Alternative Decompositions for Distributed Maximization of Network Utility: Framework and Applications

Daniel P. Palomar and Mung Chiang
Electrical Engineering Department, Princeton University, NJ 08544, USA

***Abstract*— Network utility maximization (NUM) problems provide an important approach to conduct network resource management such as end-to-end rate allocation. In the existing literature, distributed implementations are typically achieved by the means of the so-called dual decomposition technique. However, the span of decomposition possibilities includes many other elements which thus far have not been fully exploited such as the use of the primal decomposition technique, the versatile introduction of auxiliary variables, and the potential of multilevel decompositions. This paper presents a systematic framework to exploit the potential of the alternative decomposition structures as a way to obtain different distributed algorithms, each with a different tradeoff among convergence speed, message passing amount and asymmetry, and distributed computation architecture. Many specific applications are considered to illustrate the proposed framework, including resource-constrained and direct-control rate allocation, and rate allocation among QoS classes and with multipath routing. For each of these applications, the associated generalized NUM formulation is first presented, followed by the development of novel alternative decompositions and numerical experiments on the resulting new distributed algorithms.**

Keywords: Rate control, Congestion control, Resource allocation, Mathematical programming/optimization, Network utility maximization, Distributed algorithm, Network control by pricing.

I. Introduction

A. Motivation

Why would one care about a systematic theory of alternative decompositions for variants of Network Utility Maximization (NUM) problems? There are two main reasons: it leads to the most appropriate distributed algorithm for a given network resource allocation problem, and it quantifies the comparison across architectural alternatives of distributed, layered network control.

First, since the publication of the seminal paper [1] by Kelly, Maulloo, and Tan in 1998, the framework of NUM has found many applications in network resource allocation algorithms and Internet congestion control protocols, e.g., [2], [3], [4], [5], [6], [7]. The key innovation from this series of work is to interpret source rates as primal variables, link congestion prices as dual variables, and a TCP–Active Queue Management (AQM) protocol as a distributed algorithm over the Internet to solve an implicit, global utility maximization and its Lagrange dual problem. Different TCP-AQM protocols solve for different concave utility functions using different link prices. This model implies that the equilibrium properties of a large network under TCP/AQM control, such as throughput, delay, queue lengths, loss probabilities, and fairness, can be readily understood by studying the underlying nonlinear utility maximization problem. In addition to this reverse engineering direction, allocation of limited network resources, such as power, bandwidth, and rate, among competing users can also be formulated by generalizing the basic NUM in [1] to more sophisticated formulations.

Almost all the papers in the vast, recent literature on NUM use a standard dual-based distributed algorithm. Contrary to the apparent impression that such a decomposition is the only possibility, there are in fact many alternatives to solve a given network utility problem in different but all distributed manners. Each of the alternatives provides a possibly different tradeoff among three important considerations: convergence speed, amount and asymmetry of message passing's communication overhead, and architecture of distributed computation. There is no universally 'best' way to distribute the solution process across a network: which alternative is the most desirable depends on the specific problem formulation. Thus motivated, we develop a systematic framework of alternative decompositions in this paper and apply it to four network rate allocation problems motivated by practical needs and constraints.

Second, the framework of NUM has recently been substantially extended from an analytic tool of reverse-engineering TCP congestion control to a general approach of understanding interactions across layers. One possible perspective to rigorously and holistically understand layering is to integrate the various protocol layers into a single coherent theory, by regarding them as carrying out an asynchronous distributed computation over the network to implicitly solve a global optimization problem. Different layers iterate on different subsets of the decision variables using local information to achieve individual optimality. Taken together, these local algorithms attempt to achieve a global objective. This approach exposes the interconnection between protocol layers and can be used to study rigorously the performance tradeoff in protocol layering, as different ways to distribute a centralized computation.

Since the design of a complex system will always be broken down into simpler modules, a 'layering as optimization decomposition' theory will allow us to systematically carry out this layering process and explicitly trade off design objectives.

Each different decomposition represents a new possibility of network architecture. But to develop such a theory, alternative decompositions must be fully explored to understand architectural possibilities, both 'vertically' across functional modules, i.e., the layers, and 'horizontally' across disparate network elements. This paper primarily studies alternatives of horizontal decompositions, although some results are directly applicable to vertical decompositions as well, e.g., the results in Section VI can be readily applied to joint TCP and MAC design in [8].

B. Existing Work

There are at least three levels of understanding as to what it means to 'efficiently solve' a utility maximization problem. First, a convex optimization (minimizing a convex function over a convex constraint set) is easy to solve because a local optimum must also be globally optimal, whereas a nonconvex one is very difficult [9]. Second, there are provably polynomial-time but centralized algorithms, such as the interior-point method, to solve a convex optimization. Third, distributed algorithms can be found to converge to the global optimum. It is the third level that we concern ourselves in this paper.

There is indeed a large body of results on distributed computation, some of which are summarized in standard textbooks such as [10], [11], [12], [13] and others. Our goal here is certainly not to survey these known general results in linear programming, graph theory, or decomposable problems. Instead we focus on the engineering problems of network rate allocation through problems in the form of nonlinear, coupled NUM, and develop novel distributed algorithms through a systematic method of alternative decompositions.

The seminal paper [1] in 1998 outlines two major classes of approaches to solve the basic version of NUM: primal-based and dual-based. It is important to note that both approaches in [1] adopt a differential equation technique, analyzed through penalty function and Lyapunov argument, thus different from the language of primal and dual decomposition in this paper.

Similar to one of the first publications in reverse-engineering TCP congestion control [6], many recent papers on distributed resource allocation with optimization models are based on Lagrangian relaxation and one-level, full dual decomposition. In fact, as illustrated in this paper through many applications, this standard dual decomposition is only one of the many choices one can make, including multi-level, indirect, and hybrid primal-dual decompositions. Despite its popularity, the standard dual decomposition may not be the best choice. It is also important to notice that the term 'primal-dual algorithm' is used in [3] to describe the purely dual-based algorithm because both the primal problem and the Lagrange dual problem are being solved simultaneously. This is different from both the primal-dual interior-point method in centralized solution of convex optimization [9] and the primal/dual decompositions for distributed algorithms developed in this paper.

Primal decomposition has remained in the shadow of dual decomposition and its employment is scarce, although it is just starting to take off in wireless transceiver design and power control problems. Recent examples include: [14], where linear transceivers for communication through MIMO channels were designed to minimize the average BER; [15], where linear MIMO transceivers were designed for multicarrier systems; [16] where different distributed algorithms were obtained in the context of wireless power allocation; and [17], where both a primal and a dual decomposition were considered for resource allocation. However, none of these publications present the decomposition alternatives for distributed rate allocation problems in Sections IV, V, and VI of this paper.

TABLE I

SUMMARY OF THE DECOMPOSITIONS CONSIDERED IN THE APPLICATIONS (● DENOTES EXISTING ALGORITHM AND ✓ NEW ALGORITHM).

Section	Primal	Full Dual	Partial Dual	Primal-Dual	Dual-Dual
II-E	-	●	-	-	-
III	-	-	-	●	●
IV	-	-	✓	✓	-
V	●	-	✓	-	-
VI	-	✓	●	✓	-

C. Summary of Results

We first present a systematic framework in Section II for alternative decompositions and how that would lead to an array of choices of distributed algorithms. Section II thus serves both as a review of the necessary background and a summary of our new extensions in decomposition theorems. In particular, Lemmas 1 and 3 extend existing results on subgradients, and the techniques of multilevel and indirect primal/dual decompositions are systematically introduced in the context of NUM problems.

The core of this paper then consists of Sections III to VI, covering four applications of distributed rate allocation: power-constrained rate allocation in Section III, rate allocation among different quality-of-service (QoS) groups in Section IV, hybrid rate-based and pricing-based rate allocation in Section V, and rate allocation with multipath routing in Section VI. In particular, the distributed algorithms obtained in Subsections IV-B, IV-C, V-C, VI-B, and VI-D are new. The types of decompositions developed in each application are summarized in Table I (when there are two levels of decompositions, they are separated by a dash, and for simplicity of terminology, we differentiate between full and partial dual decomposition in the name only in decompositions with one level).

In all applications, after the optimization formulation is clearly explained, we develop alternative decompositions and show the benefits of fully exploring the space of possible distributed algorithms. In some cases the distribution of computational load and asymmetry of message passing are much more desirable in one of the possible alternatives, and in other cases the convergence can be accelerated as confirmed in the numerical examples in Section VII.

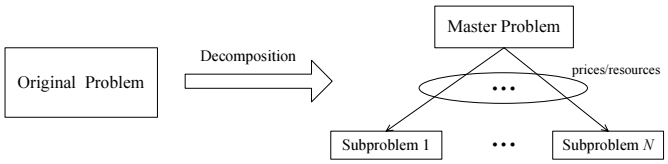

Fig. 1. Decomposition of a problem into several subproblems controlled by a master problem through prices (dual decomposition) or direct resource allocation (primal decomposition).

II. Systematic Framework for Decompositions: Review and Extensions

We first present a systematic framework to decompose a given optimization problem. In the rest of this paper after this section, we will see how different combinations of the basic elements in Subsections II-A to II-C lead to different distributed algorithms in network utility problems, among which one will typically be preferable to the others depending on the specific application.

While most of the concepts in this section are quick summaries of known results (e.g., Subsections II-D and II-E), a couple of extensions are also carried out (e.g., Lemmas 1 and 3) and some new techniques that will be very useful later in this paper are introduced (e.g., Subsections II-B and II-C).

The basic idea of a decomposition is to decompose the original large problem into distributively solvable subproblems which are then coordinated by a master problem by means of some kind of signalling (see Fig. 1) [13], [18], [10]. Most of the existing decomposition techniques can be classified into *primal decomposition* and *dual decomposition* methods. The former is based on decomposing the original primal problem, whereas the latter based on decomposing the Lagrange dual of the problem [19], [18]. Primal decomposition methods have the interpretation that the master problem directly gives each subproblem an amount of resources that it can use; the role of the master problem is then to properly allocate the existing resources. In dual decomposition methods, the master problem sets the price for the resources to each subproblem which has to decide the amount of resources to be used depending on the price; the role of the master problem is then to obtain the best pricing strategy.

