# Distributed Rate Allocation for Inelastic Flows: Optimization Frameworks, Optimality Conditions, and Optimal Algorithms

Mung Chiang, Shengyu Zhang, and Prashanth Hande
EE and CS Departments, Princeton University, NJ 08544

*Abstract*—**A common assumption behind most of the recent research on network utility maximization is that traffic flows are elastic, which implies that their utility functions are concave and there are no hard limits on the rate allocated to each flow. These critical assumptions lead to tractability of the analytic models of utility maximization, but also limits applicability of the resulting rate allocation protocols. This paper focuses on *inelastic flows* and removes these restrictive and often invalid assumptions. We present several optimization frameworks, optimality conditions, and optimal algorithms.**

**First, we consider nonconcave utility functions, which turn utility maximization into nonconvex, constrained optimization problems that are well-known to be difficult. We present conditions under which the current standard price-based distributed algorithm can still converge to the globally optimal rate allocation despite nonconcavity of utility functions. In particular, continuity of price-based rate allocation at all the optimal prices is a sufficient condition for global convergence of rate allocation by the standard algorithm, and continuity at at least one optimal price is a necessary condition. In the second part of the paper, we provide a general problem formulation of rate allocation among time-sensitive flows from real-time and streaming applications, as well as a decomposition into subproblems coordinated by pricing. After simplifying the subproblems by leveraging the optimization structures, we highlight the difficult issues of causality and time-scale, and propose an effective price-based heuristics for admission control and an optimal algorithm for a special case formulation.**

**Keywords:** Optimization, Congestion control, Network control by pricing, Resource allocation, Network utility maximization.

## I. Introduction and Review

### A. Overview

In the seminal paper published a decade ago, Shenker [13] discussed Internet service models to support applications beyond best-effort, in the framework of network utility models for four types of traffic. In particular, two major characteristics of 'elastic traffic' were highlighted: "These applications are rather elastic in nature, in that they tolerate packet delays and packet losses rather gracefully... Moreover, because of this elasticity, they can decrease their transmission rate in the presence of congestion." Utility functions for elastic traffic were modelled as smooth, concave functions of data rates, which lead to the conclusion in [13] that network utility is always maximized when no users are denied access. Maximization of concave utility functions and distributed rate allocation for elastic traffic have gained extensive attention over the last decade. Elegant analytic results and rigorous mathematical frameworks for a 'canonical' price-based distributed algorithm become possible because of the concavity assumption on utility functions and the elasticity assumption on application traffic.

However, it may appear that the other classes of *inelastic* traffic in [13] are not amenable to the utility maximization framework of rate allocation. In addition to real-time traffic, delay-adaptive traffic, and rate-adaptive traffic briefly described in [13], streaming applications have also become a major traffic class on the Internet. Utility functions for these four types of inelastic flows are nonconcave or non-smooth, and they can only tolerate a limited amount of packet delay or fluctuations in rate allocation transients. Furthermore, when there is a mixture of these inelastic flows with the elastic data flows, network utility may be maximized only with the help of admission control.

While inelastic flows represent important applications, their rate allocation methods scarcely have any mathematical foundation because of the intrinsic intractability in the utility maximization framework. This paper shows how some of the technical difficulties may be tackled, and presents a series of results on new optimization frameworks, optimality conditions, and optimal algorithms for utility maximization of inelastic traffic.

In the first part of this paper, consisting of section II and the appendices, we tackle the issue of nonconcavity of utility functions. Complementary to the recent approach of proposing suboptimal distributed heuristics [9] and that of calculating optimal solution by centralized computation [7], we answer the question: "Can the canonical distributed algorithm converge to the globally optimal rate allocation even if source utility functions are nonconcave?" Surprisingly, the answer is positive, and several conditions for global convergence are proved and illustrated.

In the second part of this paper, consisting of sections III and IV, we tackle the issue of time-sensitivity of the flows. After presenting an optimization formulation and problem decomposition, we highlight several technical difficulties that hinder the development of distributed solutions. Despite these difficulties, for some special cases of the general formulation, we develop a price-based admission control heuristics and evaluate its performance, and present a globally optimal algorithm for rate allocation among groups of flows with different elasticities.

Some of the results in this paper are applicable to general nonconvex optimization with a separable objective function and linear constraints. However, this paper does not treat the modeling aspect of constructing utility functions (concave or otherwise) from empirical data of user experience. The theory and algorithms developed here focus on equilibrium behaviors at fluid level, but not stochastic stability or behaviors at packet level.

In the rest of this section, we briefly review the current framework of concave network utility maximization. The difficulties associated with nonconcave utility maximization and time-sensitive flow modeling are then discussed.

### B. *Concave network utility maximization*

Since the publication of the seminal paper [8] by Kelly, Maulloo, and Tan in 1998, the framework of Network Utility Maximization (NUM) has found many applications in network rate allocation algorithms, Internet congestion control protocols, user behavior models, and network efficiency-fairness characterization. Consider a communication network with $L$ links, each with a fixed capacity of $c_l$ bps, and $S$ sources, each transmitting at a source rate of $x_s$ bps. Each source emits one flow, using a fixed set $L(s)$ of links in its path, and has a utility function $U_s(x_s)$. The basic version of NUM is the problem of maximizing the total network utility $\sum_s U_s(x_s)$, over the source rates $\mathbf{x}$, subject to linear flow feasibility constraints $\sum_{s:l\in L(s)} x_s \le c_l$ for all links $l$:

$$\begin{array}{ll} \text{maximize} & \sum_s U_s(x_s) \\ \text{subject to} & \sum_{s:l\in L(s)} x_s \le c_l, \quad \forall l, \\ & \mathbf{x} \succeq 0 \end{array} \tag{1}$$

where the variables are $\mathbf{x}$. Among many of its applications, this optimization problem has been extensively studied as a model for distributed rate allocation (*e.g.*, [8]) and TCP congestion control (*e.g.*, [10]). In this paper, we are primarily concerned with extensions of the basic NUM problem (1) for nonconcave or discontinuous utility functions $\{U_s\}$.

The following basic assumption on utility functions will still be maintained in this paper: **Assumption 1.** utilities are functions of the allowed rates (rate-dependency), network utility is the sum of source utilities (additivity), and each source utility is an increasing (monotonicity) and local function of its own rate (locality).

Assuming that $U_s(x_s)$ becomes concave for large enough $x_s$ is reasonable, because the law of diminishing marginal utility eventually will be effective. However, $U_s$ may not be concave throughout its domain. Despite deficiency in the concavity assumption, almost all papers in the NUM literature for Internet rate allocation assume that utility functions are concave. Part of the reason is that the concavity assumption significantly simplifies the structure of the basic NUM problem (1) and leads to a distributed rate allocation algorithm.

### C. *Dual problem and canonical distributed algorithm*

Assuming utility functions are concave, (1) is maximizing a concave function over linear constraints, which is a special case of convex optimization (minimizing a convex objective function over convex constraints) [3] called monotropic programming [11]. Thus a local optimum is also a global optimum, and the duality gap is zero. [1] Zero duality gap means that the minimized objective value of the Lagrange dual problem is equal to the maximized total utility in the primal problem (1).

The Lagrange dual problem is readily derived. We first form the Lagrangian of (1):

$$L(\mathbf{x},\boldsymbol{\lambda}) = \sum_s U_s(x_s) + \sum_l \lambda_l \left( c_l - \sum_{s:l\in L(s)} x_s \right)$$

where $\lambda_l \ge 0$ is the Lagrange multiplier (*i.e.*, link price) associated with the linear flow constraint on link $l$. Additivity of total utility and linearity of flow constraints lead to a Lagrangian dual decomposition into individual source terms:

$$\begin{aligned} L(\mathbf{x},\boldsymbol{\lambda}) &= \sum_s \left[ U_s(x_s) - \left( \sum_{l\in L(s)} \lambda_l \right) x_s \right] + \sum_l c_l \lambda_l \\ &= \sum_s L_s(x_s, \lambda^s) + \sum_l c_l \lambda_l \end{aligned}$$

where $\lambda^s = \sum_{l\in L(s)} \lambda_l$. For each source $s$, $L_s(x_s,\lambda^s) = U_s(x_s) - \lambda^s x_s$ only depends on local rate $x_s$ and the path price $\lambda^s$ (*i.e.*, sum of $\lambda_l$ on links used by source $s$).

The Lagrange dual function $g(\boldsymbol{\lambda})$ is defined as the maximized $L(\mathbf{x},\boldsymbol{\lambda})$ over $\mathbf{x}$ for a given $\boldsymbol{\lambda}$. This 'net utility' maximization [2] obviously can be conducted distributively by the each source, as long as the aggregate link price $\lambda^s$ is feedback to source $s$, where source $s$ maximizes $L_s(x_s,\lambda^s)$ over $x_s$ for a given $\lambda^s$:

$$x_s^*(\lambda^s) = \text{argmax}\left[ U_s(x_s) - \lambda^s x_s \right], \quad \forall s. \tag{2}$$

Such Lagrangian maximizer $\mathbf{x}^*(\boldsymbol{\lambda})$ will be referred to as price-based rate allocation (for a given price $\boldsymbol{\lambda}$). When the concavity assumption on utility functions is removed, there can be multiple $\mathbf{x}^*(\boldsymbol{\lambda})$. In such cases, when we say that a property (*e.g.*, continuity) holds for $\mathbf{x}^*(\boldsymbol{\lambda})$, it means that it holds no matter which one of the multiple possible values is chosen.

