Non-convex Optimization and Rate Control for Multi-class Services in the Internet

Jang-Won Lee, Ravi R. Mazumdar, and Ness B. Shroff

School of Electrical and Computer Engineering

Purdue University

West Lafayette, IN 47907, USA

{lee46, mazum, shroff}@ecn.purdue.edu

Abstract

In this paper, we investigate the problem of distributively allocating transmission data rates to users in the Internet. We allow users to have concave as well as sigmoidal utility functions as appropriate for different applications. In the literature, for simplicity, most works have dealt only with the concave utility function. However, we show that applying rate control algorithms developed for concave utility functions in a more realistic setting (with both concave and sigmoidal types of utility functions) could lead to instability and high network congestion. We show that a pricing based mechanism that solves the dual formulation can be developed based on the theory of subdifferentials with the property that the prices "self-regulate" the users to access the resources based on the net utility. We discuss convergence issues and show that an algorithm can be developed that is *efficient* in the sense of achieving the global optimum when there are many users.

I. Introduction

Over the last decades, there has been a significant amount of interest in the area of Internet rate control, which aims at providing satisfactory services and alleviating congestion in the Internet. Currently, most services in the Internet are elastic to some degree, i.e., the sources can adjust their transmission

This research has been supported in part by NSF grants ANI-0073359, ANI-9805441, and ANI-0207728.

data rates in response to congestion levels within the network. Hence, by appropriately exploiting the elasticity through rate control, one can maintain high network efficiency while at the same time alleviating network congestion. To that end, it is necessary to have an appropriate model to characterize the elasticity of the service. This is typically done using the well-known concept of a utility function that represents the level of user satisfaction or Quality of Service (QoS) at the allocated rate.

We can classify services in the Internet into two classes based on the shape of the utility function. One corresponds to traditional data services, such as file transfer and email. These services can adjust their transmission data rates gradually, resulting in graceful degradation of the QoS in the presence of network congestion. The elasticity of these services can be modeled by concave utility functions [1]. The other corresponds to delay and rate adaptive services, such as streaming video and audio services. These services are less elastic than data services. In response to network congestion, they can decrease their transmission data rates up to a certain level with a corresponding graceful degradation in the QoS. However, decreasing the transmission data rate below a certain threshold results in a significant drop in the QoS (e.g., below a certain bit rate, the quality of audio communication falls dramatically). The elasticity of these services can be modeled by using sigmoidal-like utility functions [1]. We call an increasing function $f(x)$ a *sigmoidal-like function*, if it has one inflection point x_o, and $\frac{d^2f(x)}{dx^2} > 0$, for $x < x_o$ and $\frac{d^2f(x)}{dx^2} < 0$, for $x > x_o$, as shown in Fig. 1.

In the past few years, utility based rate control problems have begun to be addressed by the networking research community [2], [3], [4], [5], [6], [7], [8], [9]. They have almost exclusively dealt with the situation where the utilities are concave for which there exist extensive theories and algorithms such as

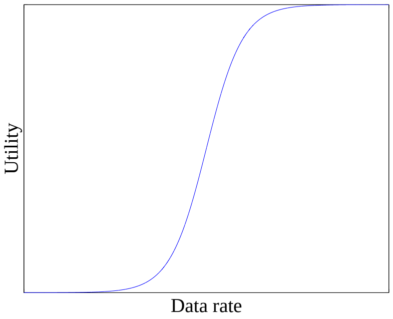


Fig. 1. A sigmoidal-like function.

the Karush-Kuhn-Tucker (KKT) conditions and the duality theorem. However, as mentioned before, concave utility functions are appropriate only for modeling traditional data services, and do not capture the characteristics of services such as audio and video that are becoming increasingly popular in the Internet. Hence, for the efficient allocation of transmission rates among services with diverse characteristics, a rate control algorithm must be able to efficiently handle delay and rate adaptive services with sigmoidal-like utility functions as well as data services with concave utility functions. But this results in the non-convexity of the system, which is, in general, difficult to handle.

A natural and logical approach to dealing with the issue of non-convexity is to simply approximate a sigmoidal-like utility function with a concave function and use one of the algorithms developed for concave utility functions. However, this approach could result in a highly inefficient solution. For example, suppose that a system has a single bottleneck link with capacity 10 Mbps and 11 users. Further, suppose that each user has the same utility function $U(x)$ that is a step function described below:

$$U(x) = \begin{cases} 0, & \text{if } x < 1 \text{ Mbps} \\ 1, & \text{if } x \geq 1 \text{ Mbps} \end{cases} .$$

Note that the step function is an extreme case of a sigmoidal-like function. If we approximate $U(x)$ with a concave function, $U'(x)$, we can apply an algorithm for concave utility functions that has been proposed in the literature to maximize the total system utility. In this case, since all users have the same concave utility functions, at the global optimal solution, each user is allocated the same amount of rate, $x^* = \frac{10}{11}$ Mbps, which provides $U(x^*) = 0$. Hence, with this approach, we achieve zero total system utility for the original utility function. However, by allocating 1 Mbps to 10 users and zero to one user, we can achieve a total system utility of 10 units. Even though this example considers an extreme case, it emphasizes that to efficiently accommodate diverse services in the Internet, it is necessary to develop a rate allocation algorithm that takes into account the properties of both concave and sigmoidal-like utility functions.

Recently, the implications of non-convexity of the system have begun to be addressed in the literature [10], [11], [12]. In fact, in [12], the author showed that even though each individual user has a concave utility function, the overall problem might be non-convex in a system with multi-class services. This implies that non-convexity is an important issue for rate control in the Internet. In this paper, we will study this problem by considering a situation similar to the current Internet. In the Internet, there is no central authority in the system that performs admission control or rate control and each user behaves in a selfish manner. Thus, a rate control algorithm must be implemented in a distributed manner taking into account the selfish behavior of users. In the papers mentioned earlier, it has been shown that if all users have concave utility functions, efficient distributed rate allocations can be obtained by using an appropriate congestion indicator in the network, in spite of the selfish behavior of users. However, as we will show later, if such algorithms developed for concave functions are now applied to non-concave functions, they may result in the situation when the system cannot support all the users causing instability and excessive congestion in the system. To prevent this situation from happening, some users must be turned off and this decision must be made by the user itself, since there is no central authority in the system. To this end, we will develop the algorithm with which each user "self-regulates" its access to the network based on the local information.

The rest of the paper is organized as follows. In Section II, we describe the system model and present the problem that is being considered in this paper. We develop and study the rate control algorithm in Section III. We provide numerical results for our algorithm in Section IV and conclude in Section V.

II. System Description and Basic Problem

We consider a system that consists of L links and N users. Each link l has capacity C_l, and each user i has a utility function U_i and maximum transmission data rate M_i ($0 < M_i < \infty$). We assume that U_i has the following properties.

Properties of the utility function:

(U1) U_i is an increasing function of x_i the *allocated rate* for user i.

(U2) U_i is twice continuously differentiable.

(U3) U_i is a sigmoidal-like or strictly concave function.

(U4) $\frac{dU_i(x_i)}{dx_i} < \infty$ for all $0 \leq x \leq M_i$.

In the following, if U_i is a sigmoidal-like function, we let x_i^o be its inflection point. Otherwise (i.e., if U_i is a concave function), we let $x_i^o = 0$.

Our objective is to obtain a data rate allocation for the users that maximizes the total system utility. This is formulated as:

$$\begin{aligned} \text{(A)} \qquad & \max \sum_{i=1}^{N} U_i(x_i) \\ & \text{subject to} \quad A\bar{x} \le \bar{C} \\ & \qquad\qquad\quad \bar{0} \le \bar{x} \le \bar{M}, \end{aligned}$$

where $\bar{x} = (x_1, x_2, \cdots, x_N)^T$[1], $\bar{C} = (C_1, C_2, \cdots, C_L)^T$, $\bar{0} = (0, 0, \cdots, 0)^T$, $\bar{M} = (M_1, M_2, \cdots, M_N)^T$, and $A = (a_{ij})$ is an $L \times N$ matrix such that

$$a_{ij} = \begin{cases} 1, & \text{if user } j \text{ uses link } i \\ 0, & \text{otherwise} \end{cases} .$$

We define

$$T(i) \;=\; \{l \mid a_{li} = 1,\ 1 \le l \le L\},\ i = 1, 2, \cdots, N$$

and

$$S(l) \;=\; \{i \mid a_{li} = 1,\ 1 \le i \le N\},\ l = 1, 2, \cdots, L.$$

Hence, $T(i)$ is a set of links that user i is using and $S(l)$ is a set of users that are using link l.

Note that since we allow non-concave utility functions, problem (A) is a non-convex programming problem, which, is usually more difficult to solve than a convex programming problem. In [11], [13], similar problems to (A) were studied. In [13], the problem was studied in the context of the power allocation in wireless environment. However, in [13], we looked at this problem only in the context of a single cell, which can be viewed as corresponding to a single link in the Internet. Further, the algorithm in [13] requires a central controller, such as a base-station in cellular systems, which is clearly not applicable to decentralized networks, such as the Internet. In [11], which is an earlier conference version of the this paper, the rate control problem in the Internet for a single link was considered. In this paper, we have studied the rate control problem in the Internet with multiple links.

[1] $\bar{x}^T$ is a transpose of a vector $\bar{x}$.

III. RATE CONTROL

In this section, we develop a distributed rate control algorithm for problem (A) by using the theory of subdifferentials. For background, we first provide definitions and properties of subdifferentials. We refer readers to [14], [15], [16] for details.

Definition 1: A vector $d \in R^n$ is a subgradient of a convex function $f : R^n \rightarrow R$ at $x \in R^n$, if

$$f(z) \geq f(x) + (z - x)^T d, \text{ for all } z \in R^n.$$

Definition 2: The set of all subgradients of a convex function f at $x \in R^n$ is called the subdifferential of f at x and denoted by $\partial f(x)$.

Properties of the subgradient:

(S1) A function $f(x)$ is differentiable at x, if and only if it has a unique subgradient at x. In this case, the subgradient is equal to the gradient of f at x.

(S2) $x \in X \subset R^n$ minimizes a convex function f over a convex set X, if and only if there exists a subgradient d such that $d^T(z - x) \geq 0,$ for all $z \in X$.

(S3) If x is an interior point of X, then (S2) implies that x minimizes a convex function f over a convex set X, if and only if $0 \in \partial f(x)$.