A. Direct Primal and Dual Decompositions

A primal decomposition is appropriate when the problem has a coupling variable such that, when fixed to some value, the rest of the optimization problem decouples into several subproblems. Consider, for example, the following problem:

$$\begin{array}{ll} \underset{\mathbf{y},\{\mathbf{x}_i\}}{\text{maximize}} & \sum_i f_i(\mathbf{x}_i) \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i \qquad \forall i \\ & \mathbf{A}_i\mathbf{x}_i \le \mathbf{y} \\ & \mathbf{y} \in \mathcal{Y}. \end{array} \tag{1}$$

Clearly, if variable $\mathbf{y}$ were fixed, then the problem would decouple. Therefore, it makes sense to separate the optimization in (1) into two levels of optimization. At the lower level, we have the subproblems, one for each i, in which (1) decouples when $\mathbf{y}$ is fixed:

$$\begin{array}{ll} \underset{\mathbf{x}_i}{\text{maximize}} & f_i(\mathbf{x}_i) \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i \\ & \mathbf{A}_i\mathbf{x}_i \le \mathbf{y}. \end{array} \tag{2}$$

At the higher level, we have the master problem in charge of updating the coupling variable $\mathbf{y}$ by solving:

$$\begin{array}{ll} \underset{\mathbf{y}}{\text{maximize}} & \sum_i f_i^\star(\mathbf{y}) \\ \text{subject to} & \mathbf{y} \in \mathcal{Y} \end{array} \tag{3}$$

where $f_i^\star(\mathbf{y})$ is the optimal objective value of problem (2) for a given $\mathbf{y}$. If the original problem (1) is convex (meaning that the objective function is concave and the feasible set is convex), then the subproblems as well as the master problem are all convex.

If the function $\sum_i f_i^\star(\mathbf{y})$ is differentiable, then the master problem (3) can be solved with a gradient method. In general, however, the objective function $\sum_i f_i^\star(\mathbf{y})$ may be nondifferentiable and a subgradient method is a convenient approach which only requires the knowledge a subgradient[1] for each $f_i^\star(\mathbf{y})$ as given by [18, Sec. 6.4.2][13, Ch. 9]

$$\mathbf{s}_i(\mathbf{y}) = \boldsymbol{\lambda}_i^\star(\mathbf{y}), \tag{4}$$

where $\boldsymbol{\lambda}_i^\star(\mathbf{y})$ is the optimal Lagrange multiplier corresponding to the constraint $\mathbf{A}_i\mathbf{x}_i \le \mathbf{y}$ in problem (2). The global subgradient is then $\mathbf{s}(\mathbf{y}) = \sum_i \mathbf{s}_i(\mathbf{y}) = \sum_i \boldsymbol{\lambda}_i^\star(\mathbf{y})$. The subproblems in (2) can be locally and independently solved with the knowledge of $\mathbf{y}$.

A dual decomposition is appropriate when the problem has a coupling constraint such that, when relaxed, the optimization problem decouples into several subproblems. Consider, for example, the following problem:

$$\begin{array}{ll} \underset{\{\mathbf{x}_i\}}{\text{maximize}} & \sum_i f_i(\mathbf{x}_i) \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i \qquad \forall i, \\ & \sum_i \mathbf{h}_i(\mathbf{x}_i) \le \mathbf{c}. \end{array} \tag{5}$$

Clearly, if the constraint $\sum_i \mathbf{h}_i(\mathbf{x}_i) \le \mathbf{c}$ were absent, then the problem would decouple. Therefore, it makes sense to relax the coupling constraint in (5) as

$$\begin{array}{ll} \underset{\{\mathbf{x}_i\}}{\text{maximize}} & \sum_i f_i(\mathbf{x}_i) - \boldsymbol{\lambda}^T\left(\sum_i \mathbf{h}_i(\mathbf{x}_i) - \mathbf{c}\right) \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i \quad \forall i \end{array} \tag{6}$$

such that the optimization separates into two levels of optimization. At the lower level, we have the subproblems, one for each i, in which (6) decouples:

$$\begin{array}{ll} \underset{\mathbf{x}_i}{\text{maximize}} & f_i(\mathbf{x}_i) - \boldsymbol{\lambda}^T\mathbf{h}_i(\mathbf{x}_i) \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i. \end{array} \tag{7}$$

[1]Given a convex function f, a vector $\mathbf{s}$ is a *subgradient* of f at a point $\mathbf{x}$ if $f(\mathbf{z}) \ge f(\mathbf{x}) + (\mathbf{z}-\mathbf{x})^T\mathbf{s}$, $\forall \mathbf{z}$ [13], [18]. For a concave function, the inequality in the previous condition is in the opposite direction.

At the higher level, we have the master dual problem in charge of updating the dual variable $\boldsymbol{\lambda}$ by solving the dual problem:

$$\begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{minimize}} & g(\boldsymbol{\lambda}) = \sum_i g_i(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^T \mathbf{c} \\ \text{subject to} & \boldsymbol{\lambda} \geq \mathbf{0} \end{array} \tag{8}$$

where $g_i(\boldsymbol{\lambda})$ is the dual function obtained as the maximum value of the Lagrangian solved in (7) for a given $\boldsymbol{\lambda}$. This approach is in fact solving the dual problem instead of the original primal one. Hence, it will only give appropriate results if strong duality holds (e.g., when the original problem is convex optimization and there exists strictly feasible solutions [9]).

If the dual function $g(\boldsymbol{\lambda})$ is differentiable, then the master dual problem in (8) can be solved with a gradient method. In general, however, it may not be nondifferentiable and a subgradient method is a convenient approach which only requires the knowledge a subgradient for each $g_i(\boldsymbol{\lambda})$ as given by [18, Sec. 6.1]

$$\mathbf{s}_i(\boldsymbol{\lambda}) = -\mathbf{h}_i\left(\mathbf{x}_i^\star(\boldsymbol{\lambda})\right), \tag{9}$$

where $\mathbf{x}_i^\star(\boldsymbol{\lambda})$ is the optimal solution of problem (7) for a given $\boldsymbol{\lambda}$. The global subgradient is then $\mathbf{s}(\boldsymbol{\lambda}) = \sum_i \mathbf{s}_i(\mathbf{y}) + \mathbf{c} = \mathbf{c} - \sum_i \mathbf{h}_i\left(\mathbf{x}_i^\star(\boldsymbol{\lambda})\right)$). The subproblems in (7) can be locally and independently solved with knowledge of $\boldsymbol{\lambda}$.

General Results. We now present (skipping the proof due to space limit) the following new result to be used later in the paper, which generalizes the known result in [18, Sec. 6.4.2][13, Ch. 9] (where the particular result in (4) is obtained) and gives the subgradient for a more general case of primal decomposition:

Lemma 1: Consider the following function defined as the optimal value of a maximization problem:

$$f^\star(\mathbf{x}) \triangleq \sup_{\mathbf{y}: f_i(\mathbf{x},\mathbf{y}) \leq 0} f_0(\mathbf{x},\mathbf{y}) \tag{10}$$

where f_0 is concave, the f_i's are convex, and strong duality[2] holds for any given $\mathbf{x}$.

Then, $f^\star(\mathbf{x})$ is concave[3] and a subgradient is given by

$$\mathbf{s}_{\mathbf{x}}^\star(\mathbf{x}) = \mathbf{s}_{0,\mathbf{x}}\left(\mathbf{x},\mathbf{y}^\star(\mathbf{x})\right) - \mathbf{S}_{\mathbf{x}}\left(\mathbf{x},\mathbf{y}^\star(\mathbf{x})\right)\boldsymbol{\lambda}^\star(\mathbf{x}) \tag{11}$$

where $\mathbf{s}_{0,\mathbf{x}}(\mathbf{x},\mathbf{y})$ is a subgradient of $f_0(\mathbf{x},\mathbf{y})$ with respect to $\mathbf{x}$, $\mathbf{S}_{\mathbf{x}}(\mathbf{x},\mathbf{y})$ is a matrix containing in the ith column a subgradient of $f_i(\mathbf{x},\mathbf{y})$ with respect to $\mathbf{x}$, $\mathbf{y}^\star(\mathbf{x})$ is the value of $\mathbf{y}$ that achieves the supremum in (10) (assumed to exist) for a given $\mathbf{x}$, and $\boldsymbol{\lambda}^\star(\mathbf{x})$ is the optimal Lagrange multiplier associated with the constraints $f_i(\mathbf{x},\mathbf{y}) \leq 0$, $\forall i$, of the maximization in (10) (which is obtained 'for free' each time that $f^\star(\mathbf{x})$ is evaluated at some point).

We will also later need the following well-known result:

[2]Strong duality can be shown, for example, by Slater's condition, which simply requires (for any given $\mathbf{x}$) the existence of a point $\mathbf{y}$ that satisfies the constraints with strict inequality $f_i(\mathbf{x},\mathbf{y}) < 0$, $\forall i$.

[3]Proving concavity of $f^\star$ only requires concavity of f_0 and convexity of f_i, $\forall i$.

Lemma 2: Consider the following dual function defined as the supremum of a partial Lagrangian:

$$g(\boldsymbol{\lambda}) \triangleq \sup_{\mathbf{x}: g_i(\mathbf{x}) \leq 0} \left\{ f_0(\mathbf{x}) - \sum_i \lambda_i f_i(\mathbf{x}) \right\}. \tag{12}$$

Then, $g(\boldsymbol{\lambda})$ is convex and a subgradient, denoted by $\mathbf{s}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$, is given by

$$s_{\lambda_i}(\boldsymbol{\lambda}) = -f_i\left(\mathbf{x}^\star(\boldsymbol{\lambda})\right) \tag{13}$$

where $\mathbf{x}^\star(\boldsymbol{\lambda})$ is the value of $\mathbf{x}$ that achieves the supremum in (12) (assumed to exist) for a given $\boldsymbol{\lambda}$ (which is obtained 'for free' each time that $g(\boldsymbol{\lambda})$ is evaluated at some point).