The Lagrange dual problem of (1) is

$$\begin{array}{ll} \text{minimize} & g(\boldsymbol{\lambda}) = L(\mathbf{x}^*(\boldsymbol{\lambda}),\boldsymbol{\lambda}) \\ \text{subject to} & \boldsymbol{\lambda} \succeq 0 \end{array} \tag{3}$$

where the optimization variable is $\boldsymbol{\lambda}$. Since $g(\boldsymbol{\lambda})$ is the pointwise supremum of a family of affine functions in $\boldsymbol{\lambda}$, it is convex and (3) is a convex minimization problem (even if the primal problem (1) is not a concave maximization problem).

[1] Duality gap is the difference between the optimized dual objective value and the optimized primal objective value.

[2] Equivalently, computation of the conjugate function of utility function.

Since $g(\boldsymbol{\lambda})$ may be non-differentiable, an iterative *subgradient* method can be used to update the dual variables $\boldsymbol{\lambda}$ to solve the dual problem (3):

$$\lambda_l(t+1) = \left[\lambda_l(t) - \alpha(t)\left(c_l - \sum_{s:l\in L(s)} x_s(\lambda^s(t))\right)\right]^+, \quad \forall l \tag{4}$$

where $c_l - \sum_{s:l\in L(s)} x_s(\lambda^s(t))$ is the $l$th component of a subgradient vector of $g(\boldsymbol{\lambda})$, $t$ is the iteration number, and $\alpha(t) > 0$ are step sizes. Certain choices of step sizes, such as $\alpha(t) = \beta/t,\ \beta > 0$, guarantee that the sequence of dual variables $\boldsymbol{\lambda}(t)$ converges to the dual optimal $\boldsymbol{\lambda}^*$ as $t \to \infty$. It can be shown that the primal variable $\mathbf{x}^*(\boldsymbol{\lambda}(t))$ also converges to the primal optimal variable $\mathbf{x}^*$. For a primal problem that is a convex optimization, the convergence is towards a global optimum.

The following assumption will be made: **Assumption 2**. Rates $x_s$ are implicitly assumed to be upper bounded by finite numbers. Therefore, the global optimum can be attained since (1) is maximizing a continuous function over a compact set.

The sequence of source and link algorithms (2,4) forms a *canonical distributed algorithm* that globally solves NUM (1) and the dual problem (3), and computes an optimal rate vector $\mathbf{x}^*$ and optimal link price vector $\boldsymbol{\lambda}^*$.

### D. Nonconcave network utility maximization

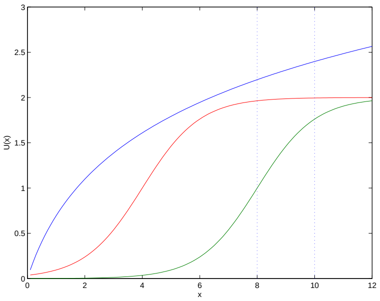

Fig. 1. Some examples of utility functions $U_s(x_s)$: it can be concave or sigmoidal as shown in the graph, or any general nonconcave function. If the bottleneck link capacity used by the source is small enough, *i.e.*, if the dotted vertical line is pushed to the left, a sigmoidal utility function effectively becomes a convex utility function.

Suppose we remove the critical assumption that $\{U_s\}$ are concave functions, and allow them to be any nonlinear functions satisfying Assumption 1. [3] The resulting NUM becomes nonconvex optimization and significantly harder to be analyzed and solved, even by centralized computational methods. In particular, a local optimum may not be a global optimum and the duality gap can be strictly positive. The standard distributive algorithms that solve the dual problem may produce infeasible or suboptimal rate allocation. Global maximization of nonconcave functions is an intrinsically difficult problem of nonconvex optimization. Indeed, over the last two decades, it has been widely recognized that "*in fact the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity*" (Quote from Rockafellar [12]).

It may appear that the canonical distributed algorithm will not converge or will converge to an infeasible or locally optimal rate allocation if some source utilities are nonconcave, because it is based on dual descent and the duality gap can be strictly positive for a nonconvex optimization problem. However, we show in section II that, even when all source utilities are nonconcave functions, the canonical distributed algorithm may still converge to a globally optimal rate allocation, duality gap may still be zero, and it is continuity of price-based rate allocation at all the optimal prices, rather than zero duality gap alone, that provides a sufficient condition for global convergence. For strictly convex (and possibly non-differentiable) $g(\boldsymbol{\lambda})$, this continuity is proved to be equivalent to the convergence of the canonical distributed algorithm.

### E. Rate allocation for time sensitive flows

Most work on NUM-based rate allocation algorithms focus only on the equilibrium properties. However, some inelastic flows cannot tolerate arbitrary transients of rate allocation. For example, real-time IP applications or Constant Bit Rate (CBR) ATM flows must maintain a certain constant rate, which allows them to achieve a constant utility. Utility functions become discontinuous for such flows. Another example is streaming applications, where slight fluctuations of data rate may be acceptable if there is enough buffered packets in the playback buffer at the receiving end host. This type of traffic is more elastic than constant rate flows, but certainly not fully elastic since the playback buffer will be depleted after a long enough period of low data rate. In general, a flow's utility depends not just on the equilibrium rate $x_i^*$, but the entire vector of rate $\{x_i(t)\}$ over time.

In section III, we present a comprehensive formulation of an optimization framework, together with an appropriate duality-based decomposition, to capture the above issues of rate allocation among time-sensitive flows. A heuristics of price-based admission control is then investigated in section IV. We also present an optimal algorithm to solve a special case of the general formulation.

## II. Nonconcave Utility Flows: Optimality Conditions

With nonconcave utility functions, the canonical distributed algorithm may fail to converge to the primal optimal solution $\mathbf{x}^*$. One reason is that solving the dual problem (3) is no longer equivalent to solving the primal problem (1). In the case of allocating rates through NUM, it is the primal problem that we care about. The canonical distributed algorithm may even fail to converge to a feasible rate allocation, as shown by Lee, Mazumdar, and Shroff [9], where they focus on the special case of sigmoidal utility functions, and show that the canonical distributed

[3] Sometimes a nonconcave function can be easily turned into a concave one by a simple transformation, for example in the case of the "pseudo-nonconvexity" in the power control problems in [5], [6]. Here we are concerned with nonconcave functions that cannot be readily turned into concave ones.

algorithm may cause link congestion as well as produce suboptimal rate allocation. A ‘self-regulating’ heuristics is proposed, and is shown to avoid link congestion caused by sigmoidal utilities but does not attain the optimal rate allocation $\mathbf{x}^*$ (except in the asymptotic case when the proportion of sources with nonconcave utilities vanishes). The proof techniques used in [9] depend on the assumption that a sigmoidal utility function has only one inflexion point.

In this section, we study the general case where $\{U_s\}$ are nonconcave functions. The goal of our study is to prove sufficient and necessary conditions under which the canonical distributed algorithm still converges to the globally optimal rate allocation, *i.e.*, the primal optimizer $\mathbf{x}^*$.

In general, these conditions do not hold for nonconcave NUM. In such cases, a centralized computational method based on the sum-of-squares approach has recently been studied [7] and is empirically found to compute the globally optimal rate allocation very efficiently.

These three different approaches: proposing distributed but suboptimal heuristics (for sigmoidal utilities) in [9], determining optimality conditions for the canonical distributed algorithm to converge globally (for all nonlinear utilities) in this section, and proposing efficient but centralized method to compute the global optimum (for a wide class of utilities that can be transformed into polynomial utilities) in [7], are complementary in the study of distributed rate allocation by nonconcave NUM, a difficult class of nonlinear optimization.

### A. *When will canonical distributed algorithm work?*

Is it true that the canonical distributed algorithm only converges to a globally optimal rate allocation for concave utility maximization? The following counter-example answers the question in the negative.

**Example 1.** There are three flows over three links as shown in Figure 2. The small number of flows and links allow exhaustive search to compute the global optimum and check against distributed algorithm’s solution. All three utility functions are nonconcave: $U_s(x_s) = \left(1 - 2Q\sqrt{2x_s}\right)^{\beta_s}$ where $Q$ is the complementary cumulative distribution of standard Gaussian variable and $\beta_s$ are positive parameters. The link capacity vector is varied within the set $\mathcal{C} = \{\mathbf{c}(\theta) = \theta[5,\ 10,\ 6] + (1-\theta)[8,\ 6,\ 7]\}, \theta \in [0, 1]$. Each capacity vector gives one realization of a nonconcave NUM. The canonical distributed algorithm is executed for this problem and the resulting rate allocation is indeed found to be globally optimal for all $\mathbf{c} \in \mathcal{C}$. As shown in Figure 3 for some of the choices of $\mathbf{c}$, the maximized network utility through the canonical distributed algorithm matches precisely with that from exhaustive search.