A. Dual Problem

As mentioned before, problem (A) is a non-convex programming problem, which is difficult to solve. Hence, we will consider its dual since the dual has some advantages over the primal problem. For example:

- The dual is a convex programming problem and thus easier to solve.
- The separable property of the dual makes it easy to implement the algorithm in a distributed fashion.
- From a networking perspective, the dual will usually have a smaller dimension and simpler constraints than the primal. This will reduce the complexity of the algorithm. In our case, the primal has a dimension of N and the dual has a dimension of L, where N is the number of users in the network and L is the number of links in the network. In general, we have $L \ll N$.

However, since the primal is not a convex programming problem (e.g., if some of the utility functions are sigmoidal), there could be a duality gap between the primal and its dual. Hence, by solving the dual, we may not obtain the optimal primal solution. This is one of the difficulties that we will overcome in this work, especially in the context of many users.

We now define a Lagrangian function associated with problem (A) as:

$$L(\bar{x},\bar{\lambda}) \;=\; \sum_{i=1}^{N} U_i(x_i) + \bar{\lambda}^T(\bar{C} - A\bar{x}), \tag{1}$$

where $\bar{\lambda}^T = (\lambda_1, \lambda_2, \cdots, \lambda_L)$. Then, the dual of problem (A) is defined by

$$\text{(B)} \qquad \min Q(\bar{\lambda})$$
$$\text{subject to} \quad \bar{\lambda} \geq \bar{0},$$

where

$$Q(\bar{\lambda}) \;=\; \max_{\bar{0}\leq\bar{x}\leq\bar{M}} L(\bar{x},\bar{\lambda}). \tag{2}$$

It can easily be shown that $Q(\bar{\lambda})$ is a convex function of $\bar{\lambda}$ [16]. However, as we will show later, $Q(\bar{\lambda})$ may not be everywhere differentiable. Hence, even though $Q(\bar{\lambda})$ is a convex function, we cannot use a simple gradient based algorithm to find a minimizer, as in [4] and [5], since clearly $Q(\bar{\lambda})$ does not have a gradient at the point where it is not differentiable.

To solve problem (B), we first study the properties of $Q(\bar{\lambda})$ by using the theory of the subdifferentials. We now characterize the subdifferentials of $Q(\bar{\lambda})$. First, we can rewrite $L(\bar{x},\bar{\lambda})$ in (1) as:

$$\begin{aligned} L(\bar{x},\bar{\lambda}) &= \sum_{i=1}^{N} U_i(x_i) - \sum_{i=1}^{N} x_i \sum_{j\in T(i)} \lambda_j + \sum_{j=1}^{L} \lambda_j C_j \\ &= \sum_{i=1}^{N} U_i(x_i) - \sum_{i=1}^{N} \lambda_{T(i)} x_i + \sum_{j=1}^{L} \lambda_j C_j, \end{aligned}$$

where

$$\lambda_{T(i)} = \sum_{j\in T(i)} \lambda_j.$$

Since it is separable in $\bar{x}$, $\bar{x}(\bar{\lambda}) = (x_1(\lambda_{T(1)}), x_2(\lambda_{T(2)}), \cdots, x_N(\lambda_{T(N)}))^T$ solves (2) if and only if it solves the following:

$$x_i(\lambda_{T(i)}) \;=\; \arg\max_{0\leq x\leq M_i} \{NU_i(\lambda_{T(i)}, x)\}, \text{ for } i = 1, 2, \cdots, N, \tag{3}$$

where

$$NU_i(\lambda_{T(i)}, x) = U_i(x) - \lambda_{T(i)} x.$$

The properties of $x_i(\lambda_{T(i)})$ were studied in [13]. First, we define λ_i^{max} for user i as:

$$\lambda_i^{max} = \min\{\lambda_{T(i)} \geq 0 \mid \max_{0 \leq x \leq M_i} \{NU_i(\lambda_{T(i)}, x)\} = 0\}. \tag{4}$$

We can calculate λ_i^{max} by solving the following equation [13]:

$$\lambda_i^{max} = \begin{cases} \frac{dU_i(x)}{dx}|_{x=0}, & \text{if } U_i \text{ is a concave function} \\ \frac{dU_i(x)}{dx}|_{x=x'}, & \text{if } U_i \text{ is a sigmoidal-like function and } x' \text{ exists} \\ \frac{U_i(M_i)}{M_i}, & \text{otherwise} \end{cases},$$

where x' is a solution of the following equation:

$$U_i(x) - x\frac{dU_i(x)}{dx} \quad = \quad 0, \qquad x_i^o \leq x \leq M_i,$$

and x_i^o is the inflection point of U_i, when U_i is sigmoidal. Also, define λ_i^{min} for user i as:

$$\lambda_i^{min} \quad = \quad \max\{\lambda_{T(i)} \geq 0 | x_i(\lambda_{T(i)}) = M_i\}.$$

Obviously, $0 < \lambda_i^{max} < \infty$ and $\lambda_i^{max} \geq \lambda_i^{min}$. Then, $x_i(\lambda_{T(i)})$ has the following properties [13]:

Properties of $x_i(\lambda_{T(i)})$:

(R1) If U_i is a sigmoidal-like function (i.e., $0 < x_i^o < M_i$), then $x_i(\lambda_{T(i)})$ has two values (zero and positive) and is discontinuous at λ_i^{max}. Otherwise, $x_i(\lambda_{T(i)})$ has a unique value and is continuous.

(R2) $x_i(\lambda_{T(i)})$ is positive and a decreasing function of λ, for $\lambda_i^{min} \leq \lambda_{T(i)} < \lambda_i^{max}$.

(R3) $x_i(\lambda_{T(i)})$ is zero, for $\lambda_{T(i)} > \lambda_i^{max}$.

(R4) $x_i(\lambda_{T(i)})$ is M_i, for $\lambda_{T(i)} \leq \lambda_i^{min}$.

(R5) $U_i(x_i(\lambda_i^{max}))$ is achieved at the concave region of U_i.

Note that, if U_i is a concave function, $x_i(\lambda_{T(i)})$ is a continuous and non-increasing function. However, if U_i is a sigmoidal-like function, $x_i(\lambda_{T(i)})$ is not only discontinuous but also has two values at λ_i^{max}. One is zero and the other is positive. In the sequel, unless explicitly mentioned otherwise, $x_i(\lambda_{T(i)})$ will denote a positive value, if (3) has two solutions.

Since the Lagrangian function, $L(\bar{x}, .)$ is differentiable for all $\bar{0} \leq \bar{x} \leq \bar{M}$, and $\nabla_{\bar{\lambda}} L(., \bar{\lambda})$ is continuous for all $\bar{0} \leq \bar{x} \leq \bar{M}$, by Danskin's Theorem [16], the subdifferential of $Q(\bar{\lambda})$, $\partial Q(\bar{\lambda})$, is obtained as:

$$\begin{aligned} \partial Q(\bar{\lambda}) &= \operatorname{conv}(\{\nabla_{\bar{\lambda}} L(\bar{x}, \bar{\lambda}) \mid \bar{x} \in \bar{x}(\bar{\lambda})\}) \\ &= \operatorname{conv}(\{(C_1 - \sum_{i \in S(1)} x_i, \cdots, C_L - \sum_{i \in S(L)} x_i)^T \mid \bar{x} \in \bar{x}(\bar{\lambda})\}) \end{aligned}$$

where $\bar{x}(\bar{\lambda})$ is a set of solutions of (3) at $\bar{\lambda}$, and $\operatorname{conv}(G)$ is a convex hull of a set G. Hence, by using the properties of $x_i(\lambda_{T(i)})$, the subdifferential of $Q(\bar{\lambda})$, $\partial Q(\bar{\lambda})$, is obtained as follows. Let $\bar{q}(\bar{\lambda}) = (q_1(\bar{\lambda}), q_2(\bar{\lambda}), \cdots, q_L(\bar{\lambda}))^T \in \partial Q(\bar{\lambda})$. Then, for each $l = 1, 2, \cdots, L$,

$$q_l(\bar{\lambda}) \in \begin{cases} \{d_l \mid C_l - \sum_{i \in S^H(l,\bar{\lambda}) \cup S^S(l,\bar{\lambda})} x_i(\lambda_{T(i)}) \leq d_l \leq C_l - \sum_{i \in S^H(l,\bar{\lambda})} x_i(\lambda_{T(i)})\}, & \text{if there exists a user } i,\ i \in S^S(l,\bar{\lambda}) \text{ such that } 0 < x_i^o < M_i \\ \{C_l - \sum_{i \in S^H(l,\bar{\lambda})} x_i(\lambda_{T(i)})\}, & \text{otherwise} \end{cases} ,\quad (5)$$

where we divided the set of users into three subsets associated with link l and $\bar{\lambda}$ as

$$\begin{aligned} S^H(l, \bar{\lambda}) &= \{i \mid \lambda_i^{max} > \lambda_{T(i)},\ i \in S(l)\}, \\ S^S(l, \bar{\lambda}) &= \{i \mid \lambda_i^{max} = \lambda_{T(i)},\ i \in S(l)\}, \text{ and} \\ S^L(l, \bar{\lambda}) &= \{i \mid \lambda_i^{max} < \lambda_{T(i)},\ i \in S(l)\}. \end{aligned} \quad (6)$$

Hence, by the properties of $x_i(\lambda_{T(i)})$, $x_i(\lambda_{T(i)}) = 0$, if $i \in S^L(l, \bar{\lambda})$, $x_i(\lambda_{T(i)}) > 0$, if $i \in S^H(l, \bar{\lambda})$, and $x_i(\lambda_{T(i)})$ has two values (zero and positive), if $i \in S^S(l, \bar{\lambda})$.

We now solve the dual problem (B). As shown in (5), if there exists user i whose utility function is a sigmoidal-like function (i.e., $0 < x_i^o < M_i$), then the subgradient of $Q(\bar{\lambda})$ is not unique for all $\bar{\lambda}$. Hence, by properties (S1) and (R1), $Q(\bar{\lambda})$ may not be differentiable everywhere and we cannot use a gradient based method to solve problem (B). To overcome this, we will consider a subgradient projection method, which is formulated using an iterative algorithm such as:

$$\bar{\lambda}^{(n+1)} = [\bar{\lambda}^{(n)} - \alpha^{(n)}(\bar{C} - A\bar{x}(\bar{\lambda}^{(n)}))]^+, \quad (7)$$

where $\bar{x}(\bar{\lambda}^{(n)})$ is a solution of (3) at $\bar{\lambda} = \bar{\lambda}^{(n)}$ and $[\bar{a}]^+ = \max\{\bar{a}, \bar{0}\}$ in component-wise sense. By (5), $\bar{C} - A\bar{x}(\bar{\lambda}^{(n)})$ is a subgradient of $Q(\bar{\lambda})$ at $\bar{\lambda} = \bar{\lambda}^{(n)}$. To make $\bar{\lambda}^{(n)}$ in (7) converge to $\bar{\lambda}^o$, the optimal solution of the dual problem (B), we must have an appropriate sequence of $\alpha^{(n)}$. In gradient based algorithms in [4] and [5], there exists a constant step size, $\alpha^{(n)} = \alpha$, which ensures that $\lambda^{(n)}$ converges to λ^o. However, in the subgradient based algorithm, we cannot guarantee the convergence of $\bar{\lambda}^{(n)}$ with a constant step size, since the subgradient, $\bar{C} - A\bar{x}(\lambda^{(n)})$ that we use in (7), may not be zero at $\bar{\lambda}^o$. Hence, we will consider the following sequence:

$$\alpha^{(n)} \to 0, \text{ as } n \to \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \alpha^{(n)} = \infty. \tag{8}$$

Then, $\bar{\lambda}^{(n)}$ in (7) converges to the optimal solution λ^o of the dual problem (B), with the sequence that satisfies the conditions in (8) [17].