Note that if there is a unique value $\mathbf{x}^\star(\boldsymbol{\lambda})$ that achieves the supremum in (12) for any given $\boldsymbol{\lambda}$, then $g(\boldsymbol{\lambda})$ is differentiable and $\nabla g(\boldsymbol{\lambda}) = \mathbf{s}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$ (this happens, for example, if f_0 is strictly concave and the f_i;s are linear) [18, Prop. 6.1.1].

B. Indirect Primal and Dual Decompositions

Often the problem can be reformulated and more effective primal and dual decompositions can be indirectly applied. The introduction of auxiliary variables provides much flexibility in terms of choosing a primal or a dual decomposition and the resulting distributed algorithm.

The basic techniques are illustrated as follows. Problem (1) contains a coupling variable and was decoupled in (2)-(3) via a primal decomposition approach. However, it can also be solved with a dual decomposition by first introducing the additional variables $\{\mathbf{y}_i\}$:

$$\begin{array}{lll} \underset{\{\mathbf{y}_i\},\{\mathbf{x}_i\}}{\text{maximize}} & \sum_i f_i(\mathbf{x}_i,\mathbf{y}_i) & \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i & \forall i \\ & \mathbf{A}_i\mathbf{x}_i \leq \mathbf{y}_i & \\ & \mathbf{y}_i = \mathbf{y} & \\ & \mathbf{y} \in \mathcal{Y}. & \end{array} \tag{14}$$

This way, we have transformed the coupling variable $\mathbf{y}$ into a set of coupling constraints $\mathbf{y}_i = \mathbf{y}$ which can be dealt with using a dual decomposition.

Consider now problem (5) which contains a coupling constraint and was decoupled in (7)-(8) via a dual decomposition. By introducing again additional variables $\{\mathbf{y}_i\}$ the problem becomes:

$$\begin{array}{lll} \underset{\{\{\mathbf{y}_i\}\mathbf{x}_i\}}{\text{maximize}} & \sum_i f_i(\mathbf{x}_i) & \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i & \forall i, \\ & \mathbf{h}_i(\mathbf{x}_i) \leq \mathbf{y}_i & \\ & \sum_i \mathbf{y}_i \leq \mathbf{c}. & \end{array} \tag{15}$$

This way, we have transformed the coupling constraint $\sum_i \mathbf{h}_i(\mathbf{x}_i) \leq \mathbf{c}$ into a coupling variable $\mathbf{y} = \left[\mathbf{y}_1^T, \cdots, \mathbf{y}_N^T\right]^T$ which can be dealt with using a primal decomposition.

C. Multilevel Primal and Dual Decompositions

An important technique that leads to alternatives of distributed architectures is to apply primal/dual decompositions recursively: The basic decompositions are repeatedly applied to a problem to obtain smaller and smaller subproblems as

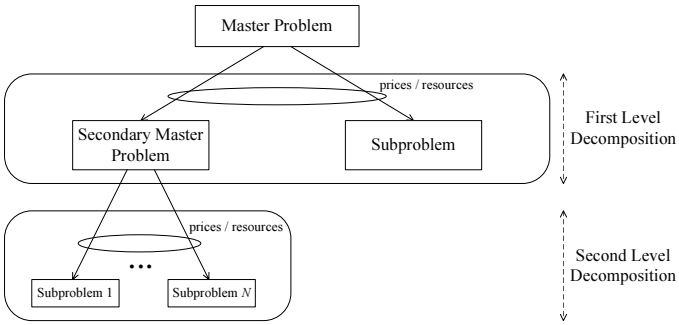

Fig. 2. Example of a multilevel primal/dual decomposition with two levels.

illustrated in Fig. 2. For example, consider the following problem which includes both a coupling variable and a coupling constraint:

$$\begin{array}{ll} \underset{\mathbf{y},\{\mathbf{x}_i\}}{\text{maximize}} & \sum_i f_i(\mathbf{x}_i, \mathbf{y}) \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i \qquad \forall i \\ & \sum_i \mathbf{h}_i(\mathbf{x}_i) \leq \mathbf{c} \\ & \mathbf{A}_i \mathbf{x}_i \leq \mathbf{y} \\ & \mathbf{y} \in \mathcal{Y}. \end{array} \tag{16}$$

One way to decouple this problem is by first taking a primal decomposition with respect to the coupling variable $\mathbf{y}$ and then a dual decomposition with respect to the coupling constraint $\sum_i \mathbf{h}_i(\mathbf{x}_i) \leq \mathbf{c}$. This would produce a three-level optimization problem: a master primal problem, a secondary master dual problem, and the subproblems. Observe that an alternative approach would be to first take a dual decomposition and then a primal one.

Another example that shows flexibility in terms of different decompositions is the following problem with two sets of constraints:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{maximize}} & f_0(\mathbf{x}) \\ \text{subject to} & f_i(\mathbf{x}) \leq 0 \qquad \forall i \\ & g_i(\mathbf{x}) \leq 0. \end{array} \tag{17}$$

One way to deal with this problem is via the dual problem with a full relaxation of both sets of constraints to obtain the dual function $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$. At this point, instead of minimizing g directly with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, it can be minimized over only one set of Lagrange multipliers first and then over the remaining one: $\min_{\boldsymbol{\lambda}} \min_{\boldsymbol{\mu}} g(\boldsymbol{\lambda}, \boldsymbol{\mu})$. This approach corresponds to first applying a full dual decomposition and then a primal one on the dual problem. The following new result (proved through Lemmas 1 and 2) characterizes the subgradient of the master problem at the top level.

Lemma 3: Consider the following partial minimization of the dual function

$$g(\boldsymbol{\lambda}) = \inf_{\boldsymbol{\mu}} g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{18}$$

where $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is the dual function defined as

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \triangleq \sup_{\mathbf{x} \in \mathcal{X}} \left\{ f_0(\mathbf{x}) - \sum_i \lambda_i f_i(\mathbf{x}) - \sum_i \mu_i g_i(\mathbf{x}) \right\}. \tag{19}$$

Then, $g(\boldsymbol{\lambda})$ is convex and a subgradient, denoted by $\mathbf{s}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$, is given by

$$s_{\lambda_i}(\boldsymbol{\lambda}) = -f_i(\mathbf{x}^\star(\boldsymbol{\lambda}, \boldsymbol{\mu}^\star(\boldsymbol{\lambda}))) \tag{20}$$

where $\mathbf{x}^\star(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is the value of $\mathbf{x}$ that achieves the supremum in (19) (assumed to exist) for a given $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, and $\boldsymbol{\mu}^\star(\boldsymbol{\lambda})$ is the value of $\boldsymbol{\mu}$ that achieves the infimum in (18) (also assumed to exist).

Alternatively, problem (17) can be approached via the dual but with a partial relaxation of only one set of constraint, say $f_i(\mathbf{x}) \leq 0 \ \forall i$, obtaining the dual function $g(\boldsymbol{\lambda})$ to be minimized by the master problem. Observe now that in order to compute $g(\boldsymbol{\lambda})$ for a given $\boldsymbol{\lambda}$, the partial Lagrangian has to be maximized subject to the remaining constraints $g_i(\mathbf{x}) \leq 0 \ \forall i$, for which yet another relaxation can be used. This approach corresponds to first applying a partial dual decomposition and then, for the subproblem, another dual decomposition.

When there is more than one level of decomposition, and all levels conduct some type of iterative algorithms, such as the subgradient method, convergence and stability are guaranteed if the lower level master problem is solved on a faster timescale than the higher level master problem, so that at each iteration of a master problem all the problems at a lower level have already converged. If the updates of the different subproblems operate on similar timescales, convergence of the overall system can still be guaranteed under certain technical conditions [20], [10], and indeed is observed empirically in the numerical examples to be presented later in this paper.

D. Review: Subgradient Method

After performing a decomposition, the resulting master problem is generally nondifferentiable. Subgradient methods arise then as excellent approaches to solve these nondifferentiable problems: they simply require the value of a subgradient at any given point [19], [18]. Subgradient methods are distinguished by their simplicity, little requirements of memory usage, and amenability for parallel implementation [19], [18], which are precisely the main interests in this paper. Consider the following general concave maximization over convex constraint set:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{maximize}} & f_0(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{X}. \end{array} \tag{21}$$

The subgradient method generates a sequence of feasible points $\{\mathbf{x}(t)\}$ as [18, Sec. 6.3.1]:

$$\mathbf{x}(t+1) = [\mathbf{x}(t) + \alpha(t)\mathbf{s}(t)]_{\mathcal{X}} \tag{22}$$

where $\mathbf{s}(t)$ is a subgradient of $f_0(\mathbf{x})$ at $\mathbf{x}(t)$, $[\cdot]_{\mathcal{X}}$ denotes the projection onto the feasible convex set $\mathcal{X}$, and $\alpha(t)$ is a positive scalar stepsize. Such an iteration looks like a gradient projection method except that a subgradient is used instead of the gradient (which may not exist). However, there is a fundamental difference: each new iteration may not improve the objective value as happens with a gradient method. What makes the subgradient method work is that for sufficiently

small stepsize $\alpha(t)$, the distance of the current solution $\mathbf{x}(t)$ to the optimal solution $\mathbf{x}^\star$ decreases.

There are many results on convergence of the subgradient method [19], [18]. For constant step size $\alpha(t) = \alpha$ and constant step length $\alpha(t) = \alpha/\|\mathbf{s}(t)\|$, the subgradient algorithm is guaranteed to converge to within some range of the optimal value; in other words, the subgradient method finds an ϵ-suboptimal point within a finite number of steps. For the diminishing step size rule

$$\alpha(t) = \frac{1+m}{t+m},$$

where m is a fixed nonnegative number, the algorithm is guaranteed to converge to the optimal value.