A natural question thus arises: under what conditions will the canonical distributed algorithm converge to the globally optimal rate allocation for nonconcave NUM? In short, when will the canonical distributed algorithm ‘work’ for inelastic flow rate allocation? A sufficient condition is provided in the following theorem, proved in Appendix A.

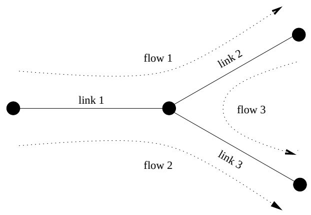

Fig. 2. Network topology for Example 1.

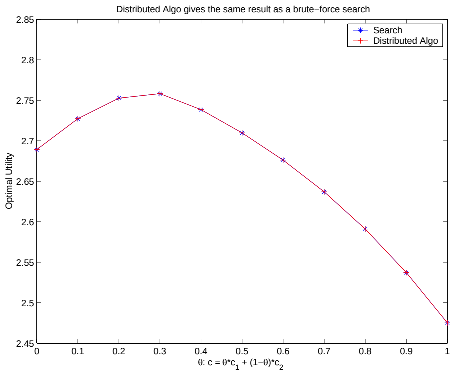

Fig. 3. Canonical distributed algorithm still converges to globally optimal rate allocation despite nonconcavity of $U_s(x_s)$.

Let $U^*$ be the (globally) optimal primal objective value and $\mathbf{x}^*$ a maximizer, *i.e.*, $U^* = \sum_s U_s(x_s^*)$, and $D^*$ be the (globally) optimal dual objective value and $\boldsymbol{\lambda}^*$ a minimizer, *i.e.*, $D^* = g(\boldsymbol{\lambda}^*)$. The duality gap for (1) is $\eta = D^* - U^* \geq 0$.

**Theorem 1.** The canonical distributed algorithm (2,4) converges to the globally optimal rate allocation if the following two conditions are satisfied: i) duality gap $\eta$ is zero, and ii) the price-based rate allocation $\mathbf{x}^*(\boldsymbol{\lambda}) = \text{argmax} L(\mathbf{x}, \boldsymbol{\lambda})$, as a function of $\boldsymbol{\lambda}$, is continuous at $\boldsymbol{\lambda}^*$.

In the next two subsections, we will further explore sufficient conditions for these two conditions in Theorem 1: zero duality gap and continuity of price-based rate allocation.

### B. *Zero duality gap*

While concavity of all utility functions $\{U_s(x_s)\}$ is a sufficient condition to guarantee zero duality gap in (1), it is *not* a necessary condition. Duality gap can be zero even for nonconvex optimization problems. However, proving zero duality gap in these cases can be much more difficult and requires arguments beyond the standard ones in convex optimization [3]. In this subsection, we provide a more general sufficient condition for duality gap to be zero, which includes the concavity condition of utility functions as a special case, but also shows that duality gap can be zero even for nonconcave utilities.

Consider the optimized total utility $U^* = \sum_s U_s(x_s^*)$ as a function of link capacities: $U^*(\mathbf{c})$. This is a completely different function from the utility functions $\{U_s(x_s)\}$, which are functions of source rates. As long as $U^*(\mathbf{c})$ is a concave function, the duality gap of (1) is zero under mild technical conditions. There is a subtle but significant difference between concavity of $U_s(x_s)$, *i.e.*, concavity of each user's utility as a function of rate, and the more general concavity of $U^*(\mathbf{c})$, *i.e.*, concavity of the optimized total utility as a function of link capacities. The former implies the latter, but not vice versa. [4]

This condition is stated formally as follows. For each $\mathbf{c}$ in a set $\mathcal{C}$ of possible link capacity vectors, there is a set of feasible $\mathbf{x}$ for (1), which defines a set of achievable values of total utility $U(\mathbf{c})$. Assume the following technical condition holds: for every sequence of $\{\mathbf{c}^k\}$ with $\mathbf{c}^k \to \mathbf{c}$, there exists a sequence of some feasible $U^k$ satisfying $\limsup_{k\to\infty} U^k \leq U^*$. The following fact from nonlinear optimization theory can be proved as shown in Appendix B.

**Fact 1.** If $U^*(\mathbf{c})$ is a concave function, duality gap $\eta = D^* - U^*$ for (1) is zero.

### C. Continuity of price-based rate allocation

It is in general difficult to test for continuity of price-based rate allocation for nonconcave utility maximization. In particular, zero duality gap does not imply this continuity property, as shown through a counter-example in Appendix C.

On the other hand, for many types of nonconcave utility functions, it is easy to characterize the set of $\boldsymbol{\lambda}$ at which $\mathbf{x}^*(\boldsymbol{\lambda})$ is *discontinuous*. For example, for a sigmoidal utility function, the slope of the tangent of a straight line from the origin with the utility curve is the only $\lambda^s$ at which $x^*_s(\lambda^s)$ is discontinuous [9]. Therefore, if we can bound the range of $\boldsymbol{\lambda}^*$ and verify that the ranges exclude these points of discontinuity, we will have guaranteed continuity of price-based rate allocation.

For example, one way to generate such bounds on $\boldsymbol{\lambda}^*$ is to use the following inequality for minimization of a strongly convex function $f(\mathbf{y})$ [3]:

$$\|\mathbf{y} - \mathbf{y}^*\|_2 \leq \frac{2}{m}\|\nabla f(\mathbf{y})\|_2, \ \ \forall \mathbf{y}$$

where $m$ is the strong convexity constant.

Consequently, we provide the following method, through centralized computation, to bound $\boldsymbol{\lambda}^*$:

- Evaluate the Hessian matrix $\nabla^2 g(\boldsymbol{\lambda})$ of the dual function $g(\boldsymbol{\lambda})$. Compute $m > 0$ such that $\nabla^2 g(\boldsymbol{\lambda}) \succeq m\mathbf{I}$ where $\mathbf{I}$ is an identity matrix.
- For a given $\boldsymbol{\lambda} \succeq 0$, compute $\|\nabla g(\boldsymbol{\lambda})\|_2 = \sum_l (c_l - \sum_{s:l\in L(s)} x_s)^2$. Compute $K(\boldsymbol{\lambda}) = \frac{2}{m}\|\nabla g(\boldsymbol{\lambda})\|_2$.
- Each $\lambda^*_l$ is (lower and upper) bounded by $|\lambda^*_l - \lambda_l| \leq \sqrt{K(\boldsymbol{\lambda})}$ for the given $\boldsymbol{\lambda}$. This pair of bounds in turn lead to upper and lower bounds on $\lambda^s$ for sources with nonconcave utilities. [5]

By picking different $\boldsymbol{\lambda}$, the above bounds can be tightened. If the upper bound is smaller than, or the lower bound larger than, the $\lambda^s$ at which $x^*_s(\lambda^s)$ is discontinuous, then the condition for continuity of price-based rate allocation holds.

### D. Another sufficient condition and a necessary condition

In this subsection, we provide a stronger sufficient condition for the canonical distributed algorithm to converge in rate allocation. It turns out that one of the two conditions in Theorem 1, the zero duality gap condition, is not needed if the other condition, the continuity condition, is satisfied.

**Theorem 2.** Continuity of price-based rate allocation $\mathbf{x}^*(\boldsymbol{\lambda})$ at the optimal prices $\boldsymbol{\lambda}^*$ implies that the canonical distributed algorithm converges to a globally optimal rate allocation.

This Theorem, as proved in Appendix D, states that continuity property *alone* is a sufficient condition for the canonical distributed algorithm to 'work'. Theorems 1 and 2 together show that continuity implies zero duality gap.

**Example 2.** An illustrative example is summarized below for a simple topology: two flows sharing a link with capacity $c$. One flow is elastic data traffic, with logarithmic concave utility function, and the other is inelastic traffic, with sigmoidal utility function shown in Figure 4. The critical dual variable $\lambda^0$ is the slope of the tangent of the straight line from the origin with the sigmoidal curve. In this example, following similar development in Appendix C, we can show that the canonical distributed algorithm converges to a globally optimal rate allocation for a large enough link capacity, *i.e.*, when $c \geq c_{min}$.

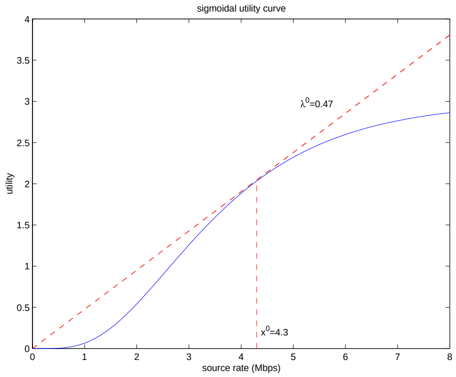

Fig. 4. Sigmoidal utility function for one of the two sources in Example 2.