B. Distributed Algorithm for the Dual Problem

In the previous subsection, we have established that the solution of (3) and (7) with coefficients satisfying (8) converges to $\bar{\lambda}^o$, the dual optimal solution. This algorithm can be implemented in a distributed way. At iteration n, user i transmits its data at a rate determined by solving (3) with $\lambda_{T(i)} = \lambda_{T(i)}^{(n)}$. In this case, we can interpret $\lambda_l^{(n)}$ as the price per unit rate at link l at iteration n, and $\lambda_{T(i)}^{(n)}$ as the price per unit rate that user i must pay to use the links in set $T(i)$ at iteration n. With this interpretation, by solving (3), user i tries to maximize $NU_i(\lambda_{T(i)}, x)$, its net utility, at the price $\lambda = \lambda_{T(i)}^{(n)}$ without considering other users. This is a natural property of selfishness (i.e., the non-cooperative property) of the user in a public environment, such as the Internet. Also, we can interpret λ_i^{max} as the maximum willingness to pay per unit rate of user i, since if the price per unit rate $\lambda_{T(i)}$ is higher than λ_i^{max}, $x_i(\lambda_{T(i)})$ will be zero by property (R3) (i.e., user i does not transmit its data). Note that the utility and the net utility must be calculated with the received rate (allocated rate). However, the user does not know its received rate before it transmits data. Thus, the user maximizes its net utility with the transmission data rate assuming that the received rate is same as the transmission data rate.

Based on the aggregate transmission data rate of users that use link l, link l updates $\lambda_l^{(n+1)}$, the price per unit rate of the next iteration, by solving the following equation:

$$\lambda_l^{(n+1)} = [\lambda_l^{(n)} - \alpha^{(n)}(C_l - \sum_{i \in S(l)} x_i(\lambda_{T(i)}^{(n)})]^+, \ l = 1, 2, \cdots, L. \tag{9}$$

Note that solving (9) for each link is equivalent to solving (7). This implies that a link tries to obtain the optimal price per unit rate that solves the dual problem by adjusting the price based on its congestion level (i.e., the aggregate transmission rate of the users that use the link). Also, the link tries to maximize the utilization of its capacity without causing congestion by equating the aggregate transmission data rate of users with its capacity.

C. Properties of the Primal Solution

Thus far, we have considered the dual of problem (A) and developed an algorithm that converges to an optimal solution $\bar{\lambda}^o$ of the dual. When there is no duality gap between the primal and its dual, the dual solution also solves the optimal primal problem. However, when some of the utilities are non-concave, the primal problem (A) is not a convex programming problem. In this case, there could exist a duality gap between the primal and its dual, i.e., the solution of problem (B) need not result in the optimal solution of problem (A). In this paper, we are more interested in the rate allocation (the primal solution) than the price (the dual solution). Thus, it is important to study how “good” a primal solution can be obtained by solving the dual. To this end, we next study the properties of the primal solution corresponding to the dual optimal solution.

Proposition 1: Suppose that $\bar{\lambda}^o$ is an optimal solution of the dual problem (B). Then, if $Q(\bar{\lambda})$ is differentiable at $\bar{\lambda}^o$, $\bar{x}(\bar{\lambda}^{(n)})$ converges to $\bar{x}(\bar{\lambda}^o)$. Moreover, $\bar{x}(\bar{\lambda}^o)$ is an optimal solution of the primal problem (A). However, otherwise, $\bar{x}(\bar{\lambda}^{(n)})$ may not converge even though $\bar{\lambda}^{(n)}$ converges to $\bar{\lambda}^o$.

Proof: See Appendix A. ■

Proposition 2: If $Q(\bar{\lambda})$ is not differentiable at $\bar{\lambda}^o$, then there exists a link l^* that satisfies one of the following conditions:

$$\sum_{i \in S^H(l^*, \bar{\lambda}^o)} x_i(\lambda_{T(i)}^o) < C_{l^*} - \epsilon_1 \quad \textit{and} \quad \sum_{i \in S^H(l^*, \bar{\lambda}^o) \cup S^S(l^*, \bar{\lambda}^o)} x_i(\lambda_{T(i)}^o) > C_{l^*} + \epsilon_2,$$

$$\sum_{i \in S^H(l^*, \bar{\lambda}^o)} x_i(\lambda^o_{T(i)}) \le C_{l^*} \quad \textit{and} \quad \sum_{i \in S^H(l^*, \bar{\lambda}^o) \cup S^S(l^*, \bar{\lambda}^o)} x_i(\lambda^o_{T(i)}) > C_{l^*} + \epsilon_3, \ \textit{or} \tag{10}$$
$$\sum_{i \in S^H(l^*, \bar{\lambda}^o)} x_i(\lambda^o_{T(i)}) < C_{l^*} - \epsilon_4 \quad \textit{and} \quad \sum_{i \in S^H(l^*, \bar{\lambda}^o) \cup S^S(l^*, \bar{\lambda}^o)} x_i(\lambda^o_{T(i)}) \ge C_{l^*},$$

where ϵ_1, ϵ_2, ϵ_3, and ϵ_4 are some positive constants, and subsets of users, $S^H(l^*, \bar{\lambda}^o)$ and $S^S(l^*, \bar{\lambda}^o)$ are defined in (6).

Proof: See Appendix B. ■

Propositions 1 and 2 imply that when $\bar{\lambda}^{(n)}$ converges to $\bar{\lambda}^o$, the rate allocation may oscillate between two cases. In one case, the constraint is satisfied (i.e., the aggregate transmission rate of the users does not exceed the capacity of the link), while in the other case, the constraint cannot be satisfied (i,e., the aggregate transmission rate of the users exceeds the capacity of the link). Since the aggregate transmission data rate of users can exceed the capacity of the link, congestion may occur at the link. Note that one of the conditions in (10) is satisfied only if there exists some user i such that $x_i(\lambda_{T(i)})$ is discontinuous at $\bar{\lambda}^o$, i.e., the oscillation happens because of the discontinuity of $x_i(\lambda_{T(i)})$ when U_i is a sigmoidal-like function. *Thus, if there exist users having sigmoidal-like utility functions, the rate allocation resulting from solving the dual problem, such as the algorithms in [4], [5] (that converges to an efficient rate allocation with concave utility functions), may cause congestion without convergence.* This implies that the system cannot accommodate all the users and some of them must be interrupted to alleviate the congestion in the system. Since there is no central authority in the Internet, this must be done in a distributed way. Hence, to resolve this situation, we impose a “self-regulating” property on the users. In the next subsection, we will study the “self-regulating” property and show that using the this property, the algorithm converges to the solution that satisfies the constraint and is also an asymptotically optimal rate allocation.

D. “Self-regulating” Property

To study the “self-regulating” property, we assume that the condition in Proposition 2 is satisfied in this subsection. Thus, there exists a link l^* that satisfies one of the conditions in (10). We first define what we mean by the “self-regulating” property and make additional assumptions on the convergence of the algorithm having this property.

Self-regulating property: The property of a user that it does not transmit data even though the price is less than its maximum willingness to pay, if it will always receive net utility that is less than or equal to δ in the future.

We will show how to implement the "self-regulating" property in practice later. Note that, if $\delta = 0$, with the "self-regulating" property, users continue to be selfish, i.e., they still preserve the non-cooperative property. We call it the non-cooperative property (selfishness) in a strict sense. If $\delta > 0$, but δ can be made arbitrarily close to zero, we call it the non-cooperative property (selfishness) in a wide sense.

To exploit the "self-regulating" property of users in the rate control, we assume that the system has the following properties.

Assumptions on the "self-regulating" property:

(A1) Each user is "self-regulating", i.e., it satisfies the "self-regulating" policy.

(A2) Each user i has thresholds of tolerance th_i and δ_i such that if it receives net utility less than δ_i by transmitting data for th_i iterations consecutively, it stops transmitting data.

(A3) Link l allocates a rate $x'_i(\lambda_{T(i)})$ to each user $i \in S(l)$ that is defined by

$$x'_i(\lambda_{T(i)}) = \begin{cases} x_i(\lambda_{T(i)}), & \text{if } \sum_{j \in S(l)} x_j(\lambda_{T(j)}) \le C_l \\ f_i^l(\bar{x}^l(\bar{\lambda})), & \text{if } \sum_{j \in S(l)} x_j(\lambda_{T(j)}) > C_l \end{cases},$$

where $x_i(\lambda_{T(i)})$ is the transmission data rate of user i, $\bar{x}^l(\bar{\lambda})$ is a vector for the transmission rates of users in $S(l)$, and f_i^l is a continuous function of $\bar{x}^l(\bar{\lambda})$ that satisfies the following conditions:

$$f_i^l(\bar{x}^l(\bar{\lambda})) \le x_i(\lambda_{T(i)}) \quad \text{and} \quad \sum_{j \in S(l)} f_j^l(\bar{x}^l(\bar{\lambda})) = C_l.$$

A good candidate for function f_i^l is

$$f_i^l(\bar{x}^l(\bar{\lambda})) = \frac{x_i(\lambda_{T(i)})}{\sum_{j \in S(l)} x_j(\lambda_{T(j)})} C_l,$$

which can be achieved by the First Come First Service (FCFS) policy.