E. Review: Standard Dual-Based Algorithm for Basic NUM

Before concluding this section on a systematic framework of alternative decompositions, we briefly review the standard way [3] to solve the basic NUM problem [1] for distributed end-to-end rate allocation, which illustrates a simple application of the one-level, full dual decomposition. In the rest of this paper, we will see a number of more sophisticated NUM formulations motivated by new application contexts and a much richer array of decomposition alternatives, beyond the well-known problem and solution method in this subsection.

Consider a communication network with L links, each with a fixed capacity of c_l, and S sources or nodes, each transmitting at a source rate of x_s. Each source s emits one flow, using a fixed set of links $L(s)$ in its path, and has a utility function $U_s(x_s)$. NUM is the problem of maximizing the total utility $\sum_s U_s(x_s)$, over the source rates $\mathbf{x}$, subject to linear flow constraints $\sum_{s:l\in L(s)} x_s \le c_l$ for all links l:

$$\begin{array}{ll} \underset{\mathbf{x}\ge\mathbf{0}}{\text{maximize}} & \sum_s U_s(x_s) \\ \text{subject to} & \sum_{s:l\in L(s)} x_s \le c_l \quad \forall l \end{array} \tag{23}$$

where the utilities U_s are strictly concave functions (the problem is therefore a convex optimization).

The standard distributed solution to (23) is based on a dual decomposition. We first form the Lagrangian of (23):

$$\begin{aligned} L(\mathbf{x},\boldsymbol{\lambda}) &= \sum_s U_s(x_s) + \sum_l \lambda_l \left(c_l - \sum_{s:l\in L(s)} x_s\right) \\ &= \sum_s \left[U_s(x_s) - \left(\sum_{l\in L(s)} \lambda_l\right) x_s\right] + \sum_l c_l\lambda_l \\ &= \sum_s L_s(x_s,\lambda^s) + \sum_l c_l\lambda_l \end{aligned} \tag{24}$$

where $\lambda_l \ge 0$ is the Lagrange multiplier (link price) associated with the linear flow constraint on link l, $\lambda^s = \sum_{l\in L(s)} \lambda_l$ is the aggregate path congestion price of those links used by source s, and $L_s(x_s,\lambda^s) = U_s(x_s) - \lambda^s x_s$ is the sth Lagrangian to be maximized by the sth source.

The dual decomposition results then in each source s solving, for the given λ^s:

$$x_s^\star(\lambda^s) = \arg\max_{x_s\ge 0} [U_s(x_s) - \lambda^s x_s] \quad \forall s \tag{25}$$

which is unique due to the strict concavity of U_s. The master dual problem is

$$\begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{minimize}} & g(\boldsymbol{\lambda}) = \sum_s g_s(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^T\mathbf{c} \\ \text{subject to} & \boldsymbol{\lambda} \ge \mathbf{0} \end{array} \tag{26}$$

where $g_s(\boldsymbol{\lambda}) = L_s(x_s^\star(\lambda^s),\lambda^s)$. Since the solution in (25) is unique, it follows that the dual function $g(\boldsymbol{\lambda})$ is differentiable and the following gradient method can be used:

$$\lambda_l(t+1) = \left[\lambda_l(t) - \alpha\left(c_l - \sum_{s:l\in L(s)} x_s^\star(\lambda^s(t))\right)\right]^+ \quad \forall l \tag{27}$$

where t is the iteration index, $\alpha > 0$ is a positive stepsize, and $[\cdot]^+$ denotes the projection onto the nonnegative orthant. Note that the term $c_l - \sum_{s:l\in L(s)} x_s^\star(\lambda^s(t))$ is the gradient of $g(\boldsymbol{\lambda})$ with respect to λ_l.

The dual variable $\boldsymbol{\lambda}(t)$ will converge to the dual optimal $\boldsymbol{\lambda}^\star$ as $t\to\infty$ and, since the duality gap for (23) is zero and the solution to (25) is unique, the primal variable $\mathbf{x}^\star(\boldsymbol{\lambda}(t))$ will also converge to the primal optimal variable $\mathbf{x}^\star$.

Summarizing, the original basic NUM problem in (23) can be distributively solved with the subgradient update in (27) carried out independently by each the link and the maximization in (25) solved independently by each source. Notice that there is no need for explicit message passing since λ^s can be measured by each source s as queuing delay and $\sum_{s:l\in L(s)} x_s$ can be measured by each link l as the total traffic load.

III. Application 1: Power-Constrained Rate Allocation

We start the applications sections with the simplest and recently studied extension of the basic NUM, before moving on to more involved formulations and novel solutions in Sections IV, V, and VI.

A. Problem Formulation

In some applications such as wireless broadcast or DSL access, distributed rate allocation can be carried out over transmission 'pipes' of different sizes, with the help of adaptive resource allocation in the physical layer. This is an example of balancing 'supply' of resources and 'demand' of link capacities 'built' from the limited resources.

Consider now the basic NUM in (23) but with variable link capacities $\{c_l(p_l)\}$, each of which depends on the allocated resource p_l, such as transmit power, with a constraint on the maximum total resource P_T. For many models such as TDMA

or FDMA, c_l is a strictly concave function[4]

$$\begin{array}{ll}\underset{\mathbf{x},\mathbf{p}\geq\mathbf{0}}{\text{maximize}} & \sum_s U_s(x_s) \\ \text{subject to} & \sum_{s:l\in L(s)} x_s \leq c_l(p_l) \quad \forall l \\ & \sum_l p_l \leq P_T.\end{array} \tag{28}$$

Although only slightly more sophisticated than the basic NUM, this problem already contains sufficient elements such that one can try different decompositions. We will consider two decompositions: a primal decomposition with respect to the power allocation, and a dual decomposition with respect to the flow constraints.

B. Primal-Dual Decomposition

Consider first a primal decomposition of (28) by fixing the power allocation $\mathbf{p}$. Clearly, the link capacities become fixed numbers and problem (28) becomes a basic NUM like (23), which can be solved via a dual decomposition as explained in Subsection II-E. The master primal problem is

$$\begin{array}{ll}\underset{\mathbf{p}\geq\mathbf{0}}{\text{maximize}} & U^\star(\mathbf{p}) \\ \text{subject to} & \sum_l p_l \leq P_T,\end{array} \tag{29}$$

where $U^\star(\mathbf{p})$ is the optimal objective value of (28) for a given $\mathbf{p}$. Since a subgradient of $U^\star(\mathbf{p})$ with respect to c_l is given by the Lagrange multiplier λ_l associated with the constraint $\sum_{s:l\in L(s)} x_s \leq c_l$ in (28), it follows that a subgradient of $U^\star(\mathbf{p})$ with respect to p_l is given by $\lambda_l c_l'(p_l)$. Therefore, the master primal problem (29) can be solved with a subgradient method by updating the powers as

$$\mathbf{p}(t+1) = \left[\mathbf{p}(t) + \alpha \begin{bmatrix} \lambda_1^\star(\mathbf{p}(t))\, c_1'(p_1(t)) \\ \vdots \\ \lambda_L^\star(\mathbf{p}(t))\, c_L'(p_L(t)) \end{bmatrix}\right]_{\mathcal{P}} \tag{30}$$

where $[\cdot]_{\mathcal{P}}$ denotes the projection onto the feasible convex set $\mathcal{P} \triangleq \{\mathbf{p} : \mathbf{p}\geq\mathbf{0}, \sum_l p_l \leq P_T\}$, which is a simplex. Due to the projection, this subgradient update cannot be performed independently by each link and requires some centralized approach. The projection of a point $\mathbf{p}_0$ (the expression inside the outer bracket in (30)) onto the simplex $\mathcal{P}$, i.e., $\mathbf{p} = [\mathbf{p}_0]_{\mathcal{P}}$, can be easily obtained in the following waterfilling form [15]:

$$p_l = \left(p_l^0 - \gamma\right)^+ \quad \forall l \tag{31}$$

where the waterlevel γ is chosen as the minimum nonnegative value such that $\sum_l p_l \leq P_T$. Observe that only the computation of γ requires a central node since the update of each power p_l can be done at each link.

C. Dual-Dual Decomposition

Consider now a dual decomposition of (28) by relaxing the flow constraints $\sum_{s:l\in L(s)} x_s \leq c_l(p_l)$:

$$\begin{array}{ll}\underset{\mathbf{x},\mathbf{p}\geq\mathbf{0}}{\text{maximize}} & \sum_s \left[U_s(x_s) - \left(\sum_{l\in L(s)} \lambda_l\right) x_s\right] + \sum_l c_l(p_l)\,\lambda_l \\ \text{subject to} & \sum_l p_l \leq P_T.\end{array} \tag{32}$$

This problem decomposes into one maximization for each source, as (25) in the basic NUM, plus the following additional maximization to update the power allocation:

$$\begin{array}{ll}\underset{\mathbf{p}\geq\mathbf{0}}{\text{maximize}} & \sum_l \lambda_l c_l(p_l) \\ \text{subject to} & \sum_l p_l \leq P_T\end{array} \tag{33}$$

which can be further decomposed via a second-level dual decomposition yielding the following subproblems

$$\underset{p_l\geq 0}{\text{maximize}} \quad \lambda_l c_l(p_l) - \gamma p_l \tag{34}$$

with solution given by

$$p_l = (c_l')^{-1}(\gamma/\lambda_l) \tag{35}$$

and a secondary master dual problem that updates the dual variable γ as

$$\gamma(t+1) = \left[\gamma(t) - \alpha\left(P_T - \sum_l p_l^\star(\gamma(t))\right)\right]^+. \tag{36}$$

The master dual problem is updated as in the standard NUM (27).

D. Summary

We have obtained two different distributed algorithms for power-constrained rate allocation in (28):

- primal-dual decomposition: the master problem (29) is solved with the subgradient power update in (30) carried out by the links with a small central coordination (due to the projection on the simplex) and then, for a given set of powers, the resulting basic NUM is solved via the standard dual-based decomposition in (25) and (27). This implies two levels of decompositions: on the highest level there is a master primal problem, on a second level there is a secondary master dual problem, and on the lowest level the subproblems.
- dual-dual decomposition: the master dual problem is solved with the standard price update in (27) which is carried out independently by each link and then, for a given set of prices, each source solves its own subproblem as in (25) and subproblem (33) is solved with some central node updating the price with (36) and each link obtaining the optimal power with (35). This approach contains two levels of decompositions: on the highest level there is a master dual problem, on a second level there are rate subproblems and a secondary master dual problem, and on the lowest level the power subproblems.