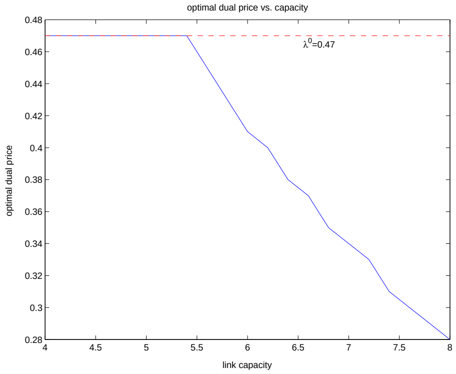

Fig. 5. The dual optimal $\lambda^*$ as a function of the link capacity $c$.

As shown in Figures 5 and 6, respectively, as link capacity becomes larger than the threshold $c_{min} = 5.38$, the optimal

[4] In Example 1, through exhaustive search it can be verified that $U^*(\mathbf{c})$ is concave for $\mathbf{c} \in \mathcal{C}$.

[5] Bounding $\lambda^s$ for sources with concave utilities is unnecessary, since their rates are always continuous in $\boldsymbol{\lambda}$.

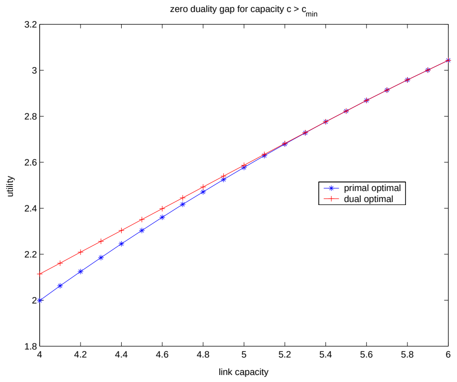

Fig. 6. Duality gap $\eta$ as a function of the link capacity $c$.

dual variable is smaller than $\lambda^0$, thus ensuring continuity of price-based rate allocation, and duality gap becomes zero. It can be verified that the canonical distributed algorithm indeed converges to the globally optimal rate allocation for $c \geq c_{min}$.

If $\mathbf{x}^*(\boldsymbol{\lambda})$ is discontinuous at all optimal prices $\boldsymbol{\lambda}^*$, the canonical distributed algorithm certainly cannot converge. If the dual objective function $g(\boldsymbol{\lambda})$ is strictly convex (and it is certainly always convex), even if it is non-differentiable, there is a unique optimal price. Therefore, the next Theorem and Corollary follow:

**Theorem 3.** Continuity of $\mathbf{x}^*(\boldsymbol{\lambda})$ at at least one of the optimal prices $\boldsymbol{\lambda}^*$ is a necessary condition for the canonical distributed algorithm to converge to a globally optimal rate allocation.

**Corollary 1.** If $g(\boldsymbol{\lambda})$ is strictly convex, continuity of $\mathbf{x}^*(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}^*$ is a necessary and sufficient condition for the canonical distributed algorithm to converge to the globally optimal rate allocation.

In concluding this section, we mention that there is an alternative link price update method by sequential bisection. In Appendix E, we show that the bisection method is never better than the canonical distributed algorithm.

## III. Time Sensitive Flows: Optimization Framework

We turn to a different modeling approach for inelastic flows in this and the next section. Instead of using nonconcave utility functions, we explicitly model different types of inelastic flows, and incorporate the models into the constraints and objective functions of NUM.

### A. Traffic models

We consider three types of time sensitive flows in this section. The assumptions on discrete time slots and continuous flow intensity remain. Time slots are uniformly spaced, each assumed, without loss of generality, to have a duration of 1 time unit.

First is the type of flows from real-time IP applications or CBR ATM traffic. This type of flows requires constant playback rate, and playback is started at the requested starting time. We denote by $R$ the set of Real-time flows, and use $r$ to index these sources.

- Required starting time is $t_r^s$. Total file size is $l_r$. Playback rate is required to be a constant of $v_r$ bits per time unit. These are constant parameters.
- For each source, an admission decision $a_r$ is made. Utility obtained is a constant $\bar{U}_r$ if the flow is admitted and 0 otherwise.
- The optimization variables are $a_r \in \{0, 1\}$. It is an admission control problem for R-type flows.

Second is the type of flows from streaming applications. Like real-time flows, they also require constant playback rates. But a playback buffer at the receiving end host can absorb fluctuations of the source rate to some extent. We denote by $B$ the set of Buffered streaming flows, and use $b$ to index these sources.

- Requested starting time is $t_b^s$, and the file size is $l_b$. Each flow $b$ has a playback buffer of size $f_b$ at the receiving end host, with a constant playback rate of $v_b$ between the actual playback start time and the end of the playback. These are constant parameters.
- The optimization variables are $w_b \geq t_b^s$, the actual time for playback to start, and $x_b(t)$, the rate allocation over time. Note that source rate $x_b(t)$ must satisfy $x_b(t) = 0$ for $t < t_b^s$. It is a joint problem of scheduling and rate allocation over time slots for B-type flows.
- To guarantee that during the playback period, the buffer is neither depleted nor overflown, we have the constraint $0 \leq \sum_{t=t_b^s}^{t_0} x_b(t) - (t_0 - w_b)v_b \leq f_b$ for all $t_0 \in [w_b, w_b + l_b/v_b]$. In particular, if $t_0 = w_b$, this condition guarantees that there is no playback buffer overflow during the preloading phase between the requested and actual playback start times.
- The utility is $U_b(w_b)$ where $U_b$ is a nonincreasing (and concave) function, since users would like to start the playback as early as possible.

Third is the type of general delay-sensitive traffic, where source utility depends on the transient behavior of rate allocation. We denote by $D$ the set of Delay-sensitive flows, and use $d$ to index these sources.

- Requested starting time is $t_d^s$. File size is $l_d$. These are constant parameters.
- The optimization variables are $x_d(t)$, constrained by $x_d(t) = 0$ if $t < t_d^s$, and $\sum_{t=t_d^s}^{\infty} x_d(t) = l_d$. It is a rate allocation problem over all time slots for D-type flows.
- The utility associated with data rate time series $\{x_d(t)\}$ is $U_d(\{x_d(t)\})$, *i.e.*, the utility function maps from the *vector* of source rates over *all* time slots to a real number. We will make a restrictive assumption in this paper that the utility is additive and memoryless: $U_{d,t}(\{x_d(t)\}) = \sum_t U_d(x_d(t))$, where $\{U_{d,t}\}$ are concave and increasing.

### B. NUM formulation

For notational simplicity, we focus on the single link case in this subsection. The formulations and decomposition method can be readily generalized to arbitrary network topologies. Suppose all three types of traffic share a single link with a fixed link capacity $c$. The problem of rate allocation (together with admission control and scheduling) among the flows can thus be expressed as the following NUM problem over the variables of $x_d(t), a_r, x_b(t), w_b, \ \forall t, d, r, b$. The first constraint is due to the limited link capacity, and the other constraints are due to the traffic models described in the last subsection. Notice that this is a constrained nonlinear optimization problem with two integer constraints.

$$\begin{array}{ll} \text{maximize} & \sum_d \sum_{t=0}^{\infty} U_{d,t}(x_d(t)) + \sum_r a_r \bar{U}_r + \sum_b U_b(w_b) \\ \text{subject to} & \sum_d x_d(t) + \sum_{r:t_r^s \le t \le t_r^s + l_r/v_r} a_r v_r + \sum_b x_b(t) \\ & \quad \le c, \ \ \forall t, \\ & x_d(t) = 0, \ \ \forall t < t_d^s, \ \ \forall d \in D, \\ & \sum_{t=t_d^s}^{\infty} x_d(t) = l_d, \ \ \forall d \in D, \\ & a_r \in \{0, 1\}, \ \ \forall r \in R, \\ & x_b(t) = 0, \forall t < t_b^s, \ \ \forall b \in B, \\ & \sum_{t=w_b}^{\infty} x_b(t) = l_b, \ \ \forall b \in B, \\ & 0 \le \sum_{t=t_b^s}^{t_0} x_b(t) - (t_0 - w_b) v_b \le f_b, \\ & \quad \forall t_0 \in [w_b, w_b + l_b/v_b], \ \ \forall b \in B, \\ & w_b \ge t_b^s, w_b \in \mathcal{Z}_+ \ \ \forall b \in B, \\ & x_d(t) \ge 0, x_b(t) \ge 0, \ \ \forall d \in D, b \in B, \ \ \forall t. \end{array} \tag{5}$$

We now make the following simplification of the models and relax one of the integer constraints:

- Assume that each receiver playback buffer is infinitely large. It may be depleted but not overflown.
- Assume that all three types of flows have infinite backlog.
- Assume that all the requested flow starting time are time 0, and the rate allocation problem exists only for $t \in [0, T]$.
- Relax the integer constraint on $w_b$. For short time slots, this relaxation introduces only small errors.

