamma r_i^1} = 0
\]

\[
\frac{\partial \mathcal{L}_1}{\partial \nu_i^-} \bigg|_{\nu_i^- = \nu_i^{-1}} = \left[ 1 - \frac{\sigma}{x_i^1} - (\nu_i^+ - \nu_i^-) - \sum_{j=1}^J \lambda_j \psi_{ij} \right] \frac{\partial x_i^1}{\partial \nu_i^-} + x_i^1
\]

\[
+ \left[ \frac{\sigma}{r_i^1(1-\gamma r_i^1)} + \frac{W}{(1-\gamma r_i^1)^2} (\nu_i^+ - \nu_i^-) - \sum_{l=1}^L \mu_l \frac{g_{il}}{g_{ib(l)}} \right] \frac{\partial r_i^1}{\partial \nu_i^-} - \frac{r_i^1 W}{1-\gamma r_i^1} = 0
\]

where the second equality in each condition comes from (12) and (13). These conditions imply:

\[
x_i^1 = \frac{r_i^1 W}{1-\gamma r_i^1} \quad \forall \, i \leq N
\]

Now, consider the first order condition for \(r_i^1\):

\[
\frac{\partial \mathcal{L}_1}{\partial r_i} \bigg|_{r_i=r_i^1} = \frac{\sigma}{r_i^1(1-\gamma r_i^1)} + \frac{W}{(1-\gamma r_i^1)^2} (\nu_i^+ - \nu_i^-) - \sum_{l=1}^L \mu_l \frac{g_{il}}{g_{ib(l)}}
\]

\[
= \frac{\sigma}{r_i^1(1-\gamma r_i^1)} + \frac{W}{(1-\gamma r_i^1)^2} \left( 1 - \frac{\sigma}{x_i^1} - \sum_{j=1}^J \lambda_j \psi_{ij} \right) - \sum_{l=1}^L \mu_l \frac{g_{il}}{g_{ib(l)}}
\]

\[
= \frac{1}{r_i^1(1-\gamma r_i^1)} - \frac{W}{(1-\gamma r_i^1)^2} \sum_{j=1}^J \lambda_j \psi_{ij} - \sum_{l=1}^L \mu_l \frac{g_{il}}{g_{ib(l)}} = 0
\]

where the second line comes from (12) and the third line comes from (14). This is exactly the first order condition given by \(\frac{\partial \mathcal{L}_2}{\partial r_i} \bigg|_{r_i=r_i^2}\) when \(r_i^2 (\lambda, \mu)\) is the maximizer of \(\mathcal{L}_2 (\lambda, \mu, \nu^+, \nu^-)\).
Similarly, we note that for \( i > N \),
\[
\frac{\partial L_1^1}{\partial x_i} \bigg|_{x_i = x_i^1} = \frac{1}{x_i^1} - \sum_{j=1}^J \lambda_j \psi_{ij} = 0
\]
Again, this is exactly the first order condition given by \( \frac{\partial^2}{\partial x_i^2} \bigg|_{x_i = x_i^1} \) when \( x_i^2(\lambda, \mu) \) is the maximizer of \( L^2(x, r, \lambda, \mu) \). So for any \( (\lambda, \mu) \), we have
\[
x_i^1(\lambda, \mu, \nu^+, \nu^-) = \frac{r_i^1(\lambda, \mu, \nu^+, \nu^-)W}{1 - \gamma r_i^1(\lambda, \mu, \nu^+, \nu^-)} = \frac{r_i^2(\lambda, \mu)W}{1 - \gamma r_i^2(\lambda, \mu)} \quad \forall \, i \leq N
\]
and
\[
x_i^1(\lambda, \mu, \nu^+, \nu^-) = x_i^2(\lambda, \mu) \quad \forall \, i > N
\]
So far we have shown that for any given dual variables \( (\lambda, \mu) \) for P1 and P2, the associated functions giving the primal variables are of similar forms. All that is left is to show that the solution to the dual problems of P1 and P2 coincide (i.e. \( (\lambda', \mu') = (\lambda^*, \mu^*) \)). To do so, recall the definition of \( (\lambda', \mu') \):
\[
(\lambda', \mu') = \arg \min_{(\lambda, \mu)} \left[ \min_{(\nu^+, \nu^-)} \max_{(x, r, \lambda, \mu)} L^1(x, r, \lambda, \mu, \nu^+, \nu^-) \right]
\]
\[
= \arg \min_{(\lambda, \mu)} L^1(x^1, r^1, \lambda, \mu, \nu^+, \nu^-)
\]
\[
= \arg \min_{(\lambda, \mu)} \sum_{i=1}^N \left[ \sigma \log \left( \frac{r_i^1 W}{1 - \gamma r_i^1} \right) + \frac{r_i^1 W}{1 - \gamma r_i^1} (\nu_i^+ - \nu_i^-) - r_i^1 \sum_{l=1}^J \mu_l g_{i(l)} \right]
\]
\[
+ \sum_{i=1}^N \left[ (1 - \sigma) \log (x_i^1) - x_i^1 (\nu_i^+ - \nu_i^-) - x_i^1 \sum_{j=1}^J \lambda_j \psi_{ij} \right]
\]
\[
+ \sum_{i=N+1}^M \left[ \log (x_i^1) - x_i^1 \sum_{j=1}^J \lambda_j \psi_{ij} \right] + \sum_{j=1}^J \lambda_j C_j + \sum_{l=1}^L \mu_l \frac{K}{\gamma (1 + K)}
\]
\[
= \arg \min_{(\lambda, \mu)} \sum_{i=1}^N \left[ \log \left( \frac{r_i^1 W}{1 - \gamma r_i^1} \right) - \frac{r_i^1 W}{1 - \gamma r_i^1} \sum_{j=1}^J \lambda_j \psi_{ij} - r_i^1 \sum_{l=1}^J \mu_l g_{i(l)} \right]
\]
\[
+ \sum_{i=N+1}^M \left[ \log (x_i^1) - x_i^1 \sum_{j=1}^J \lambda_j \psi_{ij} \right] + \sum_{j=1}^J \lambda_j C_j + \sum_{l=1}^L \mu_l \frac{K}{\gamma (1 + K)}
\]
\[
= \arg \min_{(\lambda, \mu)} L^2(x^2, r^2, \lambda, \mu) = (\lambda^*, \mu^*)
\]
where the second-to-last equality is a direct consequence of (12) and (13), and we are done.