In both approaches, the only explicit signaling required is the power-price γ from the central unit to the links and possibly the powers from the links back to the central node.

E. Special Case: Cellular Downlink Power/Rate Control

An interesting special case of the signal model in (28) arises in cellular downlink power/rate control with the flow

[4]A related and different model has been recently studied in [21]. The primal-dual solution in Subsection III.B was first proposed in [17], and that in Subsection III.C was first proposed in [22].

constraints on each downlink connection modeled in the high SNR regime of a CDMA system with orthogonal codes:

$$\begin{array}{ll}\underset{\mathbf{x},\mathbf{p}\geq\mathbf{0}}{\text{maximize}} & \sum_s U_s\left(x_s\right) \\ \text{subject to} & x_s \leq \log\left(g_s p_s\right) \quad \forall s \\ & \sum_s p_s \leq P_T\end{array} \tag{37}$$

where g_s is the channel gain of the sth user. This problem can be solved in many different combinations of multilevel primal-dual decompositions, each with a different signalling scheme and convergence speed (see Subsection VII-A for an empirical comparison of the convergence of several methods).

IV. Application 2: QoS Rate Allocation

A. Problem Formulation

Sometimes a rate allocation mechanism needs to differentiate users in different QoS classes. For example, the total link capacity received by each QoS class must lie within a range prescribed in the service level agreement. Such constraints introduce new coupling to the basic NUM problem and lead to alternative decomposition possibilities. We will see in this section two different distributed algorithms to solve this type of QoS rate allocation problem, both with a differential pricing interpretation to the new set of Lagrange multiplier introduced. Therefore, these algorithms provide an intuitive pricing alternative to the recent proposals of NUM-based rate allocation among different QoS classes in [23], [24].

Consider now the basic NUM in (23) but with different classes of users that will be treated differently. The idea of having several classes of users is, for example, to impose limits on the maximum rate and to guarantee a minimum rate for each class. To simplify the exposition we consider only two classes of users, but the results extend straightforwardly to more classes of users. Denoting by $y_l^{(1)}$ and $y_l^{(2)}$ the aggregate rates of classes 1 and 2, respectively, along the lth link, the problem formulation is

$$\begin{array}{ll}\underset{\mathbf{x},\mathbf{y}^{(1)},\mathbf{y}^{(2)}\geq\mathbf{0}}{\text{maximize}} & \sum_s U_s\left(x_s\right) \\ \text{subject to} & \sum_{s\in S_i:l\in L(s)} x_s = y_l^{(i)} \quad \forall l, i=1,2 \\ & \mathbf{y}^{(1)}+\mathbf{y}^{(2)} \leq \mathbf{c} \\ & \mathbf{c}_{\min}^{(i)} \leq \mathbf{y}^{(i)} \leq \mathbf{c}_{\max}^{(i)}.\end{array} \tag{38}$$

Observe that in the absence of the constraints $\mathbf{c}_{\min}^{(i)} \leq \mathbf{y}^{(i)} \leq \mathbf{c}_{\max}^{(i)}$, problem (38) becomes the basic NUM in (23). Also note that if problem (38) is feasible, then the equality flow constraints can be rewritten as inequality flow constraints, as we will hereinafter assume. We will consider two decompositions: a primal decomposition with respect to the aggregate rate of each class, and a dual decomposition with respect to the total aggregate rate constraints from both classes.

B. Primal-Dual Decomposition

Consider first a primal decomposition of (38) by fixing the aggregate rates $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$. Problem (38) becomes two independent subproblems, for $i=1,2$, identical to the basic NUM in (23):

$$\begin{array}{ll}\underset{\mathbf{x}\geq\mathbf{0}}{\text{maximize}} & \sum_{s\in S_i} U_s\left(x_s\right) \\ \text{subject to} & \sum_{s\in S_i:l\in L(s)} x_s \leq y_l^{(i)} \quad \forall l\end{array} \tag{39}$$

where the fixed aggregate rates $y_l^{(i)}$ play the role of the fixed link capacities in the basic NUM of (23). These two independent basic NUMs can be solved as explained in Subsection II-E.

The master primal problem is

$$\begin{array}{ll}\underset{\mathbf{y}^{(1)},\mathbf{y}^{(2)}\geq\mathbf{0}}{\text{maximize}} & U_1^\star\left(\mathbf{y}^{(1)}\right)+U_2^\star\left(\mathbf{y}^{(2)}\right) \\ \text{subject to} & \mathbf{y}^{(1)}+\mathbf{y}^{(2)} \leq \mathbf{c} \\ & \mathbf{c}_{\min}^{(i)} \leq \mathbf{y}^{(i)} \leq \mathbf{c}_{\max}^{(i)} \quad i=1,2\end{array} \tag{40}$$

where $U_i^\star\left(\mathbf{y}^{(i)}\right)$ is the optimal objective value of (39) for a given $\mathbf{y}^{(i)}$, with a subgradient given by the Lagrange multiplier $\boldsymbol{\lambda}^{(i)}$ associated to the constraints $\sum_{s\in S_i:l\in L(s)} x_s \leq y_l^{(i)}$ in (39). Observe that $\boldsymbol{\lambda}^{(i)}$ is the differential set of prices for the QoS class i. The master primal problem (40) can now be solved with a subgradient method by updating the aggregate rates as

$$\begin{bmatrix}\mathbf{y}^{(1)}(t+1)\\ \mathbf{y}^{(2)}(t+1)\end{bmatrix} = \left[\begin{bmatrix}\mathbf{y}^{(1)}(t)\\ \mathbf{y}^{(2)}(t)\end{bmatrix} + \alpha\begin{bmatrix}\boldsymbol{\lambda}^{\star(1)}\left(\mathbf{y}^{(1)}(t)\right)\\ \boldsymbol{\lambda}^{\star(2)}\left(\mathbf{y}^{(2)}(t)\right)\end{bmatrix}\right]_{\mathcal{Y}} \tag{41}$$

where $[\cdot]_{\mathcal{Y}}$ denotes the projection onto the feasible convex set $\mathcal{Y} \triangleq \left\{\left(\mathbf{y}^{(1)},\mathbf{y}^{(2)}\right):\ \mathbf{y}^{(1)}+\mathbf{y}^{(2)}\leq\mathbf{c},\ \mathbf{c}_{\min}^{(i)}\leq\mathbf{y}^{(i)}\leq\mathbf{c}_{\max}^{(i)}\ i=1,2\right\}$. Nicely enough, this feasible set enjoys the property that it already naturally decomposes into a Cartesian product for each of the links: $\mathcal{Y}=\mathcal{Y}_1\times\cdots\times\mathcal{Y}_L$. Therefore, this subgradient update can be performed independently by each link simply with the knowledge of its corresponding Lagrange multipliers $\lambda_l^{(1)}$ and $\lambda_l^{(2)}$, which in turn are also updated independently by each link as in Subsection II-E.

C. Partial Dual Decomposition

Consider now a dual decomposition of (38) by relaxing the flow constraints $\sum_{s\in S_i:l\in L(s)} x_s \leq y_l^{(i)}$:

$$\begin{array}{ll}\underset{\mathbf{x},\mathbf{y}^{(1)},\mathbf{y}^{(2)}\geq\mathbf{0}}{\text{maximize}} & \sum_{s\in S_1}\left[U_s\left(x_s\right)-\left(\sum_{l\in L(s)}\lambda_l\right)x_s\right] \\ & \quad+\sum_{s\in S_2}\left[U_s\left(x_s\right)-\left(\sum_{l\in L(s)}\lambda_l\right)x_s\right] \\ & \quad+\boldsymbol{\lambda}^{(1)T}\mathbf{y}^{(1)}+\boldsymbol{\lambda}^{(2)T}\mathbf{y}^{(2)} \\ \text{subject to} & \mathbf{y}^{(1)}+\mathbf{y}^{(2)}\leq\mathbf{c} \\ & \mathbf{c}_{\min}^{(i)}\leq\mathbf{y}^{(i)}\leq\mathbf{c}_{\max}^{(i)} \quad i=1,2.\end{array} \tag{42}$$

This problem decomposes into one maximization for each source, as (25) in the basic NUM, plus the following additional maximization to update the aggregate rates:

$$\begin{array}{ll}\underset{\mathbf{y}^{(1)},\mathbf{y}^{(2)}\geq\mathbf{0}}{\text{maximize}} & \boldsymbol{\lambda}^{(1)T}\mathbf{y}^{(1)}+\boldsymbol{\lambda}^{(2)T}\mathbf{y}^{(2)} \\ \text{subject to} & \mathbf{y}^{(1)}+\mathbf{y}^{(2)}\leq\mathbf{c} \\ & \mathbf{c}_{\min}^{(i)}\leq\mathbf{y}^{(i)}\leq\mathbf{c}_{\max}^{(i)} \quad i=1,2\end{array} \tag{43}$$

which can be solved independently by each link with knowledge of its corresponding Lagrange multipliers $\lambda_l^{(1)}$ and $\lambda_l^{(2)}$, which in turn are also updated independently by each link (c.f. Subsection II-E).

The master dual problem corresponding to this dual decomposition is updated with the following subgradient method (similarly to (27)):

$$\lambda_l^{(i)}(t+1) = \left[\lambda_l^{(i)}(t) - \alpha\left(y_l^{(i)}(t) - \sum_{s\in S_i: l\in L(s)} x_s^\star(\lambda^{(i)s}(t))\right)\right]^+ \quad \forall l, i = 1,2. \tag{44}$$

D. Summary

We have obtained two different distributed algorithms for rate allocation among QoS classes in (38):

- primal-dual decomposition: the master problem (40) is solved with the subgradient update for the aggregate rate in (41) carried out independently by each of the links and then, for a given set of aggregate rates, the two resulting basic NUMs are independently solved via the standard dual-based decomposition in (25) and (27). This implies two levels of decompositions: on the highest level there is a master primal problem, on a second level there is a secondary master dual problem, and on the lowest level the subproblems. There is no explicit signaling required.
- partial dual decomposition: the master dual problem is solved with the standard price update for each class in (44) which is carried out independently by each link and then, for a given set of prices, each source solves its own subproblem as in (25) and subproblem (43) is independently solved by each link. This approach contains only one level of decomposition and no explicit signaling is required.