We then obtain the following simplified version of NUM for rate allocation among three types of inelastic flows on a single link. The first constraint avoids exceeding the limited link capacity and the second constraint avoids depletion of receiver playback buffers. The optimization variables are $x_d(t), a_r, x_b(t), w_b$.

$$\begin{array}{ll} \text{maximize} & \sum_d \sum_{t=0}^{T} U_{d,t}(x_d(t)) + \sum_r a_r \bar{U}_r + \sum_b U_b(w_b) \\ \text{subject to} & \sum_d x_d(t) + \sum_r a_r v_r + \sum_b x_b(t) \le c, \ \ \forall t, \\ & \sum_{t=0}^{t_0} x_b(t) \ge (t_0 - w_b) v_b, \ \ \forall t_0 > w_b, \ \ \forall b, \\ & x_d(t) \ge 0, x_b(t) \ge 0, \ \ \forall d \in D, b \in B \ \ \forall t, \\ & a_r \in \{0, 1\}, \ \ \forall r \in R, \\ & w_b \in [0, T] \ \ \forall b \in B. \end{array} \tag{6}$$

### C. Decomposition

Similar to the dual decomposition that leads to the canonical distributed algorithm for elastic traffic, we would like to decompose the problem (6) into individual source problems and link problems. Such a decomposition is indeed possible. Furthermore, it turns out that $a_r$ should be either $0$ or $1$, and the integer constraints on $a_r$ do not introduce technical difficulties. The problem of having the number of constraints depend on the optimization variable $w_b$ in the second type of constraints in (6) also turns out not to introduce optimization-theoretic difficulties.

**Theorem 4.** Utility maximization for time sensitive flows (6) can be decomposed into the following individual source problems and a network problem:

The $D$-type source problem, one for each source $d$ and each time $t$:

$$\begin{array}{ll} \text{maximize} & U_{d,t}(x_d(t)) - \lambda(t) x_d(t) \\ \text{subject to} & x_d(t) \ge 0 \end{array} \tag{7}$$

Solution of each of the $D$-type source problem is $x_d^*(t) = U_{d,t}^{'-1}(\lambda(t))$, and the maximized utility is $U_{d,t}(U_{d,t}^{'-1}(\lambda(t))) - \lambda(t) U_{d,t}^{'-1}(\lambda(t))$.

The $R$-type source problem, one for each source $r$:

$$\begin{array}{ll} \text{maximize} & (\bar{U}_r - \lambda_T v_r) a_r \\ \text{subject to} & a_r \in \{0, 1\} \end{array} \tag{8}$$

where $\lambda_T = \sum_{t=0}^{T} \lambda_t$. Solution of each of the $R$-type source problem is $a_r = 1$ if $\bar{U}_r \ge \lambda_T v_r$ and $a_r = 0$ otherwise, and the maximized utility is $(\bar{U}_r - \lambda_T v_r) \mathbf{1}\{\bar{U}_r \ge \lambda_T v_r\}$.

The $B$-type source problem, one for each source $b$:

$$\begin{array}{ll} \text{maximize} & U_b(w_b) - \sum_{t=0}^{T} \lambda(t) x_b(t) \\ \text{subject to} & x_b(t) \ge 0, \ \ \forall t, \\ & w_b \in [0, T], \\ & \sum_{t=0}^{t_0} x_b(t) \ge (t_0 - w_b) v_b, \forall t_0 \end{array} \tag{9}$$

There is in general no analytic solution to this linearly constrained concave maximization, but numerical solutions can be computed locally and efficiently to obtain $U_b^* = U_b(w_b^*) - \sum_{t=0}^{T} \lambda(t) x_b^*(t)$ where $(w_b^*, x_b^*(t))$ is a solution to (9).

The network problem is to maximize the optimized values of the three types of source problems, *i.e.*, for all $t$, maximize over $\lambda_t \ge 0$ the following:

$$U_{d,t}(U_{d,t}^{'-1}(\lambda(t))) - \lambda(t) U_{d,t}^{'-1}(\lambda(t))$$
$$+ (\bar{U}_r - \lambda_T v_r) \mathbf{1}\{\bar{U}_r \ge \lambda_T v_r\} + U_b^* + \lambda_T c.$$

Solution to the master problem can be obtained by a distributed subgradient method iteratively within each time slot $t$, where the iteration number is denoted by $k$:

$$\lambda^t(k+1) = \left[\lambda^t(k) - \alpha(k) h^t(k)\right]^+,$$

where a subgradient is

$$h^t(k) = c - \sum_{d \in D} x_d^{*t}(k) - \sum_{r \in R} a_r^{*t} v_r - \sum_{b \in B} x_b^{*t}(k),$$

and step sizes $\alpha(k)$ can be chosen to ensure convergence of this algorithm based on dual decomposition.

While the above problem decomposition can be proved, there are two significant new algorithmic challenges. First, we now need to generate congestion prices for every time slot along the temporal dimension, instead of for every link along the spatial dimension (and in the case of time sensitive traffic on general networks, prices need to be generated along both dimensions). This creates a time scale problem. In order to obtain the correct price per time slot $t$, iterations indexed by $k$ need to be carried out before reaching close to the equilibrium. Furthermore, $\lambda_t^*$ depends on $\lambda_{t'}^*, t' \neq t$. Thus the optimal admission decision, playback time decision, and rate allocation cannot be made until the entire period $t = 0, \ldots, T$ is over. These two issues of *time-scale* and *causality* are the two bottlenecks in distributive and iterative rate allocation for time-sensitive flows.

## IV. Time Sensitive Flows: Admission Control Heuristics and Optimal Algorithm

In this section, we investigate two special cases by considering only one type of time sensitive flows, real-time flows indexed by $r$, sharing bandwidth with elastic TCP data flows indexed by $i$, in a network with multiple links indexed by $l$. Problem (6) reduces to

$$\begin{array}{ll} \text{maximize} & \sum_i U_i(x_i) + \sum_r a_r \bar{U}_r \\ \text{subjec to} & \sum_{i:l \in L(i)} x_i + \sum_{r:l\in L(r)} a_r v_r \leq c_l, \quad \forall l, \\ & x_i \geq 0, \quad \forall i, \\ & a_r \in \{0,1\}, \quad \forall r, \end{array} \qquad (10)$$

where the optimization variables are $x_i$ and $a_r$.

A development similar to Theorem 4 shows that to globally solve (10), the canonical distributed algorithm can be used for price update and elastic source rate adjustment, but the real-time flows should be admitted based on the equilibrium price $\boldsymbol{\lambda}^*$. A real-time flow $r$ is admitted if and only if $\lambda^{r*} \leq \frac{\bar{U}_r}{v_r}$. Therefore, to arrive at the correct admission decision, one must wait for the equilibrium to be achieved. When real-time flows may be rejected but waiting for the optimal admission decision is undesirable, a price-based admission control heuristics in subsection V.A can be used. Alternatively, if one assumes that all real-time flows are admitted, an optimal algorithm in subsection V.B can be used to ensure a fair share of bandwidth to the elastic flows.

### A. Pricing-based admission control

**Algorithm 1: Admission control heuristics.** This heuristic is conducted locally at the edge by each source, following the end-to-end principle and assuming cooperative end users. It is parameterized by nonnegative integers $m$ and $n$. If a price seen by source $r$ is smaller than $\frac{\bar{U}_r}{v_r}$ for $m$ time slots, it is *tentatively* admitted, and a message is passed to reserve $v_r$ amount of bandwidth along the path it uses. This tentative admission phase is the resource reservation phase. If the price $\lambda^r$ continues to be smaller than $\frac{\bar{U}_r}{v_r}$ for $n$ more time slots, the flow is formally admitted and transmission can start, otherwise the flow is rejected and has to wait for another window of $m$ times slots, during which the price is sufficiently low, before entering the resource reservation phase again. As is typical with other price-based heuristics (*e.g.*, [9]), larger waiting parameters $(m, n)$ enhance the probability that the correct admission decision is made but also increase the latency incurred.

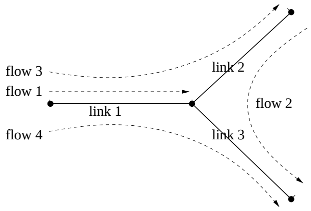

Fig. 7. Network topology for Example 3.

**Example 3.** The admission control heuristic is tested on various networks. Typical results are summarized below for the network shown in Figure 7 with three links and four flows. Link capacities are 30, 20, and 40 units respectively. Flows 1 and 2 are elastic data flows, with utility functions $U_i = \log(1 + x_i), i = 1, 2$, while flows 3 and 4 are inelastic real-time flows with $\bar{U}_3 = 1, \bar{U}_4 = 0.2$, and playback rates of $v_3 = 5, v_4 = 7$. If both flows 3 and 4 use $(m, n) = (8, 8)$ in the admission control heuristic, Figure 8 shows the resulting rate allocation iterations and convergence to the optimal solution. In this example, flow 3 is admitted in the first try, and flow 4 only enters the resource reservation phase once, during which it is rejected. The optimal solution for this utility maximization problem is indeed to admit flow 3 and reject flow 4.

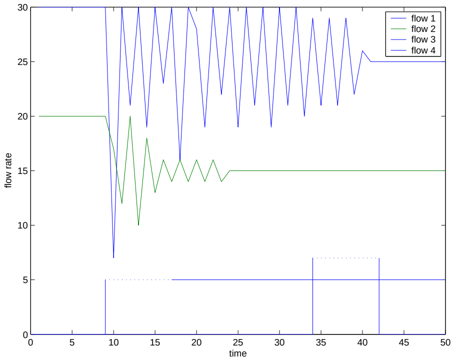

Fig. 8. Rate allocation based on the admission control heuristic.