In this subsection, we focus on link l^* that satisfies one of the conditions in (10) and divide users in set $S(l^*)$ into three subsets, $S^H(l^*, \bar{\lambda}^o)$, $S^S(l^*, \bar{\lambda}^o)$, and $S^L(l^*, \bar{\lambda}^o)$, as in (6). We now assume that the algorithm is at the m^{th} iteration such that for all $n \geq m$, the following conditions are satisfied:

$$\lambda_{T(i)}^{(n)} < \lambda_i^{max},\ i \in S^H(l^*, \bar{\lambda}^o) \quad \text{and} \quad \lambda_{T(i)}^{(n)} > \lambda_i^{max},\ i \in S^L(l^*, \bar{\lambda}^o).$$

Since $\bar{\lambda}^{(n)}$ converges to $\bar{\lambda}^o$, there exists an m that satisfies the above conditions. Hence, users in set $S^L(l^*, \bar{\lambda}^o)$ do not transmit data anymore and users in set $S^H(l^*, \bar{\lambda}^o)$ always transmit data after iteration m. However, users in set $S^S(l^*, \bar{\lambda}^o)$ may continue to resume and stop data transmission again. When a user in set $S^S(l^*, \bar{\lambda}^o)$ transmits data, it may obtain positive net utility. However, the next proposition implies that for any $\delta_i > 0$, if $\bar{\lambda}^{(n)}$ is enough close to $\bar{\lambda}^o$, user i, $i \in S^S(l^*, \bar{\lambda}^o)$ always obtains net utility that is less than δ_i. The users in $S^S(l^*, \bar{\lambda}^o)$ would eventually stop transmitting at $\bar{\lambda}^o$, and this limiting behavior can be equivalently captured using the $\epsilon - \delta$ definition of convergence, thus giving rise to a finite window for these users to stop transmitting.

Proposition 3: For any $\delta_i > 0$ and user i, $i \in S^S(l^*, \bar{\lambda}^o)$, there exists an $m_i(\delta_i)$ such that $NU_i(\lambda_{T(i)}^{(n)}, x_i'(\lambda_{T(i)}^{(n)})) \leq \delta_i$ for all $n \geq m_i(\delta_i)$ where $NU_i(\lambda_{T(i)}^{(n)}, x_i'(\lambda_{T(i)}^{(n)}))$ is the received net utility of user i with price $\lambda_{T(i)}^{(n)}$ and received rate $x_i'(\lambda_{T(i)}^{(n)})$.

Proof: See Appendix C ■

Hence, by "self-regulating" itself, user i, $i \in S^S(l^*, \bar{\lambda}^o)$ stops transmitting data after iteration $m_i(\delta_i)$ where δ_i is a threshold of user i in the "self-regulating" property.

This procedure will be repeated for other users in the set $S^S(l^*, \bar{\lambda}^o)$ until any condition in Proposition 2 is not satisfied (i.e., the condition in Proposition 1 is satisfied) for the remaining users. After that, rate allocation converges to a rate allocation that satisfies the constraint, since it converges to an optimal rate allocation for the remaining users by Proposition 1. Since Proposition 3 is true for any $\delta_i > 0$, we can have an arbitrary small $\delta_i > 0$. With this property, we can say that each user still preserves the non-cooperative property (i.e., selfishness) in a wide sense. Further, the next corollary shows that if the system has a single bottleneck link, with the "self-regulating" property, each user preserves the non-cooperative property (i.e., selfishness) in a strict sense [11].

Corollary 1: If the system has a single bottleneck link l^* that satisfies one of the conditions in (10), for each user i, $i \in S^S(l^*, \bar{\lambda}^o)$, there exists an m_i such that $NU_i(\lambda_{T(i)}^{(n)}, x_i'(\lambda_{T(i)}^{(n)})) \leq 0$ for all $n \geq m_i$.

However, even if there exists iteration $m_i(\delta_i)$ for user i that satisfies the condition in Proposition 3, it may not be possible for the user to ascertain this. For example, during a transient period, a user may receive net utility that is lower than δ_i, even though it would receive net utility that is much higher than δ_i in the future. Hence, it may not be a good strategy to stop transmitting data immediately after it receives net utility that is lower than δ_i. Thus, the idea behind (A2) is to not turn user i off immediately, but only after it has received net utility that is less than δ_i for th_i consecutive iterations. This implies that, by an appropriate choice of th_i, user i stops transmitting data only after th_i iterations of iteration $m_i(\delta_i)$. Note that, in this scheme, it is important to have an appropriate threshold, th_i. If it is too small, user i may stop transmitting data during the transient period even if it can receive net utility that is higher than δ_i in the future. On the other hand, if it is too large, the algorithm may take very long to converge.

As long as the users are "self-regulating," our algorithm converges to the rate allocation that satisfies the constraint. Hence, our rate control algorithm does not cause congestion within the network even with non-concave utility functions. However, we still need to study the efficiency of our method in general because even though it results in an optimal rate allocation for the remaining users, it may not result in an optimal rate allocation for all users. To study this, we first define some variables as:

- $\bar{x}^o$: the optimal primal solution, i.e., the optimal rate allocation.
- $\bar{\lambda}^o$: the optimal dual solution.
- $\bar{x}^*$: our rate allocation.
- $\bar{x}(\bar{\lambda}^o)$: the transmission data rate at $\bar{\lambda}^o$.
- R^s: a subset of users that stop transmitting data due to the "self-regulating" property in our rate control algorithm.

Proposition 4: If $\frac{\sum_{i\in R^s} U_i(x_i(\lambda^o_{T(i)}))}{\sum_{i=1}^N U_i(x_i^o)} \to 0$ *as* $N \to \infty$*, then* $\frac{\sum_{i=1}^N U_i(x_i^*)}{\sum_{i=1}^N U_i(x_i^o)} \to 1$ *as* $N \to \infty$.

Proof: See Appendix D. ■

Proposition 4 states that our rate allocation is asymptotically optimal. This means that we would expect to have a good approximation of the global optimal rate allocation, when there are many users in a system with large capacity and the number of users in the set R^s has vanishing proportion. Hence, for our algorithm to converge to an efficient rate allocation, we need the condition that the number of users that stop transmitting data due to the “self-regulating” property has vanishing proportion. We will study the effect that this condition has on the efficiency of our algorithm later and also propose methods to make this number small.

Thus far, we have shown that the algorithm based on the subgradient and the “self-regulating” property converges to an asymptotically optimal rate allocation without causing congestion within the system. As mentioned before, in the subgradient based algorithm, we cannot guarantee convergence with a constant step size. Hence, we use a step size that diminishes to zero. However, the constant step size can more efficiently track system variations, such as initiation and completion of calls than a diminishing step size. In the next proposition, we will show that if each user applies the “self-regulating” property with the following additional assumption for the utility function, there exists a constant step size α with which the algorithm in (3) and (9) converges.

(U5) $-\frac{d^2U_i(x)}{dx^2} \geq c > 0$ for all $x_i(\lambda_i^{max}) \leq x \leq M_i$, $i = 1, 2, \cdots, N$.

Proposition 5: Assuming that each user is “self-regulating,” there exists a constant step size α with which our algorithm converges.

Proof: See Appendix E. ■

E. The Worst Case

In the previous subsection, we have shown that our rate allocation could be a good approximation of the global optimal rate allocation. However, it could also be inefficient in certain cases. In this subsection, we show an example of the worst case and provide solutions to resolve it. We consider a system with a single bottleneck link l with capacity C_l. We assume that each user i has the same utility function U that is a sigmoidal-like function, the same thresholds of tolerance th and δ. By assuming that each user has the same utility function, each user has the same maximum willingness to pay λ^{max}. Further, assume that $\sum_{i=1}^{N} x_i(\lambda^{max}) > C_l$. In this case, $\lambda_l^o = \lambda^{max}$ and we have $S^H(l, \lambda_l^o) = \emptyset$, $S^L(l, \lambda_l^o) = \emptyset$, and $S^S(l, \lambda_l^o) = \{1, 2, \cdots, N\}$, and one of the conditions in (10) is satisfied, since

$\sum_{i\in S^H(l,\lambda_l^o)} x_i(\lambda^{max}) = 0 < C_l$ and $\sum_{i\in S^H(l,\lambda_l^o)\cup S^S(l,\lambda_l^o)} x_i(\lambda^{max}) = \sum_{i=1}^{N} x_i(\lambda^{max}) > C_l$. Hence, there exist some users in set $S^S(l,\lambda_l^o)$ that stop transmitting data due to the "self-regulating" property. But, since all users have the same thresholds of tolerance, all users stop transmitting data at the same time, which results in zero total system utility.

Note that in the situation above, the parameters of each user are synchronized, i.e., each user has same maximum willingness to pay and thresholds of tolerance. Furthermore, they use the same set of links, i.e., they pay the same price per unit rate. Therefore, we expect this to occur very rarely in the Internet. In general, users in the Internet may have different characteristics such as different utility functions (i.e., different maximum willingness to pays). The status of links that each user uses may differ in a high degree and, thus, each user may pay a different price per unit rate. Hence, the probability that the parameters of many users are synchronized is very small in the Internet and, in most cases, our rate allocation could be an efficient rate allocation. However, to further reduce the probability that the parameters of many users are synchronized, we can use one of the following two methods. First, we can slightly perturb (randomly) the utility function of each user. By doing this, each user i has a different maximum willingness to pay, λ_i^{max}, with high probability while making the effect on the performance of each user small. Second, we can assume that the thresholds of tolerance (th_i and δ_i) of each user depend on the preference of the user. This ensures that users stop transmitting data at different iterations even if they have the same maximum willingness to pay and the same price.

F. Complexity

In this subsection, we compare the complexity of our subgradient based algorithm that considers both concave and sigmoidal-like utility functions with that of the gradient based algorithms in [4] and [5] that consider only concave utility functions.

To calculate the price of the next iteration, the subgradient based algorithm uses a subgradient while the gradient algorithm uses a gradient. Further, in general, we cannot guarantee the convergence of the subgradient algorithm with a constant step size, while the gradient based algorithm converges with a constant step size. However, in our algorithm, a subgradient is calculated from the difference between the capacity and the aggregate transmission data rate of all users that use the link, which is identical to

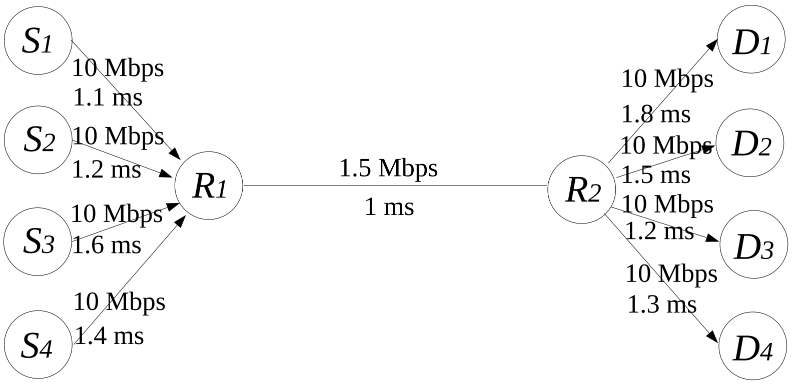


Fig. 2. A system with a single bottleneck link.

how a gradient is computed in the gradient based algorithm. Moreover, in Proposition 5, we have shown that our algorithm converges even with a constant step size when each user is "self-regulating," as in the gradient based algorithm . Thus, our algorithm and the algorithms in [4] and [5] have the same price update rule at the link. Further, both the algorithms have the same update rule for the transmission data rate in each user. Hence, the only modification that we need to make is at the end users (i.e., imposing the "self-regulating" property). This property is required because $x_i(\lambda_{T(i)})$ in (3) is not continuous at λ_i^{max}, if the utility function of user i is a sigmoidal-like function. If the utility function of user i is a concave function, $x_i(\lambda_{T(i)})$ is continuous and we do not need the "self-regulating" property for user i. Hence, compared with the algorithms in [4] and [5], we have to add the "self-regulating" property to users with sigmoidal-like utility functions in our algorithm. This requires calculating the received net utility by measuring the received rate. This can be easily done either by counting the number of ACK packets or by explicit notification of the received rate from the destination.