\[\square\]

**Lemma 2:** Problem P2 has no duality gap.
Proof: It is simple to verify that 1) the objective function in Problem P2 is strictly concave over the range \(0 \leq r_i \leq \frac{W}{W_{\gamma+W}}\), and 2) the constraints are convex over this same range. From Proposition 5.3.1 (page 512) in [30], we know that there is no duality gap if there exists a vector of primal variables \((x, r)\) such that

\[
\sum_{i=1}^{N} r_i W_1 - \gamma r_i < 0 \quad \forall \ j \in J
\]

\[
\sum_{i=1}^{N} r_i g_{il} - \frac{K}{\gamma(1+K)} < 0 \quad \forall \ l \in L
\]

This is clearly satisfied with \(r_i = 0 \ \forall \ i \leq N\) and \(x_i = 0 \ \forall \ i > N\), and we are done.

\[
\text{APPENDIX II: SUPPORTING DEFINITIONS AND LEMMAS FOR THEOREM 1}
\]

Fact 3: Given a set of Lagrangian multipliers and the corresponding dual objective function \(D(\cdot)\), the limit point generated by gradient projection with an appropriate choice of step size minimizes \(D(\cdot)\) if \(D(\cdot)\) is convex, lower bounded, continuously differentiable, and \(\nabla D(\cdot)\) is Lipschitz continuous (for proof, see [24], pages 213-214).

Definition 2: Let \(\eta\) be a \((J + L + 2N)\) by 1 vector containing the Lagrangian multipliers \(\lambda\), \(\mu\), \(\nu^+\) and \(\nu^-\). We can write the dual objective function as

\[
D(\eta) = \sum_{i=1}^{N} \left[ \max_x \left( (1 - \sigma) \log(x) - x \sum_{j=1}^{J} \psi_{ij} \eta_j - x(\eta_{J+L+i} - \eta_{J+L+N+i}) \right) \right]
\]

\[
+ \sum_{i=N+1}^{M} \left[ \max_x \left( \log(x) - x \sum_{j=1}^{J} \psi_{ij} \eta_j \right) \right]
\]

\[
+ \sum_{i=1}^{N} \left[ \max_r \left( \sigma \log(\frac{rW}{1 - \gamma r}) + \frac{rW}{1 - \gamma r} (\eta_{J+L+i} - \eta_{J+L+N+i}) - r \sum_{j=J+1}^{J+L} \frac{g_{il}}{g_{ib(i)}} \eta_j \right) \right]
\]

\[
+ \sum_{j=1}^{J} \eta_j C_j + \frac{K}{\gamma(1+K)} \sum_{j=J+1}^{L} \eta_j
\]

Lemma 3: The dual objective function \(D(\eta)\) is lower bounded, convex, and continuously differentiable over the feasibility region.