Extensive simulations are conducted on the same topology for other $(m, n)$ pairs. A variety of rate allocation scenarios are observed and classified into correct or incorrect final admission decisions by comparing against the global optimum computed through exhaustive search. The region in the $(m, n)$ plane where correct decisions are made is the shaded region (and the rest of the 2D plane upwards and to the right) in Figure 10. This connected region illustrates the following desirable and intuitive properties of the heuristic:

- When either $m$ or $n$ is larger than a threshold $m_0$ or $n_0$, the other parameter can be as small as zero. If both $m$ and $n$ are nonzero, they can be smaller than $m_0$ or $n_0$ and still

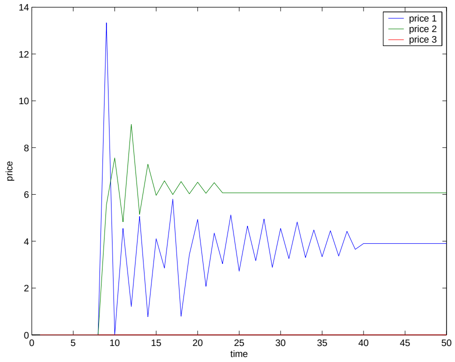

Fig. 9. Price update based on the admission control heuristic.

remain in the correct decision region.

- The Pareto optimal tradeoff curve is the boundary line between the shaded region and the unshaded region. If the total latency before formal admission needs to be minimized, it is best to operate at the point $(m = 7, n = 0)$ for the network in Figure 7.
- In practice, it is unlikely that the best $(m, n)$ will be used. Thus it is useful to observe that the latency associated with any point on the Pareto optimal tradeoff curve in the $(m, n)$ plane is only a fraction of about $20\%$ of the time it takes for all the flows to converge. This shows the effectiveness of this heuristic in reducing the time it takes to make the right admission decision.

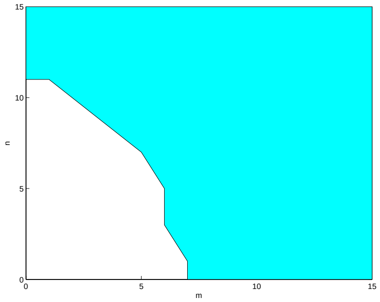

Fig. 10. The correct decision region in the $(m, n)$ plane is shaded.

There are several variations on the admission control heuristics, some requiring explicit feedback from the links and coordination among the sources. A simple one that does not require any extra communication overhead is to adapt the choices of $(m, n)$ parameters. Each inelastic flow picks large $(m, n)$ initially, since the gradient based iterations have just started. If the flow is rejected after being tentatively accepted, thus having to wait for the next round, smaller $(m, n)$ are picked because the iterations have reduced the distance between the rate allocation at that time and the optimal rate allocation.

### B. *Rate allocation among different groups of flows*

Sometimes, when link capacities are expected to be large enough to accommodate all the real-time flows, we may admit all real-time flows and ensure that their rates are never smaller than the necessary threshold $v_r$. Modeling the situation where rates higher than $v_r$ may lead to improvements in user perception of the real-time applications, we also allow the utility functions to grow beyond $\bar{U}_r$, following a concave utility $U_r(x_r)$ with $U_r(v_r) = \bar{U}_r$. We refer to these flows as enhanced real-time flows. However, to ensure that the group of elastic traffic can still be allocated at least a certain share of the bandwidth, we put an upper bound on the fraction $\theta_l$ of total link capacity that can be given to the group of enhanced real-time flows. This results in the following NUM problem:

$$\begin{array}{ll} \text{maximize} & \sum_i U_i(x_i) + \sum_r U_r(x_r) \\ \text{subjec to} & \sum_{i:l\in L(i)} x_i + \sum_{r:l\in L(r)} x_r \le c_l, \quad \forall l, \\ & \sum_{r:l\in L(r)} x_r \le \theta_l c_l, \quad \forall l, \\ & x_r \ge v_r, \quad \forall r, \\ & \mathbf{x} \succeq 0. \end{array} \qquad (11)$$

The above problem can be solved by **Algorithm 2: Rate allocation among groups.** This algorithm is the canonical distributed algorithm together with two modifications. First, at each iteration, the allocated rate to a source in the enhanced real-time group is projected onto the interval $[v_r, \min_{l\in L(r)} c_l]$, *i.e.*, if $x_r(k) < v_r$ at some iteration $k$, set $x_r(k) = v_r$. Two, we update another link price vector $\boldsymbol{\sigma}$, distributively on each link:

$$\sigma_l(k+1) = \left[ \sigma_l(k) - \alpha(k) \left( \theta_l c_l - \sum_{r:l\in L(r)} x_r^*(k) \right) \right]^+ .$$

Each source in the enhanced real-time group uses the total price $\lambda^r$ *plus* $\sigma^r = \sum_{l\in L(r)} \sigma_l$. Each source in the elastic group uses *only* price $\lambda^r$.

**Theorem 5.** Algorithm 2 converges to the globally optimal rate allocation of (11).

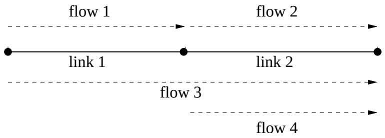

Fig. 11. Network topology for Example 4.

**Example 4.** A typical simulation for Algorithm 2 is shown for the network in Figure 11, where link capacities are 30 and 20 units, respectively. Flows 3 and 4 are the enhanced real-time flows with utility functions $U_3 = 0.871x_3^{0.1}, U_4 = 4.1808x_4^{0.9}$ and playback rate $v_3 = 5, v_4 = 4$ units. Flows 1 and 2 are elastic data flows with utility functions $U_i = \log(1 + x_1), i = 1, 2$. Figures 12 and 13 show the source rate and total path price iterations. It is observed that both enhanced real-time flows never have their source rates dropped below the minimum thresholds, and, at the same time, they do not occupy more than $\theta_1 = 30\%, \theta_2 = 50\%$ of link capacities. The equilibrium rate allocation for flow 4 at global optimality is 6 units, more than the minimum threshold $v_4$ required by this enhanced real time flow.

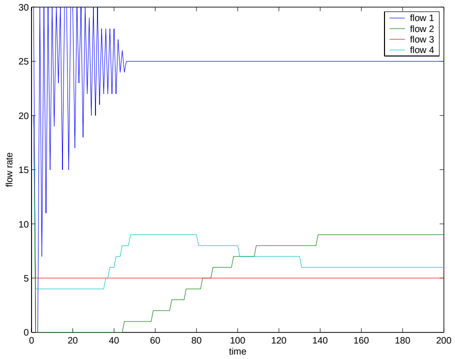

Fig. 12. Source rate allocation among four sources in two groups.

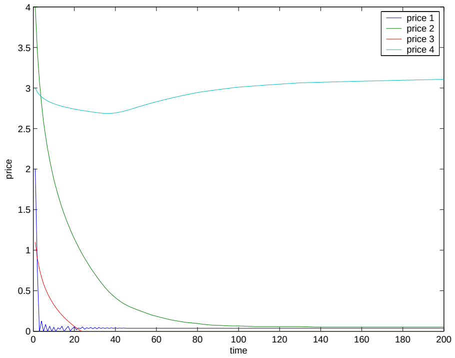

Fig. 13. Total path price iteration for four sources in two groups.

## V. Concluding Remarks

The difficult problem of distributed rate allocation for inelastic flows is tackled in this paper from two angles:

- When the concavity assumption for utility functions breaks down, we prove several sufficient conditions (*e.g.*, continuity of price-based rate allocation) and a necessary condition for the canonical distributed algorithm to converge to a globally optimal rate allocation.
- When time sensitivity is modeled in the utility maximization problem, a dual decomposition is presented. We propose a simple yet effective admission control heuristic to allocate rates between elastic TCP flows and inelastic realtime flows. When link capacities are large enough that admission control becomes unnecessary, we provide an optimal algorithm to allocate rates among traffic groups with different elasticities.

This paper tackles NUM beyond the widespread yet often invalid assumption of elastic flow. Many open issues remain to be resolved in this area, including noncausality of optimal admission control, transient behavior of the canonical distributed algorithm, and fairness and smoothness of rate allocation to inelastic flows.

The results in section II apply to other applications of linearly constrained nonconvex optimization with a separable objective function, such as spectrum management in OFDM communication systems. It will also be interesting to investigate the case when the linear flow constraints become nonlinear, for applications in wireless network power control.

## Appendix A: Proof of Theorem 1

*Proof:* Price update through subgradient descent with diminishing step sizes makes the dual variables converge to a global minimizer of the dual function $g(\boldsymbol{\lambda})$. Since $g(\boldsymbol{\lambda})$ is a convex function, it is also continuous, thus converging to $D^*$ as $\boldsymbol{\lambda}$ converges to $\boldsymbol{\lambda}^*$. Since the price-based rate allocation $\mathbf{x}^*(\boldsymbol{\lambda})$ is continuous at $\boldsymbol{\lambda}^*$, $\mathbf{x}^*(\boldsymbol{\lambda})$ converges to $\mathbf{x}^*(\boldsymbol{\lambda}^*)$ as $\boldsymbol{\lambda}$ converges to $\boldsymbol{\lambda}^*$.