IV. Numerical Results

In this section, we provide simulation results using an ns-2 simulator. For the simulation, each link updates its price per unit rate every 200 msec by solving (9) with a constant step size of 0.03. To forward the price to users, we add a field for the price in the header of a packet that has zero as its initial value. Whenever a packet passes through a link, the link adds its current price to the value in the field for the price. At the destination, the price in the received packet is copied to the field of an acknowledgment

(ACK) packet and is sent to the source. We assume that a data packet and an ACK packet consist of 500 bytes and 40 bytes, respectively. The source estimates the received rate by counting the number of ACK packets and calculates the received utility and the received net utility by using the estimated received rate. By the transmission data rate update rule, if the price becomes higher than its maximum willingness to pay, a user does not transmit data packets. If this happens in the transient period, the user cannot be informed of the price for the next iteration, since the price is conveyed by ACK packets from the destination in our simulation setting. Thus, we allow the user to transmit packets at a very low rate, even though its transmission data rate that maximizes its net utility is zero during the transient period. By doing this, the user can be informed the price for the next iteration by the ACK packets from the destination. To that end, in the simulation, a user transmits two packets, each of which consists of 40 bytes, at every iteration (200 msec).

A. A System with a Single Bottleneck Link: Comparison with a System without the "Self-regulating" Property

We first consider a system with a single bottleneck link in Fig. 2. In this figure, we provide the capacity and the propagation delay of each link. User i transmits packets from source node S_i to destination node D_i with utility function U_i. Users 1 and 4 have a sigmoidal utility function given by

$$U_i(x) \quad = \quad c_i(\frac{1}{1+e^{-a_i(x-b_i)}}+d_i), \tag{11}$$

where c_i and d_i are used for the normalization of the function and x is a rate in a unit of Megabit per second (Mbps). Users 2 and 3 have a log utility function given by

$$U_i(x) \quad = \quad c_i(\log(a_i x+b_i)+d_i). \tag{12}$$

In this simulation, we normalize the utility function such that $U_i(0) = 0$ and $U_i(M_i) = 1$, where M_i is the maximum transmission data rate of user i (it is not necessary to normalize the utility function). User i has its thresholds of tolerance, th_i and δ_i, and starts transmitting data packets at time st_i sec. We provide parameters of each user in Table I and plot the utility function of each user in Fig. 3.

We compare two systems: a system with the "self-regulating" property and a system without the "self-regulating" property. Note that the algorithm for the system without the "self-regulating" property

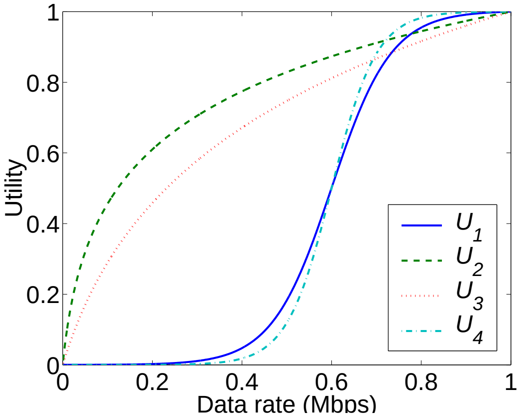


Fig. 3. The utility function of each user.

TABLE I

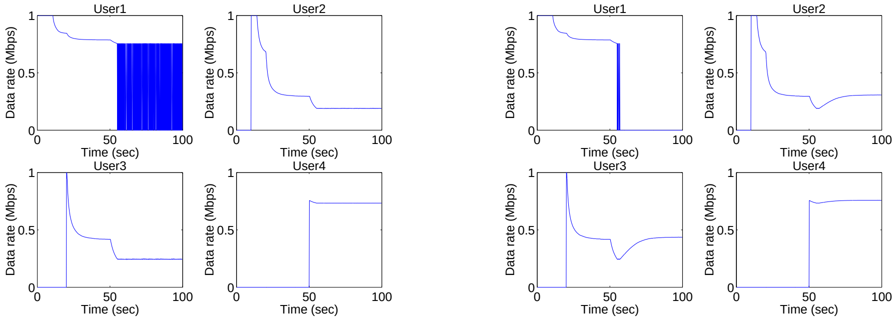
PARAMETERS FOR USERS (SINGLE BOTTLENECK LINK)

User i	Type	a_i	b_i	M_i	th_i	δ_i	st_i	λ_i^{max}	$x_i(\lambda_1^{max})$	$x_i(\lambda_4^{max})$
1	Sigmoid	15	0.6	1	20	0	0	1.210	0.756	0
2	Log	50	1	1	20	0	10	12.717	0.190	0.179
3	Log	10	1	1	20	0	20	4.170	0.245	0.226
4	Sigmoid	20	0.6	1	20	0	50	1.276	0.734	0.731

is the same as the gradient based algorithms in [4] and [5]. Thus, the results for this system show the behavior of the algorithms developed in the literature for concave utility functions when applied to a network supporting users with both concave and sigmoidal utility functions. We plot the transmission data rate, the received data rate, and the received net utility of each user in Figs. 4, 5, and 6, respectively.

The results show that before user 4 starts transmitting packets (50 sec), the two systems yield the same results. When only users 1, 2, and 3 are in the system, as shown in Table I, $\sum_{i=1}^{3} x_i(\lambda_1^{max}) = 1.191$ (Mbps) < 1.5 (Mbps), where λ_1^{max} is the smallest maximum willingness to pay among those of users in the system. Thus, we can have $\lambda^o < \lambda_1^{max}$, such that $\sum_{i=1}^{3} x_i(\lambda^o) = 1.5$ (Mbps). Then, by (S2), λ^o is a dual optimal solution and it satisfies the condition in Proposition 1. Hence, the algorithm converges to the optimal rate allocation without relying on the "self-regulating" property of users.

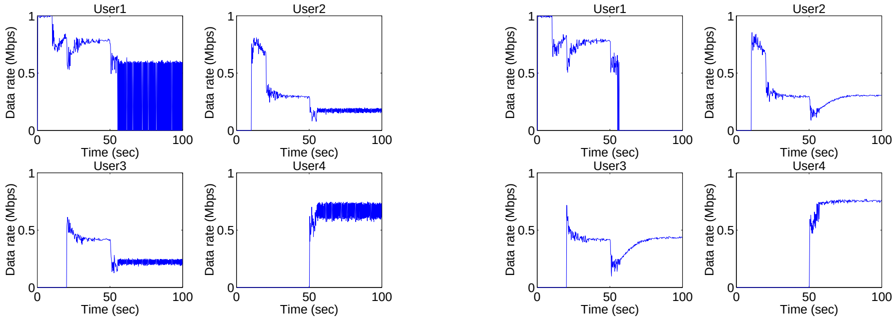
However, when all four users are in the system, as shown in Table I, $\sum_{i=1}^{4} x_i(\lambda_1^{max}) = 1.925$ (Mbps) $>$



(a) Without self-regulating.

(b) With self-regulating.

Fig. 4. Transmission data rate (a single bottleneck link).

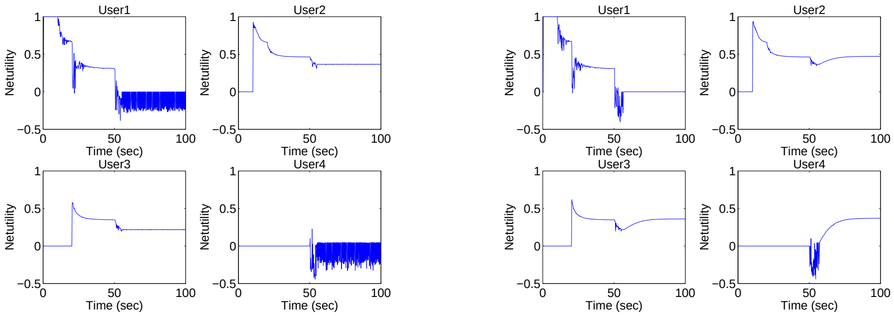


(a) Without self-regulating.

(b) With self-regulating.

Fig. 5. Received data rate (a single bottleneck link).

1.5 (Mbps) and $\sum_{i=2}^{4} x_i(\lambda_1^{max}) = 1.169$ (Mbps) < 1.5 (Mbps), where λ_1^{max} is the smallest maximum willingness to pay among users. In this case, by (S2), λ_1^{max} is a dual optimal solution and it satisfies the condition in Proposition 2 with $S^H(l^*, \lambda_1^{max}) = \{2, 3, 4\}$ and $S^S(l^*, \lambda_1^{max}) = \{1\}$. Therefore, in the system without the "self-regulating" property, after user 4 starts transmitting packets, the transmission data rate of user 1 (the primal solution) keeps oscillating, as shown in Fig. 4(a). In this case, when user 1 transmits packets, the aggregate transmission data rate of all users exceeds the capacity of the link. This causes congestion at the link and a large number of packet losses for all users. Thus, as shown in Figs. 4(a) and 5(a), each user has a large difference between the transmission data rate and the received data rate. Further, due to these packet losses, some users have negative received net utility, even though each user determines its transmission data rate by solving (3) so that it has non-negative net utility if



(a) Without self-regulating.

(b) With self-regulating.

Fig. 6. Received net utility (a single bottleneck link).

there is no packet loss. As shown in Fig. 6(a), after user 4 starts transmitting packets, the net utility of user 1 becomes non-positive and the net utility of user 4 oscillates between positive and negative values. These results show that if there exist users with non-concave utility functions in the system, using a rate control algorithm devised only for concave utility functions could result in an unstable system as well as a large amount of network congestion.