Proof: Under the allowable region of effective rates, the utility function \(\log(\frac{rW}{1-\gamma r})\) is strictly increasing, concave, and twice differentiable. This, combined with the properties of the dual objective function and weak duality, leads immediately to Lemma 3 [4].
**Fact 4:** For a given set of Lagrange multipliers ($\eta$), the rates which maximize $L_M(x, r, \eta)$ are given by

$$x_i^* = \arg \max_{0 \leq x \leq \frac{W}{\gamma}} \left( (1 - \sigma) \log(x) - x \sum_{j=1}^{J} \psi_{ij} \eta_j - x(\eta_{J+L+i} - \eta_{J+L+N+i}) \right) \forall i \leq N$$  \hspace{1cm} (15)

and

$$x_i^* = \arg \max_{0 \leq x \leq \frac{W}{\gamma}} \left( \log(x) - x \sum_{j=1}^{J} \psi_{ij} \eta_j \right) \forall i > N \hspace{1cm} (16)$$

and

$$r_i^* = \arg \max_{0 \leq r \leq \frac{rW}{1 - \gamma r}} g_i(r) \hspace{1cm} (17)$$

where $g_i(r) = \sigma \log\left(\frac{rW}{1 - \gamma r}\right) + \frac{rW}{1 - \gamma r} (\eta_{J+L+i} - \eta_{J+L+N+i}) - r \sum_{j=J+1}^{J+L} \frac{g_{ij}}{g_{(b(i))}} \eta_j$.

**Fact 5:** The gradient of the dual objective function is given by the following column vector:

$$\nabla D(\eta) = \begin{bmatrix} C_j - \sum_{i=1}^{M} x_i^* \psi_{ij} \\ \frac{K}{\gamma(1+K)} - \sum_{i=1}^{N} r_i^* g_{ij} \frac{g_{(b(i))}}{g_{(b(i))}} \\ \frac{r^*_{j,J-L}W}{1 - \gamma r^*_{j,J-L}} - x_{j-J-L}^* \\ x_{j-J-L-N}^* - \frac{r^*_{j-J-L-N}W}{1 - \gamma r^*_{j,J-L}} \end{bmatrix} \begin{bmatrix} J \times 1 \\ L \times 1 \\ N \times 1 \\ N \times 1 \end{bmatrix}$$

(see [24], page 669).

**Fact 6:** The Hessian of the dual objective function is given by the following matrix:

$$\nabla^2 D(\eta) = \begin{bmatrix} -\sum_{i=1}^{M} \frac{\partial x_i^*}{\partial \eta_k} \psi_{ij} \\ -\sum_{i=1}^{N} \frac{\partial r_i^*}{\partial \eta_k} g_{ij} \frac{g_{(b(i))}}{g_{(b(i))}} \\ \frac{W}{1 - \gamma r_{j-J-L}^*} \frac{\partial r_{j-J-L}^*}{\partial \eta_k} - \frac{\partial x_{j-J-L}^*}{\partial \eta_k} \\ \frac{\partial x_{j-J-L-N}^*}{\partial \eta_k} - \frac{W}{1 - \gamma r_{j-J-L-N}^*} \frac{\partial r_{j-J-L-N}^*}{\partial \eta_k} \end{bmatrix} \begin{bmatrix} J \times (J + L + 2N) \\ L \times (J + L + 2N) \\ (N \times J + L + 2N) \\ (N \times J + L + 2N) \end{bmatrix}$$

Now we introduce Propositions 1 and 3 which provide bounds on appropriate terms in $\nabla^2 D(\eta)$, and Proposition 2 which is an auxiliary step bounding $\eta$.

**Proposition 1:** The quantity $|\frac{\partial x_i^*}{\partial \eta_k}|$ is upper bounded by $\frac{W^2}{(1 - \sigma) \gamma}$ for every $i, k$. 

Proof: Using Fact 4, we write

\[
\frac{\partial x^*_i}{\partial \eta_k} = \begin{cases} 
-\frac{1}{1-\sigma}\psi_{ik}x^2_i & \text{if } 1 \leq k \leq J \text{ and } i \leq N \\
-\psi_{ik}x^2_i & \text{if } 1 \leq k \leq J \text{ and } i > N \\
0 & \text{if } J + 1 \leq K \leq J + L \\
-\frac{1}{1-\sigma}x^2_i & \text{if } J + L + 1 \leq k \leq J + L + N \text{ and } i \leq N \text{ and } i = k - J - L \\
\frac{1}{1-\sigma}x^2_i & \text{if } J + L + N < K \text{ and } i \leq N \text{ and } i = k - J - L - N \\\n0 & \text{else}
\end{cases}
\]

Noticing that \(0 \leq x^*_i \leq \frac{W}{\gamma}\) gives us the result.