It is known [3] that zero duality gap implies complementary slackness and that a primal maximizer (*i.e.*, a globally optimal rate allocation in this case) must also be a Lagrangian maximizer (*i.e.*, price-based rate allocation) at $\boldsymbol{\lambda}^*$. We can also show that continuity and monotonicity of $\mathbf{x}^*(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}^*$ imply its uniqueness. Thus $\mathbf{x}^*(\boldsymbol{\lambda}^*)$ is primal feasible and optimal. After the guaranteed convergence of price-base rate allocation to a feasible solution, the resulting rate allocation is globally optimal.

■

## Appendix B: Proof of Fact 1

*Proof:* We will use the 'min common max crossing duality' established in [2]. A similar technique has recently been used for spectrum management in DSL [14].

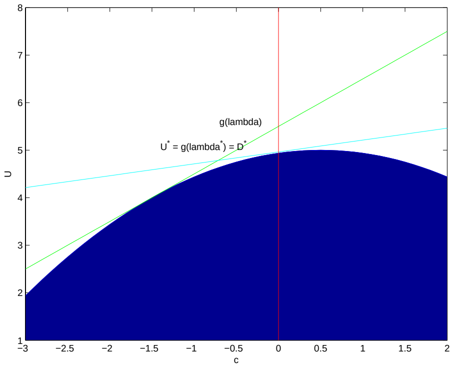

Fig. 14. Concavity of $U^*(\mathbf{c})$ implies zero duality gap.

First, we rewrite NUM (1) as

$$\begin{array}{ll} \text{maximize} & \sum_s U_s(x_s) \\ \text{subject to} & c_l - \sum_{s:l\in L(s)} x_s \geq 0, \quad \forall l. \end{array} \qquad (12)$$

We also consider a perturbed utility optimization problem, where the link capacities become $c_l + c_l'$, $\forall l$. Notice that concavity of $U^*(\mathbf{c})$ is equivalent to concavity of $U^*(\mathbf{c}')$.

Define $M$ (shaded region in Figure 14) to be the set of two-tuples of link capacity perturbation and resulting feasible total utility:

$$M = \left\{ (\mathbf{c}', U) | \exists \mathbf{x} \text{s.t.} c_l - \sum_{s:l\in L(s)} x_s \geq -c'_l, U \leq \sum_s U_s(x_s) \right\}.$$

Utility maximization (12) can now be equivalently written as optimizing over a scalar variable $U$:

$$\begin{array}{ll} \text{maximize} & U \\ \text{subject to} & (\mathbf{0}, U) \in M \end{array}$$

The optimal network utility $U^* = \sup_{(\mathbf{0},U)\in M} U$ is thus the ‘min common’ point defined in [2].

Consider the maximized total utility $U^*(\mathbf{c}')$ as a function of link capacity perturbation $\mathbf{c}'$, the tangent of $U^*(\mathbf{c}')$ at a point $\mathbf{c}'_0$ along the direction $\boldsymbol{\lambda} \succeq 0$, and the intersection of this tangent with the $y$-axis, as shown in Figure 14. The $y$-coordinate at the intersection is

$$\begin{aligned} y(\boldsymbol{\lambda}) &= U^*(\mathbf{c}'_0) + \boldsymbol{\lambda}^T(\mathbf{0} - \mathbf{c}'_0) \\ &= U^*(\mathbf{c}'_0) - \boldsymbol{\lambda}^T\mathbf{c}'_0 \\ &= \max_{\mathbf{c}'} \left[ U - \boldsymbol{\lambda}^T\mathbf{c}' \right] \end{aligned}$$

where the last equality can be readily verified either geometrically or by differentiation. Since $\boldsymbol{\lambda} \succeq 0$ and $\mathbf{c}'$ is lower bounded by $\sum_{s:l\in L(s)} x_s - c_l$, we have

$$\begin{aligned} y(\boldsymbol{\lambda}) &= \max_{\mathbf{x}} \left[ U(\mathbf{x}) + \sum_l \lambda_l \left( c_l - \sum_{s:l\in L(s)} x_s \right) \right] \\ &= \max_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) = g(\boldsymbol{\lambda}). \end{aligned}$$

Therefore, the $y$-coordinate of the intersection is the Lagrange dual function evaluated at $\boldsymbol{\lambda}$. We would like to find the minimum $y(\boldsymbol{\lambda})$ by varying over $\boldsymbol{\lambda} \succeq 0$, *i.e.*, the lowest intersection as the angle of the tangent varies. That would equal the dual optimal value $D^*$ by definition. We have thus established that the ‘min common’ point equals $U^*$ and the ‘max crossing’ point equals $D^*$.

A function $f(\mathbf{x})$ is concave if and only if its hypograph, the set $\{(\mathbf{x}, t) | t \leq f(\mathbf{x})\}$, is a convex set. Set $M$ is the hypograph of $U^*(\mathbf{c}')$, since $M$ contains all the feasible two-tuples of $\mathbf{c}'$ and $U$. If $U^*(\mathbf{c}')$ is concave, $M$ is a convex set, and by the result in [2], min common point equals max crossing point. Therefore $U^* = D^*$. ■

## Appendix C: Zero duality gap and continuity

This appendix shows that zero duality gap does not imply continuity. Consider the following problem with two variables $x_1, x_2$ and one constraint:

$$\begin{array}{ll} \text{maximize} & U_1(x_1) + U_2(x_2) \\ \text{subject to} & x_1 + x_2 \leq c \end{array}$$

where $U_1$ is a concave function and $U_2$ is a sigmoidal function. We will show that for a given $U_2$ function, we can construct a $U_1$ function such that there is zero duality gap and yet $x_2(\lambda)$ is discontinuous at the optimal dual variable $\lambda^*$.

For any given curve $U_2(x_2)$, construct a straight line from the origin to be tangent to the curve, and denote the x-coordinate and slope of the tangent point by $x_2^0$ and $\lambda^0$, respectively. It is easy to verify that $x_2^0 = x_2^*(\lambda^0)$, *i.e.*, $x_2^0$ maximizes $U_2(x_2) - \lambda^0 x_2$. Furthermore, $x_2^*(\lambda)$ is discontinuous at $\lambda^0$. We would like to construct $U_1$ so that $\lambda^0$ is the dual optimal $\lambda^*$. Since the dual problem is an unconstrained convex minimization problem, any stationary point is globally optimal. Therefore, $\lambda^0$ is the dual optimal $\lambda^*$ if the subgradient vanishes at $\lambda^0$, *i.e.*, $x_1^*(\lambda^0) + x_2^*(\lambda^0) = c$. We need $U_1$ be such that $x_1(\lambda^0) = c - x_2^*(\lambda^0) = c - x_2^0$, which, by concavity of $U_1$, is equivalent to $U_1^{'-1}(\lambda^0) = c - x_2^0$. It is always possible to construct $U_1$ such that $U_1^{'-1}$ passes through a given point $(\lambda^0, c - x_2^0)$, which makes $x_2^*(\lambda)$ discontinuous at $\lambda^0$.

With the above construction of $U_1$ to make $\lambda^0$ the dual optimal variable, we can also show that duality gap is zero. On the one hand, by the definition of dual optimality, $D^* = g(\lambda^*)$, which is in turn equal to $g(\lambda^0) = L(x^*(\lambda^0), \lambda^0) = U_1(x_1^*(\lambda^0)) + U_1(x_1^*(\lambda^0)) + 0$. Thus, by weak duality, $D^* = U_1(x_1^*(\lambda^0)) + U_1(x_1^*(\lambda^0)) \geq U^*$. On the other hand, by definition of primal optimality, $U^* \geq U_1(x_1) + U_2(x_2)$ for all $\mathbf{x}$. Therefore, $U^* = D^*$.

## Appendix D: Proof of Theorem 2

*Proof:* Continuity of price-based rate allocation $\mathbf{x}^*(\boldsymbol{\lambda})$ at the optimal price $\boldsymbol{\lambda}^*$ implies that $\mathbf{x}^*(\boldsymbol{\lambda}^*)$ and $\boldsymbol{\lambda}^*$ satisfy complementary slackness: $\lambda_l^*(\sum_{s:l\in L(s)} x_s^*(\lambda^{s*}) - c_l) = 0, \forall l$. To see this, consider the case where $\lambda_l^* > 0$, we can then find $\lambda_l^{min} = \lambda_l^* - \delta$ and $\lambda_l^{max} = \lambda_l^* + \delta$, for sufficiently small $\delta > 0$, in the neighborhood of $\lambda_l^*$. The price-based rate allocation evaluated at $\lambda_l^{min}$ must sum up to $\geq c_l$, because the subgradient at $\lambda_l^{min}$ is strictly positive due to $\lambda_l^* > \lambda_l^{min}$, and that evaluated at $\lambda_l^{max}$ must sum up to $\leq c_l$, because the subgradient at $\lambda_l^{max}$ is strictly negative due to $\lambda_l^* < \lambda_l^{max}$. By continuity of $\mathbf{x}^*(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}^*$, $\sum_{s:l\in L(s)} x_s^*(\lambda^{s*}) = c_l$.