However, in the system with the "self-regulating" property, as shown in Fig. 4(b), user 1 stops transmitting packets due to the "self-regulating" property, after having received non-positive net utility values for th_1 consecutive iterations. After user 1 stops transmitting packets, as shown in Table I, $\sum_{i=2}^{4} x_i(\lambda_4^{max}) = 1.136$ (Mbps) < 1.5 (Mbps), where λ_4^{max} is the smallest maximum willingness to pay among those of users that remain in the system. Thus, we can have $\lambda^* < \lambda_4^{max}$ such that $\sum_{i=2}^{4} x_i(\lambda^*) = 1.5$ (Mbps). This satisfies the condition in Proposition 1 for the remaining users and the algorithm converges to the global optimal rate allocation for the remaining users. In this case, the aggregate transmission data rate for users converges to the capacity of the link (1.5 Mbps). Thus, as shown in Figs. 4(b) and 5(b), the transmission data rate of each user converges and the received rate of each user is almost same as its transmission data rate. This implies that with the "self-regulating" property, the system stabilizes and congestion is alleviated.

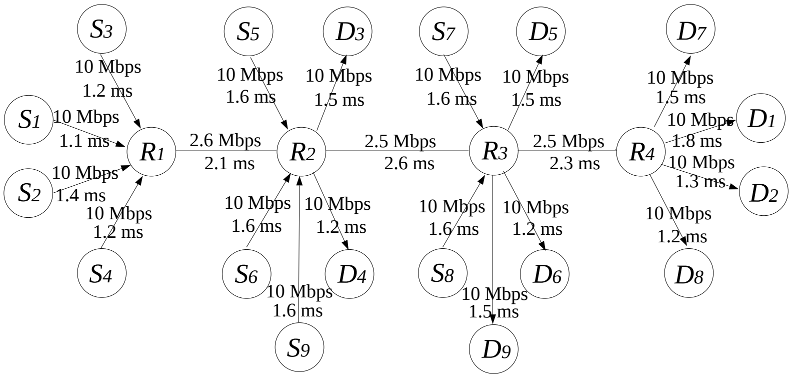


Fig. 7. A system with multiple bottleneck links.

TABLE II

PARAMETERS FOR USERS (MULTIPLE BOTTLENECK LINKS)

User ID	Type	a_i	b_i	M_i	th_i	δ_i	λ_i^{max}
ODD	Sigmoid	15	0.6	1	20	0	1.210
EVEN	Log	50	1	1	20	0	12.717

B. A System with Multiple Bottleneck Links

We now consider a system with multiple bottleneck links, as shown in Fig. 7. Each user i transmits packets from source node S_i to destination node D_i. If the user ID is an odd number, the user has a sigmoid utility function given by (11) and, otherwise, it has a log utility function given by (12). The parameters of the utility functions are provided in Table II. Users from 1 to 8 arrive at the system at time 0 sec and user 9 arrives at the system at time 50 sec. We plot the transmission data rate, the received data rate, and the net utility of each user from 1 to 8 in Figs. 8, 9, and 10, respectively. In Fig. 11, the price of each link is provided. We call a link between nodes R_l and R_{l+1} link l.

As shown in the figures, before user 9 arrives at the system (i.e., before 50 sec), the system can accommodate all users and the transmission data rate of each user converges without congestion. In this case, each link has the same demand for rate allocation, since each link has the same type of users. Hence, as shown in Fig. 11, the prices for links 2 and 3 converge to the same value, since links 2 and 3 have the same capacity. However, since link 1 has a larger capacity than links 2 and 3, the price of link

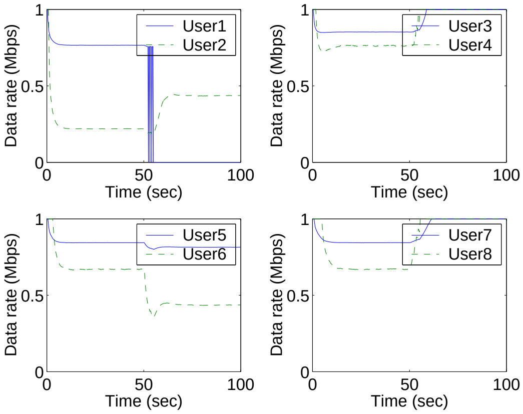


Fig. 8. Transmission data rate (multiple bottleneck links).

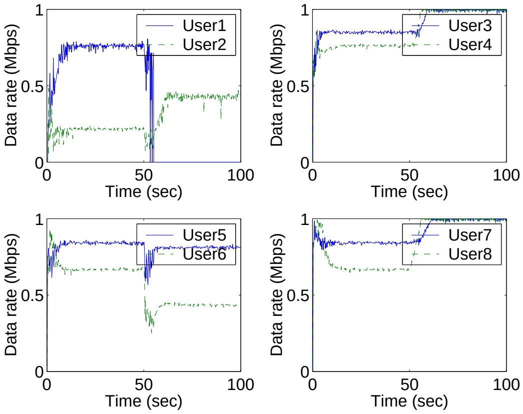


Fig. 9. Received data rate (multiple bottleneck links).

1 converges to a lower value than that for links 2 and 3.

After user 9 arrives to the system (i.e., after 50 sec), as shown in Figs. 8 and 10, the transmission data rate of user 1 begins to oscillate and user 1 obtains negative net utility. Hence, by the "self-regulating" property, it stops transmitting packets. In this case, all users with odd IDs have the same maximum willingness to pay, since they have the same utility function. However, since user 1 uses all three links while the others use only one link, the former must pay a higher price than the latter. Thus, as shown in Fig. 10, only user 1 achieves negative net utility and it stops transmitting packets due to the "self-regulating" property, even though all users with odd IDs have the same maximum willingness to pay. After user 1 stops transmitting packets, the transmission data rate for each user converges without congestion. In this case, link 2 has a larger demand for rate allocation than link 3, since link 2 has an

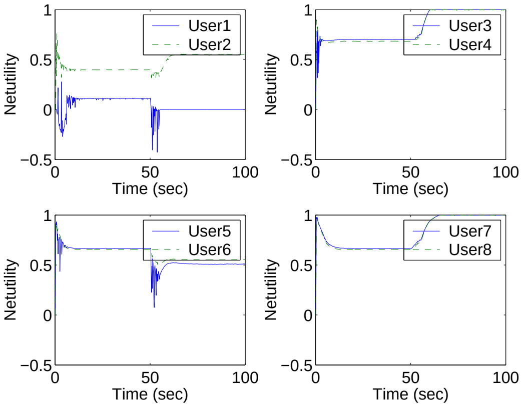


Fig. 10. Net utility (multiple bottleneck links).

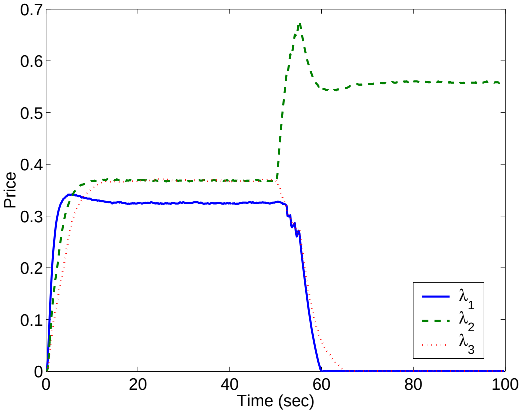


Fig. 11. Price (multiple bottleneck links).

additional user 9 compared with link 3. Hence, as shown in Fig. 11, the price of link 2 converges to a higher value than that of link 3, even though they have the same capacity.

C. Discussion

In the results, we note that a user (user 1 in both systems) that has already been in the system stops transmitting packets due to the arrival of a new user (user 4 in the system with a single bottleneck link and user 9 in the system with multiple bottleneck links). It may be undesirable to interrupt existing services.[2] However, recall that, in this paper, we consider the situation that is similar to the current

[2] This happens because of the property of utility and pricing based algorithms. Hence, this may happen even in the system in which all users have concave utility functions, if users do not have the minimum rate that must be guaranteed or their maximum willingness to pays are not infinity.

Internet in which the system does not have a central authority for call admission control and rate control, and a user adjusts its transmission data rate according to a congestion indicator from the system without considering the other users. In such a situation, as shown by the results, by continuing to transmit packets, user 1 has negative net utility value as well as a large number of packet losses that might be unsatisfactory to the service. Therefore, it may be beneficial not only to the other users but also to user 1 itself for it to stop transmitting packets. User 1 may restart its transmission after some random time or find an alternative route.[3]

The results also tell us the following. First, a service with a concave utility function can be better adapted to congestion on the link than a service with a sigmoidal-like utility function. The former can adjust its transmission data rate gradually according to the congestion level on the link, while the latter can adjust its transmission data rate gradually only up to a certain level. Further, the former has a higher degree of adaptation to the level of the congestion than the latter. This implies that by modeling traditional data services with concave utility functions and real-time streaming services as sigmoidal-like utility functions, we can exploit the characteristics of each service appropriately.

Second, from the viewpoint of the pricing, if a real-time service with a sigmoidal-like utility function wants to have a higher priority to be served than a data service with a concave utility function, it must have a higher maximum willingness to pay than the data service. In this case, in general, the real-time service pays more for the service than the data service, since real-time service keep transmitting data even though the data services stop transmitting because of the high price. This implies that the real-time service must be more expensive than the data service.

Thirdly, if there is no call admission control in the system, when a new service enters into the network, it may be inevitable to interrupt existing services to preserve the system efficiency without incurring congestion. Hence, to prevent this from happening, the system should have an appropriate, preferably distributed, call admission control that admits a new service if it does not interrupt existing (real-time) services.

Finally, as a by-product of rate control, the price of each link represents its supply and demand relationship (i.e., its congestion level). Hence, the price of each link can be used as a parameter for a

[3] Finding a good strategy for this will be a topic for future research.

QoS routing scheme.

V. Conclusion

In this paper, we have studied the distributed rate control algorithm by considering both sigmoidal-like and concave utility functions. We have shown that in the presence of sigmoidal-like utility functions, an algorithm that converges to an efficient rate allocation for a system with only concave utility functions, may not converge, exhibiting oscillatory behavior. Further, such algorithms may result in excessive congestion within the network. This implies that rate control algorithms that have been developed only for concave functions might be inefficient in more realistic settings. To overcome these difficulties, we have developed a distributed algorithm where each user has a "self-regulating" property. Our algorithm works for both sigmoidal-like and concave utility functions. We have shown that our algorithm converges to the asymptotically optimal rate allocation and that its complexity is comparable to that of algorithms developed only for concave utility functions. In this paper, we assume that there exist only controllable services within the network. However, in general, there also exist uncontrollable services within the network that may further affect the efficiency and the convergence of the rate control algorithm. Hence, as shown in [18], the study of the effect of uncontrollable services to rate control is important and will be a topic for future research.