Observe that in the primal-dual decomposition approach each link updates the aggregate rates on a slower timescale and the prices on a faster timescale, whereas in the partial dual decomposition approach each link updates the prices on a slower timescale and the aggregate rates on a faster timescale (actually in one shot); therefore, the speed of convergence of the partial dual approach should be faster in general. In both cases, the users are privy of the existence of classes and only the links have to take this into account by having one price for each class. In other words, this is a way to give each class of users a different price than the one based on the standard dual-based algorithm so that they can be further controlled. The next application hinges on this observation.

V. Application 3: Hybrid Rate-Based and Pricing-Based Rate Allocation

A. Problem Formulation

One extreme way to control the rate allocation process is to directly give each source the rate they can use, at the expense of a centralized computation. At the other extreme, we can optimize the system in a fully distributed way via pricing, as in the basic NUM of Subsection II-E, at the expense of trusting the sources even though they can be non-cooperative and try to obtain more bandwidth by using a more aggressive utility function. Neither of these two extreme approaches is completely satisfactory in all applications, and hybrid solutions between rate-based and window-based rate allocation are desirable for both robustness of fair allocation against aggressive users and speed of converging to the correct rate allocation equilibrium.

New congestion control protocols using direct rate allocation have recently been proposed, such as RCP [25] that is based on a heuristic computation of the processor-sharing type of rate allocation by each router that a flow traverses. We now describe a systematic method to perform distributed and direct rate allocation to each user. The key idea is to use the approach of Section IV but with one class for each user.

The problem formulation becomes

$$\begin{array}{ll} \underset{\mathbf{x},\{\mathbf{y}^{(s)}\}\geq\mathbf{0}}{\text{maximize}} & \sum_s U_s(x_s) \\ \text{subject to} & x_s \leq y_l^{(s)} \qquad \forall s, l\in L(s) \\ & \sum_s \mathbf{y}^{(s)} \leq \mathbf{c} \\ & \mathbf{c}_{\min}^{(s)} \leq \mathbf{y}^{(s)} \leq \mathbf{c}_{\max}^{(s)}. \end{array} \tag{45}$$

Note that if a source s does not use a path l, then $y_l^{(s)}$ is taken as zero in the constraint $\sum_s \mathbf{y}^{(s)} \leq \mathbf{c}$.

B. Primal Decomposition

If we now take a primal decomposition approach, then the master primal problem will be in charge of the update of $y_l^{(s)}$ and each user will simply choose x_s equal to the minimum of the $y_l^{(s)}$ along its path in order to satisfy $x_s \leq y_l^{(s)}\ \forall l \in L(s)$. This approach constitutes in fact one of the extreme methods in which each user is directly given the amount of bandwidth it can use.

C. Partial Dual Decomposition

We may also take a dual decomposition approach by relaxing the flow constraints

$$\begin{array}{ll} \underset{\mathbf{x},\{\mathbf{y}^{(s)}\}\geq\mathbf{0}}{\text{maximize}} & \sum_s \left[U_s(x_s) - \left(\sum_{l\in L(s)} \lambda_l^{(s)}\right) x_s\right] \\ & \quad + \sum_s \sum_{l\in L(s)} \lambda_l^{(s)} y_l^{(s)} \\ \text{subject to} & \sum_s \mathbf{y}^{(s)} \leq \mathbf{c} \\ & \mathbf{c}_{\min}^{(s)} \leq \mathbf{y}^{(s)} \leq \mathbf{c}_{\max}^{(s)} \qquad \forall s. \end{array} \tag{46}$$

This problem decomposes into one maximization for each source, as (25) in the basic NUM, with $\lambda^s = \sum_{l\in L(s)} \lambda_l^{(s)}$ being the aggregate path price specific for user s, plus the following additional rate-bounding maximization to obtain the $y_l^{(s)}$, for each link l:

$$\begin{array}{ll} \underset{\{y_l^{(s)}\}\geq 0}{\text{maximize}} & \sum_{s:l\in L(s)} \lambda_l^{(s)} y_l^{(s)} \\ \text{subject to} & \sum_{s:l\in L(s)} y_l^{(s)} \leq c_l \\ & c_{l,\min}^{(s)} \leq y_l^{(s)} \leq c_{l,\max}^{(s)} \qquad \forall s: l\in L(s). \end{array} \tag{47}$$

This problem can be solved independently by each link as a way to distribute its capacity c_l among the sources using the

link according to the weights given by the prices $\lambda_l^{(s)}$, which are different for each source.

The master dual problem corresponding to this dual decomposition is solved with the following subgradient price update step (similarly to (27)):

$$\lambda_l^{(s)}(t+1) = \left[\lambda_l^{(s)}(t) - \alpha\left(y_l^{(s)}(t) - x_s^\star\left(\lambda^s(t)\right)\right)\right]^+ \quad \forall l, s : l \in L(s). \tag{48}$$

D. Summary

We have explored different decompositions for the hybrid rate/pricing-based rate allocation in (45):

- primal decomposition: it leads to a direct rate allocation and is based on one level of decomposition. This approach requires the signaling to inform each user what rate to transmit at.
- partial dual decomposition: the master dual problem is solved with the price update in (48) which is carried out independently by each link and then, for a given set of prices, each source solves its own subproblem as in (25) and the bounding rates of subproblem (47) are also obtained independently by each link. This approach only shows one level of decomposition and does not require any explicit signaling. It is a hybrid of rate-bounding and pricing-feedback mechanisms.

VI. Application 4: Multipath-Routing Rate Allocation

A. Problem Formulation

Consider now a more general setup of the basic NUM of Subsection II-E where each source can choose among several possible paths (possibly using a weighted combination of them). The structure of a network with S sources, L links, and J paths can be summarized with the $L \times J$ path availability $0-1$ matrix $\mathbf{H}$ defined by

$$[\mathbf{H}]_{l,j} = \begin{cases} 1 & \text{if the } j\text{th path uses the } l\text{th link} \\ 0 & \text{otherwise} \end{cases}$$

together with the $J \times S$ path choice nonnegative matrix $\mathbf{W}$[5] defined by

$$[\mathbf{W}]_{j,s} = \begin{cases} w_{js} & \text{if the } s\text{th source uses the } j\text{th path} \\ 0 & \text{otherwise} \end{cases}$$

where w_{js} indicates the percentage of the rate of the sth user allocated to the jth path and has to satisfy $w_{js} > 0$ and $\sum_j w_{js} = 1$. These two matrices can be combined into the routing matrix $\mathbf{R} = \mathbf{HW}$ that tells how much each source is using each link.

[5]This notation follows that in [26]. However, the problem being considered here is to design rate allocation algorithm with a fixed $\mathbf{H}$ and $\mathbf{W}$, whereas the problem considered in [26] is to analyze the effect of joint routing and rate allocation with $\mathbf{W}$ being a variable.

To start with, the problem can be directly formulated with the routing matrix $\mathbf{R}$ like the basic NUM in (23):

$$\begin{array}{ll} \underset{\mathbf{x}\geq\mathbf{0}}{\text{maximize}} & \sum_s U_s(x_s) \\ \text{subject to} & \mathbf{Rx} \leq \mathbf{c} \end{array} \tag{49}$$

and then the standard dual-based decomposition algorithm can be used. We will later see that it may be more flexible to formulate the problem alternatively in terms of $\mathbf{H}$ and $\mathbf{W}$ as follows:

$$\begin{array}{lll} \underset{\mathbf{x},\mathbf{y}\geq\mathbf{0}}{\text{maximize}} & \sum_s U_s(x_s) & \\ \text{subject to} & \mathbf{Wx} \leq \mathbf{y} & \text{(path constraint)} \\ & \mathbf{Hy} \leq \mathbf{c} & \text{(link constraint)} \end{array} \tag{50}$$

where y_l contains the aggregate rate along the lth path.

B. Primal-Dual Decomposition

We can now consider a primal decomposition approach of (50) by fixing the path rates $\mathbf{y}$. Problem (50) becomes then a basic NUM where $\mathbf{y}$ plays the role of the link capacities in (23). This problem can be solved via the standard dual-based algorithm as reviewed in Subsection II-E.

The master primal problem is

$$\begin{array}{ll} \underset{\mathbf{y}\geq\mathbf{0}}{\text{maximize}} & U^\star(\mathbf{y}) \\ \text{subject to} & \mathbf{Hy} \leq \mathbf{c} \end{array} \tag{51}$$

where $U^\star(\mathbf{y})$ is the optimal objective value of (50) for a given $\mathbf{y}$, with subgradient given by the Lagrange multiplier $\boldsymbol{\lambda}$ associated to the constraints $\mathbf{Wx} \leq \mathbf{y}$ in (50). As usual, the master primal problem (51) can be solved with a subgradient method by updating the path rates as

$$\mathbf{y}(t+1) = \left[\mathbf{y}(t) + \alpha\boldsymbol{\lambda}^\star(\mathbf{y}(t))\right]_{\mathcal{Y}} \tag{52}$$

where $[\cdot]_{\mathcal{Y}}$ denotes the projection onto the feasible convex set $\mathcal{Y} \triangleq \{\mathbf{y} : \mathbf{y} \geq \mathbf{0}, \mathbf{Hy} \leq \mathbf{c}\}$. In principle, this subgradient update cannot be performed independently by each path due to the projection onto $\mathcal{Y}$, which makes it impractical.