Proposition 2: If the quantity \(g_i(r)\) is not maximized on the boundary of the feasibility region for a given user \(i\), then

\[
(\eta_{J+L+i} - \eta_{J+L+N+i}) < \frac{\sigma(1-\gamma r^*_i)(1-2\gamma r^*_i)}{W r^*_i (3-2\gamma r^*_i)}.
\]

Proof: If the quantity \(g_i(r)\) is not maximized on the boundary of the feasibility region, then \(\frac{\partial g_i}{\partial r} < 0\) for some \(r \in (0, \frac{1}{2\gamma} - \varepsilon)\). We know that

\[
\frac{\partial g_i}{\partial r} = \frac{\sigma(1-\gamma r^*_i)+rW(\eta_{J+L+i}-\eta_{J+L+N+i})-r(1-\gamma r^*_i)^2 \sum_{j=J+1}^{J+L} \eta_j \frac{g_{ij}}{g_{ib(i)}}}{r(1-\gamma r^*_i)^2},
\]

and the denominator of this function is positive over the allowable range of effective rates. The numerator, \(f_i(r)\), has the following characteristics:

1. \(f_i(r)\) has two imaginary roots and one real root
2. \(f_i(0) = \sigma\)
3. \(\lim_{r \to \infty} f_i(r) = -\infty\)

From this, we can infer that if \(\frac{\partial g_i}{\partial r} < 0\) for some \(r \in (0, \frac{1}{2\gamma} - \varepsilon)\), then \(f_i(\frac{1}{2\gamma}) < 0\). This is equivalent to the condition

\[
\frac{\sigma}{2} + \frac{W(\eta_{J+L+i} - \eta_{J+L+N+i})}{2\gamma} - \frac{1}{8\gamma} \sum_{j=J+1}^{J+L} \eta_j \frac{g_{ij}}{g_{ib(i)}} < 0 \quad (18)
\]

From Eqn (17) in Fact 4, we have

\[
\frac{\sigma}{r^*_i(1-\gamma r^*_i)} + \frac{W(\eta_{J+L+i} - \eta_{J+L+N+i})}{(1-\gamma r^*_i)^2} - \sum_{j=J+1}^{J+L} \eta_j \frac{g_{ij}}{g_{ib(i)}} = 0 \quad (19)
\]

Substituting Eqn (19) into Eqn (18) yields the desired result.

Proposition 3: The quantity \(\left|\frac{\partial r^*_i}{\partial \eta_k}\right|\) is upper bounded by \(\frac{W(1+\gamma\varepsilon)}{2\sigma\gamma^3(1+2\gamma\varepsilon)(1+4\gamma\varepsilon)}\) for every \(i, k\).
Proof: By convention, we set \( \frac{\partial r^*_i}{\partial \eta_k} = 0 \) for all \( k \) if \( g_i(r) \) is maximized on the boundary of the feasibility region. Otherwise, using Fact 4, we write:

\[
\frac{\partial r^*_i}{\partial \eta_k} = \begin{cases} 
0 & \text{if } 1 \leq k \leq J \\
\frac{g_{ik} r^*_i (1-\gamma r^*_i)^3}{h_i(r^*_i)} & \text{if } J + 1 \leq k \leq J + L \\
\frac{-W r^*_i (1-\gamma r^*_i)}{h_i(r^*_i)} & \text{if } J + L + 1 \leq k \leq J + L + N \text{ and } i = k - J - L \\
\frac{W r^*_i (1-\gamma r^*_i)}{h_i(r^*_i)} & \text{if } k > J + L + N \text{ and } i = k - J - L - N \\
0 & \text{else}
\end{cases}
\]

where \( h_i(r^*_i) = 2\gamma W (\eta_{J+L+i} - \eta_{J+L+N+i}) r^*_i - \sigma (1-2\gamma r^*_i)(1-\gamma r^*_i) \). Combining this with Proposition 2 gives \( h_i(r^*_i) \leq \frac{\sigma(1-\gamma r^*_i)(1-2\gamma r^*_i)(4\gamma r^*_i - 3)}{3(2\gamma r^*_i)} \). Noticing that the right hand side is increasing in \( r^*_i \) and \( r^*_i \leq \frac{1}{2\gamma} - \varepsilon \), we have \( h_i(r^*_i) \leq \frac{-\sigma \gamma (1+2\gamma \varepsilon)(1+4\gamma \varepsilon)}{2(1+\gamma \varepsilon)} \). Furthermore, since \( r^*_i \) and \( (1-\gamma r^*_i) \) are bounded, we are done.