Appendix

A. Proof of Proposition 1

If $Q(\bar{\lambda})$ is differentiable at $\bar{\lambda}^o$, $Q(\bar{\lambda})$ has a unique subgradient at $\bar{\lambda}^o$ by (S1). Hence, by (5), $\bar{x}(\bar{\lambda}^o)$ is unique and, thus, by (R1), $\bar{x}(\bar{\lambda})$ is continuous at $\bar{\lambda}^o$. This implies that $\bar{x}(\bar{\lambda}^{(n)})$ converges to $\bar{x}(\bar{\lambda}^o)$, since $\bar{\lambda}^{(n)}$ converges to $\bar{\lambda}^o$. Further, since $\bar{x}(\bar{\lambda}^o)$ is a unique maximum in $\bar{x}$ of $L(\bar{x}, \bar{\lambda}^o)$ in (2), by property 6.5 in [15], the primal problem (A) has a saddle point $(\bar{x}(\bar{\lambda}^o), \bar{\lambda}^o)$ and $\bar{x}(\bar{\lambda}^o)$ is an optimal solution of the primal problem (A).

If $Q(\bar{\lambda})$ is not differentiable at $\bar{\lambda}^o$, the subgradient of $Q(\bar{\lambda})$ at $\bar{\lambda}^o$ is not unique by (S1). Hence, by (5), there exists a user i such that $x_i^o > 0$ and $\lambda_i^{max} = \lambda_{T(i)}^o$. In this case, by (R1), $x_i(\lambda_{T(i)})$ is discontinuous at $\lambda_{T(i)}^o$. Hence, $x_i(\lambda_{T(i)}^{(n)})$ (i.e., $\bar{x}(\bar{\lambda}^{(n)})$) may not converge, even though $\bar{\lambda}^{(n)}$ converges to $\bar{\lambda}^o$.

B. Proof of Proposition 2

We first prove the following lemma.

Lemma 1: Suppose that $\bar{\lambda}^o$ *is an optimal solution of the dual problem (B) and* $\lambda^o_l > 0$ *for some link* l*. Then, there exists a subgradient of* $Q(\bar{\lambda})$, $\bar{d}(\bar{\lambda}^o)$, *at* $\bar{\lambda}^o$ *such that* $d_l(\bar{\lambda}^o) = 0$.

Proof: Since $\bar{\lambda}^o$ is a minimizer of $Q(\bar{\lambda})$, by (S2), there exists a subgradient of $Q(\bar{\lambda})$, $\bar{d}(\bar{\lambda}^o)$, at $\bar{\lambda}^o$ such that

$$\bar{d}(\bar{\lambda}^o)^T(\bar{\lambda} - \bar{\lambda}^o) \geq 0 \quad \text{for all } \bar{\lambda} \geq \bar{0}. \tag{13}$$

If we take $\bar{\lambda} = \bar{\lambda}'$, where $\lambda'_i = \lambda^o_i$, $i \neq l$ and $\lambda'_l = \lambda^o_l + \epsilon$, $\epsilon > 0$. Then, by (13), we have $d_l(\bar{\lambda}^o)\epsilon \geq 0$ and, thus, $d_l(\bar{\lambda}^o) \geq 0$. In a similar way, by taking $\epsilon < 0$, we have $d_l(\bar{\lambda}^o) \leq 0$. Hence, $d_l(\bar{\lambda}^o) = 0$. ■

We now prove Proposition 2. If $Q(\bar{\lambda})$ is not differentiable at $\bar{\lambda}^o$, from the proof of Proposition 1 in Appendix A, there exist a user i^* such that $x^o_{i^*} > 0$ and $\lambda^{max}_{i^*} = \lambda^o_{T(i^*)}$. Hence, $x_{i^*}(\lambda_{T(i^*)})$ is discontinuous at $\lambda^o_{T(i^*)}$ by Property (R1). Further, since $\lambda^{max}_{i^*} > 0$, $\lambda^o_{T(i^*)} = \sum_{l \in T(i^*)} \lambda^o_l > 0$ and, thus, there exists a link $l^* \in T(i^*)$ such that $\lambda^o_{l^*} > 0$. In this case, by Lemma 1, there exists a subgradient of $Q(\bar{\lambda})$, $\bar{d}(\bar{\lambda}^o)$, at $\bar{\lambda}^o$, such that $d_{l^*}(\bar{\lambda}^o) = 0$. Hence, from (5) and the fact that $x_{i^*}(\lambda_{T(i^*)})$ is discontinuous at $\lambda^o_{T(i)}$, one of the following conditions is satisfied at link l^*:

$$\begin{aligned}
\sum_{i \in S^H(l^*, \bar{\lambda}^o)} x_i(\lambda^o_{T(i)}) < C_{l^*} - \epsilon_1 \quad &\text{and} \quad \sum_{i \in S^H(l^*, \bar{\lambda}^o) \cup S^S(l, \bar{\lambda}^o)} x_i(\lambda^o_{T(i)}) > C_{l^*} + \epsilon_2, \\
\sum_{i \in S^H(l^*, \bar{\lambda}^o)} x_i(\lambda^o_{T(i)}) \leq C_{l^*} \quad &\text{and} \quad \sum_{i \in S^H(l^*, \bar{\lambda}^o) \cup S^S(l^*, \bar{\lambda}^o)} x_i(\lambda^o_{T(i)}) > C_{l^*} + \epsilon_3, \text{ or} \\
\sum_{i \in S^H(l^*, \bar{\lambda}^o)} x_i(\lambda^o_{T(i)}) < C_{l^*} - \epsilon_4 \quad &\text{and} \quad \sum_{i \in S^H(l^*, \bar{\lambda}^o) \cup S^S(l^*, \bar{\lambda}^o)} x_i(\lambda^o_{T(i)}) \geq C_{l^*},
\end{aligned} \tag{14}$$

where ϵ_1, ϵ_2, ϵ_3, and ϵ_4 are some positive constants.

C. Proof of Proposition 3

We first define the maximum net utility of user i at the price $\bar{\lambda}$ (i.e., at $\lambda_{T(i)}$) as

$$NU^{max}_i(\lambda_{T(i)}) \;\; = \;\; \max_{0 \leq r \leq M_i} \{U_i(r) - \lambda_{T(i)} r\}.$$

Then, to prove the proposition, we only have to show that

$$\limsup_{n\to\infty} NU_i^{max}(\lambda_{T(i)}^{(n)}) = 0,\ i \in S^S(l^*, \bar{\lambda}^o).$$

Since $NU_i^{max}(\lambda_{T(i)})$ is a continuous function of $\lambda_{T(i)}$ and $\bar{\lambda}^{(n)}$ converges to $\bar{\lambda}^o$, $NU_i^{max}(\lambda_{T(i)}^{(n)})$ converges to $NU_i^{max}(\lambda_{T(i)}^o)$. Further, since $NU_i(\lambda_{T(i)}^{(n)}, x_i'(\lambda_{T(i)}^{(n)})) \leq NU_i(\lambda_{T(i)}^{(n)}, x_i(\lambda_{T(i)}^{(n)})) = NU_i^{max}(\lambda_{T(i)}^{(n)})$ and $NU_i^{max}(\lambda_{T(i)}^o) = 0$, for $i \in S^S(l^*, \lambda^o)$,

$$\limsup_{n\to\infty} NU_i(\lambda_{T(i)}^{(n)}, x_i'(\lambda_{T(i)}^{(n)})) \leq \limsup_{n\to\infty} NU_i(\lambda_{T(i)}^{(n)}, x_i(\lambda_{T(i)}^{(n)})) = \limsup_{n\to\infty} NU_i^{max}(\lambda_{T(i)}^{(n)}) = 0,\ i \in S^S(l^*, \bar{\lambda}^o).$$

D. Proof of Proposition 4

We first divide users into three subsets at our rate allocation: R is a subset of users that keep transmitting data, R^s is a subset of users that stop transmitting data due to the "self-regulating" property, and R^c is a subset of users that stop transmitting data due to higher prices than their maximum willingness to pays.

By the weak duality theorem [16],

$$\begin{aligned}
& Q(\bar{\lambda}^o) - \sum_{i=1}^N U_i(x_i^o) \\
= \quad & \sum_{i=1}^N U_i(x_i(\lambda_{T(i)}^o)) + \bar{\lambda}^{oT}(\bar{C} - A\bar{x}(\bar{\lambda}^o)) - \sum_{i=1}^N U_i(x_i^o) \\
= \quad & \sum_{i\in R} U_i(x_i(\lambda_{T(i)}^o)) + \sum_{i\in R^c} U_i(x_i(\lambda_{T(i)}^o)) + \sum_{i\in R^s} U_i(x_i(\lambda_{T(i)}^o)) + \bar{\lambda}^{oT}(\bar{C} - A\bar{x}(\bar{\lambda}^o)) - \sum_{i=1}^N U_i(x_i^o) \\
\geq \quad & 0.
\end{aligned}$$

Since $\sum_{i\in R^c} U_i(x_i(\lambda_{T(i)}^o)) = 0$ and $\bar{\lambda}^{oT}(\bar{C} - A\bar{x}(\bar{\lambda}^o)) \leq 0$,

$$\sum_{i\in R} U_i(x_i(\lambda_{T(i)}^o)) \geq \sum_{i=1}^N U_i(x_i^o) - \sum_{i\in R^s} U_i(x_i(\lambda_{T(i)}^o)).$$

Further, since our rate allocation is a global optimal rate allocation for the remaining users,

$$\sum_{i=1}^N U_i(x_i^*) = \sum_{i\in R} U_i(x_i^*) \geq \sum_{i\in R} U_i(x_i(\lambda_{T(i)}^o)) \geq \sum_{i=1}^N U_i(x_i^o) - \sum_{i\in R^s} U_i(x_i(\lambda_{T(i)}^o))$$

and

$$\frac{\sum_{i=1}^N U_i(x_i^*)}{\sum_{i=1}^N U_i(x_i^o)} \geq 1 - \frac{\sum_{i\in R^s} U_i(x_i(\lambda_{T(i)}^o))}{\sum_{i=1}^N U_i(x_i^o)}.$$

Since $\sum_{i=1}^{N} U_i(x_i^*) \leq \sum_{i=1}^{N} U_i(x_i^o)$ and we assume that $\frac{\sum_{i \in R^s} U_i(x_i(\lambda^o_{T(i)}))}{\sum_{i=1}^{N} U_i(x_i^o)} \rightarrow 0$ as $N \rightarrow \infty$,

$$\frac{\sum_{i=1}^{N} U_i(x_i^*)}{\sum_{i=1}^{N} U_i(x_i^o)} \rightarrow 1, \text{ as } N \rightarrow \infty.$$

E. Proof of Proposition 5

Before we prove Proposition 5, we first prove the following two lemmas.