By a similar argument, but applied only to $\lambda_l^{max}$, we can show that if $\lambda_l^* = 0$, then the price-based rate allocation evaluated at $\lambda_l^{max}$ must sum up to $\leq c_l$. By continuity, $\sum_{s:l\in L(s)} x_s^*(\lambda^{s*}) \leq c_l$ in this case. In summary, continuity of $\mathbf{x}^*(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}^*$ implies primal feasibility of $\mathbf{x}^*(\boldsymbol{\lambda}^*)$

Therefore, we know $D^* \leq U^*$ as shown below:

$$\begin{aligned} D^* &\overset{(a)}{=} g(\boldsymbol{\lambda}^*) \\ &\overset{(b)}{=} \max_{\mathbf{x}} L(\mathbf{x}(\boldsymbol{\lambda}^*), \boldsymbol{\lambda}^*) \\ &\overset{(c)}{=} \sum_s U_s(x_s^*(\lambda^{s*})) + \sum_l \lambda_l^* \left( c_l - \sum_{s:l\in L(s)} x_s^*(\lambda^{s*}) \right) \\ &\overset{(d)}{=} \sum_s U_s(x_s^*(\lambda^{s*})) \\ &\overset{(e)}{\leq} U^* \end{aligned}$$

where (a) follows from the definition of dual optimal value, (b) from the definition of Lagrange dual function, (c) from the definition of Lagrangian, (d) from complementary slackness, and (e) from the definition of primal optimal value. Of course, by weak duality, $D^* \geq U^*$. Thus $D^* = U^*$, *i.e.*, inequality (e) must be an equality. Therefore, the the canonical distributed algorithm converges to the globally optimal rate allocation. ■

## APPENDIX E: COMMENTS ON BISECTION ALGORITHM

Consider the following price update by a bisection method executed *sequentially* over the links (see also the recent result for DSL spectrum management [4]):

1) Each link maintains an upper bound $\lambda_l^{max}$ and a lower bound $\lambda_l^{min}$, sets its price $\lambda_l = \frac{\lambda_l^{max}+\lambda_l^{min}}{2}$, and passes $\{\lambda_{l'}\}_{l'=1,2,\ldots,l}$ to the next link $l+1$, for $l = 1, 2, \ldots, L-1$.
2) Based on the total path price $\lambda^s$, each source solves the local net utility maximization: $\max_{x_s} \quad [U_s(x_s) - \lambda^s x_s]$.
3) At each link, if $\sum_{s:l\in L(s)} x_s > c_l$, then $\lambda_l^{min} = \lambda_l$, otherwise $\lambda_l^{max} = \lambda_l$. Repeat 1).

**Theorem 6.** If the above bisection price update algorithm converges, the canonical distributed algorithm converges to the globally optimal rate allocation.

*Proof:* First, we show that continuity of price-based rate allocation implies convergence of the bisection algorithm. The key idea is that after link prices $\lambda_l$ for links $l$ in a set $\mathcal{L}$ converge, the resulting rate allocation is the globally optimal solution of the following partially constrained utility maximization [4]:

$$\begin{array}{ll} \text{maximize} & \sum_s U_s(x_s) \\ \text{subject to} & c_l - \sum_{s:l\in L(s)} x_s \geq 0, \quad \forall l \in \mathcal{L}. \end{array} \tag{13}$$

To see this, first notice that, as can be readily verified, $\sum_{s:l\in L(s)} x_s(\lambda_l)$ is a decreasing function of $\lambda_l$ (unless some $x_s$ is infinity, which is ruled out by Assumption 2). Now consider the two possible cases for each link $l \in \mathcal{L}$: $\sum_{s:l\in L(s)} x_s(\lambda_l^{min}) > c_l$ at $\lambda_l^{min} = 0$ and $\sum_{s:l\in L(s)} x_s(\lambda_l^{min}) \leq c_l$ at $\lambda_l^{min} = 0$. In the first case, since the initial condition must satisfy the constraint $\sum_{s:l\in L(s)} x_s(\lambda_l^{max}) \leq c_l$, throughout the iterations, we must have $\sum_{s:l\in L(s)} x_s(\lambda_l^{max}) \leq c_l$ and $\sum_{s:l\in L(s)} x_s(\lambda_l^{min}) > c_l$. As $\lambda_l^{min}$ and $\lambda_l^{max}$ both converge to a fixed value with $\lambda_l$ sandwiched in between, $\lambda_l$ also converges to a point $\tilde{\lambda}_l$. If $\sum_{s:l\in L(s)} x_s(\lambda_l)$ is continuous at $\tilde{\lambda}_l$, it converges to $c_l$. In the second case, it is easy to verify that $\lambda_l$ converges to $\tilde{\lambda}_l = 0$. Therefore, after convergence to $\tilde{\boldsymbol{\lambda}}$ (which we have not yet shown is the dual optimal variable $\boldsymbol{\lambda}^*$) and the associated rate allocation, we have either $\tilde{\lambda}_l = 0$ or $\sum_{s:l\in L(s)} x_s(\tilde{\lambda}_l) = c_l$ or both, which is the complementary slackness condition of the partially constrained utility maximization problem (13). Together with the global search at each source in Step 2, the rate allocation after the link prices $\lambda_l, l \in \mathcal{L}$, converge maximizes the Lagrangian of (13) and also globally solves (13).

By induction, after all the link prices converge and $\mathcal{L}$ contains all the links in the network, problem (13) becomes (1), and the resulting rate allocation solves (1) globally.

Now by a similar development as in Appendix D, we can show that global convergence of bisection method implies zero duality gap, which further implies that $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}^*$. Since global convergence of bisection method implies continuity at $\boldsymbol{\lambda}^*$, by Theorem 2, the canonical distributed algorithm also converges. ■

Unlike the canonical distributed algorithm based on simultaneous subgradient update, the above bisection algorithm is based on sequential price update (*i.e.*, when one link changes its price, the other links do not change their prices), which is exponential time complexity in the number of links and has much slower convergence. Together with Theorem 6, the conclusion is that bisection method is never preferred to the canonical distributed algorithm.

## ACKNOWLEDGEMENT

The authors would like to acknowledge very helpful discussions with Jang Won Lee at Princeton, Wei Yu at the U. Toronto, Ness Shroff at Purdue, Venkat Anantharam at Berkeley, Dah Ming Chiu at CUHK, and Steven Low, John Doyle and their students at Caltech.

This work was in part supported by NSF Grants CCF-0440443, CNS-0417607, and CNS-0427677.

## REFERENCES

[1] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.
[2] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar, “Min common max crossing duality: a simple geometric framework for convex optimization and minimax theory,” *J. Op. Th. and App.*, Jan. 2002.
[3] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
[4] R. Cendrillon, W. Yu, M. Moonen, J. Verlinden, and T. Bostoen, “Optimal multiuser spectrum management for digital subscriber lines,” *Proc. IEEE ICC*, Paris, France, June 2004.
[5] M. Chiang, “Balacing transport and physical layers in wireless multihop networks: jointly optimal congestion control and power control”, *IEEE J. Sel. Area Comm.*, vol. 23, no. 1, 2005.
[6] D. Julian, M. Chiang, D. O’Neill, and S. Boyd, “QoS and fairness constrained convex optimization of resource allocation for wireless cellular and ad hoc networks,” *Proc. IEEE Infocom*, New York, USA, June 2002.
[7] M. Fazel, M. Chiang, and J. Doyle, “Network rate allocation by nonconcave utility maximization,” Preprint, 2005.
[8] F. P. Kelly, A. Maulloo, and D. Tan, “Rate control for communication networks: shadow prices, proportional fairness and stability,” *Journal of Operations Research Society*, vol. 49, no. 3, pp.237-252, March 1998.
[9] J. W. Lee, R. R. Mazumdar, and N. Shroff, “Non-convex optimization and rate control for multi-class services in the Internet,” *Proc. IEEE Infocom*, Hong Kong, China, March 2004.
[10] S. H. Low, “Duality congestion control,” *IEEE/ACM Tran. Networking*, Aug. 2003.
[11] R. T. Rockafellar, *Network Flows and Monotropic Programming*, Athena Scientific, 1998.
[12] R. T. Rockafellar, “Lagrange multipliers and optimality,” *SIAM Review*, vol. 35, pp. 183-283, 1993.
[13] S. Shenker, “Fundamental design issues for the future Internet,” *IEEE J. Sel. Area Comm.*, vol. 13, no. 7, pp. 1176-1188, Sept. 1995.
[14] W. Yu, R. Lui, and R. Cendrillon, “Dual optimization methods for multiuser orthogonal frequency division multiplex systems,” *Proc. IEEE Globecom*, Dallas, TX, Dec. 2004.