Lemma 4: The quantity \( \|\nabla^2 D(\eta)\|_1 = \max_k \sum_{j=1}^{J+L+2N} \nabla^2 D_{Mjk} \) is upper bounded by \( (J + L + 2N) \left( \frac{2W^2N(1+\gamma \varepsilon)}{\sigma \gamma^3(1+2\gamma \varepsilon)(1+4\gamma \varepsilon)} + \frac{MW^2}{(1-\sigma)\gamma^2} \right) \).

Proof: From Fact 6 and Propositions 1 and 3, it is easy to verify that the absolute value of every element in the matrix \( \nabla^2 D(\eta) \) is upper bounded by \( \frac{2W^2N(1+\gamma \varepsilon)}{\sigma \gamma^3(1+2\gamma \varepsilon)(1+4\gamma \varepsilon)} + \frac{MW^2}{(1-\sigma)\gamma^2} \). This lemma is then an immediate consequence of the definition of norm-1 (see [24], pages 626, 634-635).

Lemma 5: The Hessian of the dual objective function \( \nabla D(\eta) \) is Lipschitz continuous (i.e. \( \exists K \) such that \( \|\nabla D(\eta) - \nabla D(\eta')\|_2 \leq K\|\eta - \eta'\|_2 \forall \eta, \eta' \)).

Proof: Using the mean value theorem (see [24] page 639) and norm definitions, we have

\[
\|\nabla D(\eta) - \nabla D(\eta')\|_2 = \|\nabla^2 D(\eta)(\eta - \eta')\|_2 \\
\leq \|\nabla^2 D(\eta)\|_2\|\eta - \eta'\|_2
\]
Now, using the properties of the norm (see [24], pages 626, 634-635), we can write

\[ \|\nabla^2 D(\eta_1)\|_2^2 \leq \|\nabla^2 D(\eta_1)\|_{\infty} \|\nabla^2 D(\eta_1)\|_1 \]
\[ = \|\nabla^2 D(\eta_1)\|_1^2 \]

This quantity is bounded from Lemma 4, and we are done.

This theorem is now an immediate consequence of Fact 3 and Lemmas 3 and 5.

**Appendix III: Supporting Definitions and Lemmas for Corollary 1**

**Fact 7:** Given a set of Lagrangian multipliers and the corresponding dual objective function \(D(\cdot)\), the limit point generated by scaled gradient projection with an appropriate choice of step size minimizes \(D(\cdot)\) if \(D(\cdot)\) is convex, lower bounded, continuously differentiable, \(\nabla D(\cdot)\) is Lipschitz continuous, and the invertible, symmetric scaling matrix \(M(t)\) is chosen such that it satisfies \((x - y)^T M(t)(x - y) \geq A \|x - y\|_2^2 \forall x, y \in \text{Domain}(D(\cdot))\), where \(A\) is some positive constant (for proof, see [24], pages 213-217).

**Lemma 6:** Choose the scaling matrix \(M(t)\) to be a diagonal matrix whose entries correspond to the sequence \(C_1, \ldots, C_J, 1_1, \ldots, 1_L, \alpha_1, \ldots, \alpha_N, x_1, \ldots, x_N\) (where \(1_i\) represents the \(i^{th}\) 1 in the sequence). Then \(M(t)\) is invertible, symmetric, and satisfies the positivity constraint \((x - y)^T M(t)(x - y) \geq A \|x - y\|_2^2 \forall x, y \in \text{Domain}(D(\eta))\) for some positive constant \(A\).

**Proof:** It is clear from inspection that \(M(t)\) is symmetric. Fact 4 gives the equations for finding \(x_i^*\) and \(r_i^*\). Since \(\log(0) = -\infty\), we know that \(x_i^* = 0\) and \(r_i^* = 0\) can never be solutions to these equations. This guarantees that \(M(t)\) is invertible. Finally, since \(M(t)\) is a diagonal matrix with non-zero entries, we know that it also satisfies the positivity constraint for some positive constant \(A\).

This corollary is now an immediate consequence of Fact 7 and Lemma 6, along with Lemmas 3 and 5 from Appendix II.
REFERENCES