C. Partial Dual Decomposition

We can also take a partial dual decomposition of (50) by relaxing only the constraint $\mathbf{Wx} \leq \mathbf{y}$ (similarly to [8]):

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{y}\geq\mathbf{0}}{\text{maximize}} & \sum_s U_s(x_s) + \boldsymbol{\gamma}^T(\mathbf{y} - \mathbf{Wx}) \\ \text{subject to} & \mathbf{Hy} \leq \mathbf{c}. \end{array} \tag{53}$$

This problem decomposes into one maximization for the sources as in (25) for the basic NUM:

$$\underset{\mathbf{x}\geq\mathbf{0}}{\text{maximize}} \ \sum_s \left[U_s(x_s) - \gamma^s x_s\right], \tag{54}$$

where $\gamma^s = \boldsymbol{\gamma}^T\mathbf{W}_{:,s} = \sum_{j\in J(s)} \gamma_j w_{js}$ is the aggregate price for the sth source, plus one maximization for the path rates:

$$\begin{array}{ll} \underset{\mathbf{y}\geq\mathbf{0}}{\text{maximize}} & \boldsymbol{\gamma}^T\mathbf{y} \\ \text{subject to} & \mathbf{Hy} \leq \mathbf{c} \end{array} \tag{55}$$

which has to be solved in a centralized way.

The master dual problem updates the prices as

$$\boldsymbol{\gamma}(t+1) = \left[\boldsymbol{\gamma}(t) - \alpha\left(\mathbf{y} - \mathbf{W}\mathbf{x}(\boldsymbol{\gamma}(t))\right)\right]^+. \tag{56}$$

D. Full Dual Decomposition

Yet another different way to solve problem (50) is with a full dual decomposition by relaxing both constraints $\mathbf{Wx} \leq \mathbf{y}$ and $\mathbf{Hy} \leq \mathbf{c}$:

$$\underset{\mathbf{x},\mathbf{y}\geq\mathbf{0}}{\text{maximize}} \ \sum_s U_s(x_s) + \boldsymbol{\gamma}^T(\mathbf{y} - \mathbf{Wx}) + \boldsymbol{\lambda}^T(\mathbf{c} - \mathbf{Hy}) \tag{57}$$

which can be rewritten as

$$\underset{\mathbf{x},\mathbf{y}\geq\mathbf{0}}{\text{maximize}} \ \sum_s [U_s(x_s) - x_s\gamma^s] + \sum_j y_j(\gamma_j - \lambda^j) + \boldsymbol{\lambda}^T\mathbf{c} \tag{58}$$

where $\lambda^j = \boldsymbol{\lambda}^T\mathbf{H}_{:,j} = \sum_{l\in L(j)} \lambda_l$ is the aggregate price of the jth path and $\gamma^s = \boldsymbol{\gamma}^T\mathbf{W}_{:,s} = \sum_{j\in J(s)} \gamma_j w_{js}$ is the aggregate price for the sth source. This problem separates into a maximization over $\mathbf{x}$, as in (25) for the basic NUM, and a maximization over $\mathbf{y}$, which is unbounded unless $\gamma_j = \lambda^j$. Therefore, the optimal choice for the master dual problem is $\gamma_j = \lambda^j$ and then $\gamma^s = \sum_{j\in J(s)} \lambda^j w_{js} = \sum_{j\in J(s)} w_{js} \sum_{l\in L(j)} \lambda_l$. Hence, this approach reduces to the standard dual-based algorithm applied to problem (49).

Now consider a variant of this rate allocation problem with multipath-routing, where the objective of Internet Service Provider (ISP) is combined with the end user utility objective. In today's operating environment of the Internet, the ISP controlling each Autonomous System tries to minimize a total convex cost function of the link utilizations [27]. Suppose the cost function is quadratic, and the network utility maximization is now formulated as maximizing the weighted difference between end user utility and ISP cost:

$$\sum_s U_s(x_s) - \theta\mathbf{y}^T\mathbf{y} \tag{59}$$

where θ is the weight. Observe that by taking θ sufficiently small the quadratic term becomes negligible and we are back to the original problem (50).

Repeating the same full relaxation as before, one gets the following maximization problem:

$$\begin{aligned}\underset{\mathbf{x},\mathbf{y}\geq\mathbf{0}}{\text{maximize}} \ & \sum_s [U_s(x_s) - x_s\gamma^s] \\ & + \sum_j y_j(\gamma_j - \lambda^j) - \theta\mathbf{y}^T\mathbf{y} + \boldsymbol{\lambda}^T\mathbf{c}.\end{aligned} \tag{60}$$

This problem separates as before into a maximization over $\mathbf{x}$, as in (25) for the basic NUM, and a maximization over $\mathbf{y}$ with optimal solution given by

$$y_j = \frac{1}{2\theta}(\gamma_j - \lambda^j) \qquad \forall j. \tag{61}$$

Then, the master dual problem has to update two sets of prices:

$$\boldsymbol{\lambda}(t+1) = [\boldsymbol{\lambda}(t) - \alpha(\mathbf{c} - \mathbf{Hy}(t))]^+ \tag{62}$$

$$\boldsymbol{\gamma}(t+1) = [\boldsymbol{\gamma}(t) - \alpha(\mathbf{y}(t) - \mathbf{Wx}^\star(\boldsymbol{\gamma}(t)))]^+ \tag{63}$$

where $\mathbf{y}(t) \triangleq \mathbf{y}^\star(\boldsymbol{\lambda}(t), \boldsymbol{\gamma}(t))$ is the optimal $\mathbf{y}$ for the given $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\gamma}(t)$ as in (61). Note that the matrix-vector products can be conveniently written as $[\mathbf{Hy}]_l = \sum_{j:l\in L(j)} y_j$ and $[\mathbf{Wx}]_j = \sum_{s:j\in J(s)} w_{js}x_s$.

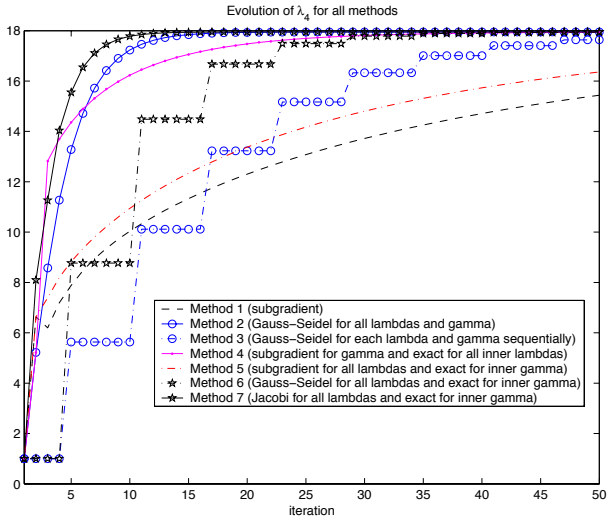

Fig. 3. Evolution of λ_4 for the seven methods based on a dual decomposition.

E. Summary

We have explored several possibilities for distributed algorithms for rate allocation with multipath-routing possibilities in (50):

- standard dual decomposition: by reformulating the problem as in (49) we recover the basic NUM formulation and the standard dual-based algorithm can be readily used.
- primal-dual decomposition: the master primal problem (51) is solved with the path rate subgradient update in (52) and then, for a given set of path rates, the resulting basic NUMs is solved via the standard dual-based decomposition in (25) and (27). Unfortunately, due to the projection in (52) a centralized computation is required, which makes this approach impractical.
- partial dual decomposition: the master dual problem is solved with the price update in (56) and then, for a given set of prices, each source solves its own subproblem as in (54) and subproblem (55) is solved in a centralized way, making this approach also inconvenient.
- full dual decomposition: the master dual problem is solved with the price updates in (62)-(63) and then, for a given set of prices, each source solves its own subproblem as in (54) and the path rates are obtained as in (61). This approach contains one level of decomposition: on the higher level the master dual problem and on the lower level the source-rate and path-rate subproblems. Explicit signaling is required for the update of the price $\boldsymbol{\gamma}(t)$ in (63), and for the computation of the path rate in (61) (which can be done either at the receiver of the path or through heuristic-based computation distributed across routers along the path).

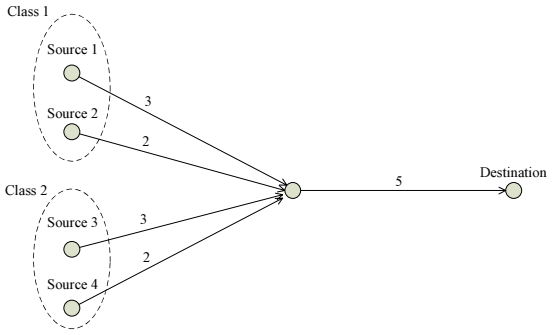

Fig. 4. Block diagram of the considered example of NUM with priorities.

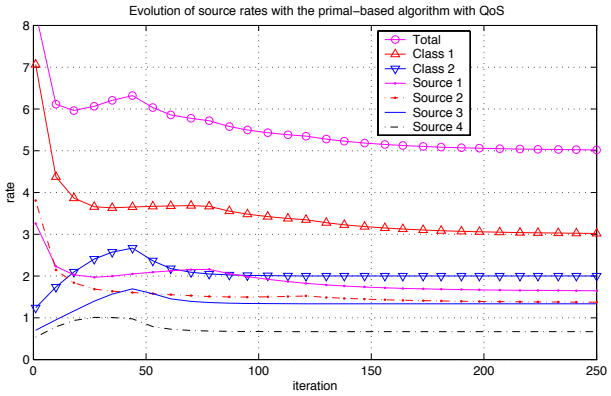

Fig. 5. Evolution of the rates with the primal-based algorithm for a NUM with QoS control.

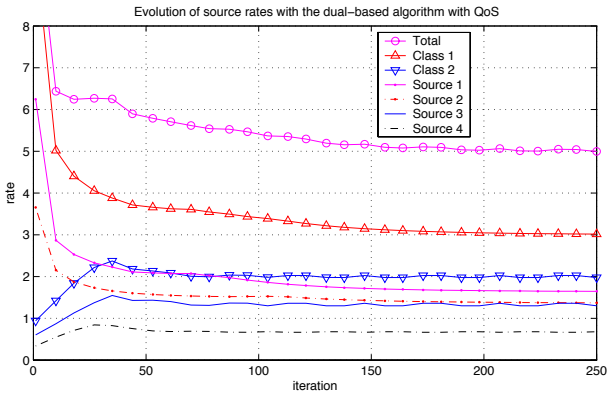

Fig. 6. Evolution of the rates with the dual-based algorithm for a NUM with QoS control.