Lemma 2: Suppose that $\bar{\lambda}^o$ is a dual optimal solution. Then, for any $\epsilon > 0$, there exists $\alpha_\epsilon > 0$ such that $||\bar{\lambda}^{(m)} - \bar{\lambda}^o|| \leq \epsilon$ for some m by solving (3) and (7) with a constant step size $0 < \alpha \leq \alpha_\epsilon$.

Proof: This can be proved in a similar way to the proof of Theorem 2.1 in [14]. ■

Lemma 3: Suppose that $\bar{\lambda}^o$ is a dual optimal solution and at iteration n_1, $||\bar{\lambda}^{(n_1)} - \bar{\lambda}^o|| \leq \epsilon_1$. Then, for any $\epsilon > \epsilon_1$ and m, there exists $\alpha_{\epsilon,m} > 0$ such that $||\bar{\lambda}^{(n)} - \bar{\lambda}^o|| \leq \epsilon$ at least for m consecutive iterations after iteration n_1 by solving (3) and (7) with a constant step size $0 < \alpha \leq \alpha_{\epsilon,m}$.

Proof: Let us define

$$r \;=\; \max\{||\bar{C} - A\bar{x}(\bar{\lambda})|| \;|\; ||\bar{\lambda} - \bar{\lambda}^o|| \leq \epsilon, \ \bar{\lambda} \geq \bar{0}\}.$$

Suppose that we have a constant step size $\alpha_{\epsilon,m}$ that satisfies the following inequality:

$$\begin{aligned} & \epsilon_1^2 + m(\alpha_{\epsilon,m}^2 r^2 + 2\alpha_{\epsilon,m}\epsilon r) - \epsilon^2 \\ = \;& m r^2 \alpha_{\epsilon,m}^2 + 2m\epsilon r \alpha_{\epsilon,m} \epsilon_1^2 + \epsilon_1^2 - \epsilon^2 \\ \leq \;& 0. \end{aligned}$$

Since $\epsilon_1^2 - \epsilon^2 < 0$, there exists an $\alpha_{\epsilon,m}$ that satisfies the above inequality. Then,

$$\begin{aligned} ||\bar{\lambda}^{(n_1+1)} - \bar{\lambda}^o||^2 &= ||[\bar{\lambda}^{(n_1)} - \alpha_{\epsilon,m}(\bar{C} - A\bar{x}(\bar{\lambda}^{(n_1)}))]^+ - \bar{\lambda}^o||^2 \\ &\leq ||\bar{\lambda}^{(n_1)} - \alpha_{\epsilon,m}(\bar{C} - A\bar{x}(\bar{\lambda}^{(n_1)})) - \bar{\lambda}^o||^2 \\ &= ||\bar{\lambda}^{(n_1)} - \bar{\lambda}^o||^2 + \alpha_{\epsilon,m}^2 ||\bar{C} - A\bar{x}(\bar{\lambda}^{(n_1)})||^2 - 2\alpha_{\epsilon,m}(\bar{\lambda}^{(n_1)} - \bar{\lambda}^o)^T(\bar{C} - A\bar{x}(\bar{\lambda}^{(n_1)})) \\ &\leq ||\bar{\lambda}^{(n_1)} - \bar{\lambda}^o||^2 + \alpha_{\epsilon,m}^2 r^2 + 2\alpha_{\epsilon,m}\epsilon r \\ &\leq \epsilon_1^2 + \alpha_{\epsilon,m}^2 r^2 + 2\alpha_{\epsilon,m}\epsilon r \\ &\leq \epsilon^2. \end{aligned}$$

Hence, $||\bar{\lambda}^{(n_1+1)} - \bar{\lambda}^o|| \leq \epsilon$. In a similar way, we can show that

$$\begin{aligned} ||\bar{\lambda}^{(n_1+k)} - \bar{\lambda}^o||^2 &\leq \epsilon_1^2 + k(\alpha_{\epsilon,m}^2 r^2 + 2\alpha_{\epsilon,m}\epsilon r) \\ &\leq \epsilon,\ k = 1, 2, \cdots, m. \end{aligned}$$

Therefore, $||\bar{\lambda}^{(n)} - \bar{\lambda}^o|| \leq \epsilon$ at least for m consecutive iterations after iteration n_1. ■

We now prove Proposition 5 considering two cases.

Case1: Suppose that the condition in Proposition 1 is satisfied, i.e., at the dual optimal solution, $\bar{\lambda}^o$, $Q(\bar{\lambda})$ is differentiable. Since $Q(\bar{\lambda})$ is differentiable almost everywhere and it is differentiable at $\bar{\lambda}^o$, there exists an $\epsilon_1 > 0$ such that, for all $\bar{\lambda}$ that satisfies $||\bar{\lambda} - \bar{\lambda}^o|| \leq \epsilon_1$, $Q(\bar{\lambda})$ is differentiable. Let us define

$$d = \max_{\{\bar{\lambda} \mid ||\bar{\lambda} - \bar{\lambda}^o|| > \epsilon_1\}} \{Q(\bar{\lambda})\}.$$

Then, since $Q(\bar{\lambda})$ is a convex function, there exists an $\epsilon \leq \epsilon_1$ such that $Q(\bar{\lambda}) \leq d$ for all $\bar{\lambda}$ that satisfies $||\bar{\lambda} - \bar{\lambda}^o|| \leq \epsilon$. Hence, for $\bar{\lambda}$ that satisfies $||\bar{\lambda} - \bar{\lambda}^o|| \leq \epsilon$, the subgradient projection algorithm is equivalent to the gradient projection algorithm, since $Q(\bar{\lambda})$ is differentiable for all $\bar{\lambda}$ that satisfies $||\bar{\lambda} - \bar{\lambda}^o|| \leq \epsilon$. With the assumption (U5), there exists a constant α_1 that makes the gradient projection algorithm converge with a descent property [4], [5]. Hence, once $||\bar{\lambda}^{(k)} - \bar{\lambda}^o|| \leq \epsilon$ for some iteration k, there exists a constant α_1 such that $||\bar{\lambda}^{(n)} - \bar{\lambda}^o|| \leq \epsilon$ for all $n \geq k$ and $\bar{\lambda}^{(n)}$ converges to $\bar{\lambda}^o$. Further, by Lemma 2, for any $\epsilon > 0$, there exists a constant α_2 that makes $||\bar{\lambda}^{(k)} - \bar{\lambda}^o|| \leq \epsilon$ for some k. Hence, by taking $\alpha = \min\{\alpha_1, \alpha_2\}$, $\bar{\lambda}^{(n)}$ converges to $\bar{\lambda}^o$. Since $Q(\bar{\lambda})$ is differentiable at $\bar{\lambda}^o$, $\bar{x}(\bar{\lambda}^{(n)})$ converges to $\bar{x}(\bar{\lambda}^o)$ by Proposition 1.

Case2: Suppose that the condition in Proposition 1 is not satisfied. Then, there exist an ϵ and an m such that when $||\bar{\lambda}^{(n)} - \bar{\lambda}^o|| \leq \epsilon$ for m consecutive iterations, the condition in Proposition 1 is satisfied for the remaining users after some users stop transmitting data due to the "self-regulating" property. By Lemmas 2 and 3, there exists a constant α_1 with which $||\bar{\lambda}^{(n)} - \bar{\lambda}^o|| \leq \epsilon$ for m consecutive iterations. After that, since the condition in Proposition 1 is satisfied for the remaining users, as in Case 1, there exists a constant α_2 that makes the algorithm for the remaining users converge. Hence, by taking $\alpha = \min\{\alpha_1, \alpha_2\}$, the algorithm converges.

References

[1] S. Shenker, “Fundamental design issues for the future Internet,” *IEEE journal on selected area in communications*, vol. 13, no. 7, pp. 1176–1188, Sept. 1995.

[2] F. P. Kelly, “Charging and rate control for elastic traffic,” *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, Jan. 1997.

[3] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, “Rate control in communication networks: shadow prices, proportional fairness and stability,” *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, Mar. 1998.

[4] H. Yäiche, R. R. Mazumdar, and C. Rosenberg, “A game theoretic framework for bandwidth allocation and pricing of elastic connections in broadband networks: theory and algorithms,” *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 667–678, Oct. 2000.

[5] S. H. Low and D. E. Lapsley, “Optimization flow control-I: basic algorithm and convergence,” *IEEE/ACM Transactions on Networking*, vol. 7, no. 6, pp. 861–874, Dec. 1999.

[6] S. Athuraliya and S. H. Low. Optimization flow control-II: Implementation. Submitted for publication. [Online]. Available: http://netlab.caltech.edu

[7] S. Kunniyur and R. Srikant, “End-to-end congestion control schemes: utility function, random losses and ECN marks,” in *IEEE Infocom’00*, vol. 3, 2000, pp. 1323–1332.

[8] R. J. La and V. Anantharam, “Utility-based rate control in the Internet for elastic traffic,” *IEEE/ACM Transactions on Networking*, vol. 10, no. 2, pp. 272–286, Apr. 2002.

[9] K. Kar, S. Sarkar, and L. Tassiulas, “A simple rate control algorithm for maximizing total user utility,” in *IEEE Infocom’01*, vol. 1, 2001, pp. 133–141.

[10] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, “Non-convex optimization and distributed pricing-based algorithms for optimal resource allocation in high speed networks,” in *17th IEEE Annual Computer Communications Workshop*, 2002.

[11] ——, “Non-convexity issues for Internet rate control with multi-class services: stability and optimality,” to appear in the proceedings of *IEEE Infocom’04*, 2004.

[12] S. Stidham. Pricing and congestion management in a network with heterogeneous users. Submitted for publication. [Online]. Available: http://www.or.unc.edu/∼sandy

[13] J. W. Lee, R. R. Mazumdar, and N. B. Shroff, "Downlink power allocation for multi-class CDMA wireless networks," in *IEEE Infocom'02*, vol. 3, 2002, pp. 1480–1489.

[14] N. Z. Shor, *Minimization methods for non-differentiable functions*. Springer-Verlag, 1985.

[15] M. Minoux, *Mathematical programming:theory and algorithms*. Wiley, 1986.

[16] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[17] K. Kar, S. Sarkar, and L. Tassiulas, "Optimization based rate control for multirate multicast sessions," in *IEEE Infocom'01*, vol. 1, 2001, pp. 123–132.

[18] D. Qiu and N. B. Shroff, "Queueing properties of feedback flow control systems," to appear in the *IEEE/ACM Transactions on Networking*, Dec. 2004.