VII. Numerical Examples

A. Downlink Power/Rate Control

The purpose of this subsection is to illustrate the convergence behavior of different decomposition approaches can be quite different, using the downlink power/rate control formulated in (37) as the context. Fig. 3 shows the evolution of one of the dual variables (λ, γ) under the iterations for seven methods based on various combinations of primal/dual multilevel decompositions and variants of Gauss-Seidel and Jacobi iterations of subgradient calculations. Without going into the details due to space limit, the seven methods are respectively based on the following decompositions: full dual + subgradient for $\boldsymbol{\lambda}$ and γ, full dual + Gauss-Seidel for $\boldsymbol{\lambda}$ and γ, full dual + Gauss-Seidel for each λ_i and γ, dual-primal (for $\boldsymbol{\lambda}$ and γ), dual-primal (for γ and $\boldsymbol{\lambda}$), dual-primal + Gauss-Seidel, dual-primal + Jacobi.

B. QoS Rate Allocation

To illustrate the distributed algorithms for a rate allocation among QoS classes (as in Section IV), we consider a simple example consisting of four sources transmitting to the same destination and sharing a common link as shown in Fig. 4. Users in class 1 are very aggressive, with utility functions $U_1(x) = 12\log(x)$ and $U_2(x) = 10\log(x)$, whereas users in class 2 are not aggressive, with utility functions $U_3(x) = 2\log(x)$ and $U_4(x) = \log(x)$. If no QoS control is included in the design and the standard dual-based distributed algorithm of Subsection II-E is used, then the aggressive users of class 1 get most of the available capacity in the common link. In particular, class 1 gets a rate of 4.5 out of the total available rate of 5, leaving class 2 only with a rate of 0.5. This is precisely the kind of unfair behavior that can be avoided with QoS control.

Figs. 5 and 6 show the evolution of the rates of the sources when QoS control is included in the distributed algorithms based on a primal decomposition and on a dual decomposition, respectively (as described in Section IV). In particular, the rate for each class has been limited to 3. As can be observed, the rate of class 1 now tends to the limit of 3 and, since the link capacity is 5, class 2 is left with a rate of 2 (as opposed to 0.5 obtained without QoS control). Hence, the distribution of the total rate between both classes is more fair. Both primal-based and dual-based algorithms show a similar convergence (a constant stepsize of 0.05 was used for all subgradient updates). Note that the primal-based algorithm contains two levels of subgradient updates and, in principle, the inner subgradient algorithm should run until convergence before updating the outer subgradient. In practice, however, this is not necessary and both subgradients can run simultaneously (in general using a smaller stepsize for the outer subgradient so that it works on a slower timescale).

C. Multipath-Routing Rate Allocation

We now consider a NUM with different grouping of the path and link constraints as described in Section VI. In particular, we generate a random network topology with $S = 4$ sources, $J = 12$ paths, and $L = 36$ links, such that each user uses 3 paths and each path uses 5 links.

Fig. 7 shows the evolution of the rates of the sources for the standard dual-based algorithm based directly on the routing matrix $\mathbf{R} = \mathbf{HW}$. Fig. 8 shows the evolution of the rates of the sources with a full dual-based algorithm (including the quadratic term $-\theta \mathbf{y}^T \mathbf{y}$ with $\theta = 0.001$), which follows closely the performance of the standard algorithm. In practice, the optimal solution for the path rates in (61) leads to a large dynamic range that can lead to instability; this can be

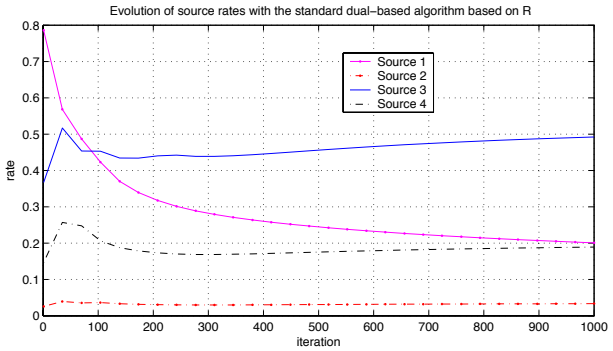

Fig. 7. Evolution of the rates with the standard dual-based algorithm for a NUM based directly on the routing matrix $\mathbf{R}$.

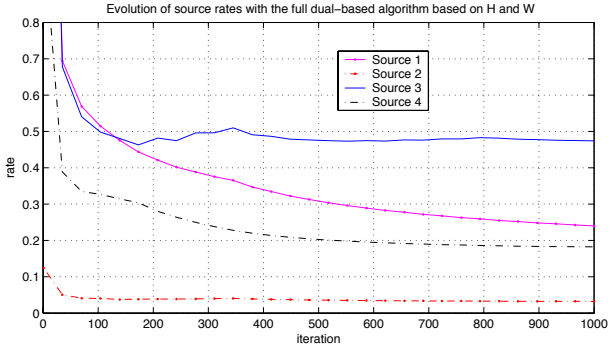

Fig. 8. Evolution of the rates with the full-dual-based algorithm for a NUM based on the path-link $\mathbf{H}$ and link-source $\mathbf{W}$ matrices.

easily avoided by providing the update with some memory (controlled by the forgetting factor β):

$$y_j(t+1) = \beta \frac{1}{2\theta}\left(\gamma_j(t) - \lambda^j(t)\right) + (1-\beta)\, y_j(t) \quad \forall j. \tag{64}$$

The other two methods described in Section VI, based on a primal decomposition and on a partial dual decomposition, provide similar convergence trajectories. However, their complexity and the need for centralized computation make them impractical (due to the projection in (52) and to the resolution of problem (55), respectively).

VIII. Conclusions

Recent focus in the literature on the standard dual-based method notwithstanding, there are more than one way to solve a network utility maximization in a distributed manner. A systematic framework is developed in this paper to explore alternative decompositions, and four specific rate allocation applications are presented. Implications of these results include designing faster distributed algorithm with the right distribution of computation and communication load across network elements, developing five new congestion control algorithms under various practical constraints, and understanding architectural tradeoffs in distributed network control.

References

[1] 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.

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

[3] S. H. Low, "A duality model of TCP and queue management algorithms," *IEEE/ACM Trans. Networking*, vol. 11, no. 4, pp. 525–536, Aug. 2003.

[4] S. Kunniyur and R. Srikant, "End-to-end congestion control: utility functions, random losses and ECN marks," *IEEE/ACM Trans. Networking*, vol. 10, no. 5, pp. 689–702, Oct. 2003.

[5] R. J. La and V. Anantharam, "Utility-based rate control in the internet for elastic traffic," *IEEE/ACM Trans. Networking*, vol. 9, no. 2, pp. 272–286, April 2002.

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

[7] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Networking*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[8] L. Chen, S. H. Low, and J. C. Doyle, "Joint congestion control and media access control design for ad hoc wireless networks," in *Proc. IEEE INFOCOM*, Miami, FL, USA, March 13-17, 2005.

[9] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.

[10] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[11] J. L. R. Ford and D. R. Fulkerson, *Flow in Networks*. Princeton, NJ: Princeton University Press, 1962.

[12] D. P. Bertsekas and R. G. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice Hall, 1987.

[13] L. S. Lasdon, *Optimization Theory for Large Systems*. New York: Macmillian, 1970.

[14] D. P. Palomar, M. Bengtsson, and B. Ottersten, "Minimum BER linear transceivers for MIMO channels via primal decomposition," *accepted in IEEE Trans. Signal Processing*, to appear 2005.

[15] D. P. Palomar, "Convex primal decomposition applied to linear MIMO transceiver design," *accepted in IEEE Trans. Signal Processing*, to appear 2005.

[16] D. P. Palomar and M. Chiang, "Choices of distributed algorithms for wireless network resource allocation," to appear in *Proc. IEEE Globecom 2005*, St. Louis, MO, USA, Nov. 2005.

[17] B. Johansson and M. Johansson, "Primal and dual approaches to distributed cross-layer optimization," in *Proc. 16th IFAC World Congress*, Prague, Czech republic, 2005.

[18] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.

[19] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*. Berlin: Springer-Verlag, 1985.

[20] R. T. Rockafellar, "Saddle-points and convex analysis," in *Differential Games and Related Topics*, H. W. Kuhn and G. P. Szego, Eds. North-Holland, 1971.

[21] M. Chiang, "Balancing transport and physical layer in wireless multihop networks: Jointly optimal congestion control and power control," *IEEE J. Select. Areas Commun.*, vol. 23, no. 1, Jan. 2005.

[22] L. Xiao, M. Johansson, and S. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *IEEE Trans. Commun.*, vol. 52, no. 7, pp. 1136–1144, July 2004.

[23] K. Chandrayana and S. Kalyanaraman, "Uncooperative congestion control," in *Proc. ACM SIGMETRICS*, New York, NY, June 2004.

[24] M. Chiang, S. Zhang, and P. Hande, "Distributed rate allocation for inelastic flows: Optimization frameworks," in *Proc. IEEE INFOCOM*, Miami, FL, USA, March 13-17, 2005.

[25] N. Dukkipati, M. Kobayashi, R. Z. Shen, and N. McKeown, "Processor sharing flows in the internet," in *Proc. International Workshop on Quality of Service*, Passau, Germany, June 2005.

[26] J. Wang, L. Li, S. H. Low, and J. C. Doyle, "Can TCP and shortest path routing maximize utility," in *Proc. IEEE INFOCOM*, San Francisco, CA, April 2003.

[27] J. Rexford, "Route optimization in IP networks," in *Handbook of Optimization in Telecommunications*. Kluwer Academic Publishers, 2005.