Recovery of Primal Solution in Dual Subgradient Schemes

by

Jing Ma

Submitted to the School of Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Computation for Design and Optimization

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2007

Author .....
School of Engineering
August 16, 2007

Certified by .....
Asuman E. Ozdaglar
Associate Professor of EECS
Class of 1943 Career Development Professor
Thesis Supervisor

Accepted by .....
Jaime Peraire
Professor of Aeronautics and Astronautics
Codirector, Computation for Design and Optimization

Recovery of Primal Solution in Dual Subgradient Schemes

by

Jing Ma

Submitted to the School of Engineering
on August 16, 2007, in partial fulfillment of the
requirements for the degree of
Master of Science in Computation for Design and Optimization

Abstract

In this thesis, we study primal solutions for general optimization problems. In particular, we employ the subgradient method to solve the Lagrangian dual of a convex constrained problem, and use a primal-averaging scheme to obtain near-optimal and near-feasible primal solutions. We numerically evaluate the performance of the scheme in the framework of Network Utility Maximization (NUM), which has recently drawn great research interest. Specifically for the NUM problems, which can have concave or nonconcave utility functions and linear constraints, we apply the dual-based decentralized subgradient method with averaging to estimate the rate allocation for individual users in a distributed manner, due to its decomposability structure.

Unlike the existing literature on primal recovery schemes, we use a constant stepsize rule in view of its simplicity and practical significance. Under the Slater condition, we develop a way to effectively reduce the amount of feasibility violation at the approximate primal solutions, namely, by increasing the value initial dual iterate; moreover, we extend the established convergence results in the convex case to the more general and realistic situation where the objective function is convex. In particular, we explore the asymptotical convergence properties of the averaging sequence, the tradeoffs involved in the selection of parameter values, the estimation of duality gap for particular functions, and the bounds for the amount of constraint violation and value of primal cost per iteration. Numerical experiments performed on NUM problems with both concave and nonconcave utility functions show that, the averaging scheme is more robust in providing near-optimal and near-feasible primal solutions, and it has consistently better performance than other schemes in most of the test instances.

Thesis Supervisor: Asuman E. Ozdaglar
Title: Associate Professor of EECS
Class of 1943 Career Development Professor

Acknowledgments

First of all, I would like to thank my supervisor Prof. Asu Ozdaglar for offering me the opportunity to work with her. I sincerely appreciate her continuous guidance and help throughout this research project. It would have not been possible for me to complete the thesis without her insightful advise and support. Things I have learned from her are invaluable and long-lasting.

Thanks are also due to Attila for the useful discussions on the second half of my thesis. Also I am very grateful to the friends I have at MIT. I truly thank them for all the encouragement and care throughout my stay here.

Last but not least, I want to express my deepest gratitude to my beloved family, Yinqiang Ma, Guirong Tang and Chinsoon Lim, for their unconditional love and endless support. It is them who teach me the meaning of life, help me understand myself better and keep me going all the way till today. I dedicate this thesis to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations and Literature Review

Over the past few decades, there have been many powerful computational tools to generate approximate solutions for nonlinear programming problems. In particular, to solve the large-scale and structured optimization problems, Lagrangian dual formulation together with the subgradient method has been efficiently used and is one of the most popular schemes in the field of computational optimization. Based on the fundamental framework, we develop the dual subgradient scheme with primal-averaging for our study in the thesis.

1.1.1 Duality and Subgradient Method

Firstly, why do we use *duality* to solve the optimization problems?

The Lagrange relaxation and dual formulation can not only efficiently generate near-optimal solutions for the convex problems, but also provide good lower bounds on the optimal cost of the original optimization problems in the nonconvex case (see M. Dür [10]). Especially for the structured problems arising in the network applications, the dual problem has at least the following advantages:

- The dual problem is always convex regardless of the primal, hence its optimal

solution can be found.

- The dual problem typically has a smaller dimension and simpler constraints than its primal.
- The dual formulation can decouple the primal problem of certain structures and thus can be implemented distributively.

Secondly, why do we use *subgradient methods*?

The subgradient method, which relies strongly on the convexity properties of the dual problem, is used as the dual function is typically nondifferentiable (cf. the gradient method applied to differentiable functions; see Bertsekas [2]). As such, dual subgradient method may not always guarantee a direction of descent, but wander around and reduce the distance to the optimal solution at each iteration. A diminishing stepsize is essential for the convergence of subgradient methods. With a constant stepsize, while the gradient algorithm can converge asymptotically (i.e., in the limit when the number of iterations increases to infinity) to the optimal value provided that the stepsize is sufficiently small, the subgradient algorithm can only approach the best value within some range of the optimal value which is proportional of the stepsize (see Bertsekas, Nedić and Ozdaglar [4] or Bertsekas [2]). In short, we cannot guarantee convergence of the subgradient method with a constant stepsize.

Then, why do we use the *constant stepsize* in our analysis?

- It is easy to implement and can be constructed at minor cost.
- It can track the system variations and set termination conditions more efficiently than other stepsize rules.
- It has practical importance.

There has been much literature on the subgradient method and its extensions since Shor first introduces it in [24]. For example, Bertsekas, Nedić, and Ozdaglar study

the convergence properties under different stepsize rules in [4]; Nedić and Bertsekas propose and analyze the incremental subgradient method in [21]; Nemirovski discusses cutting plan methods and mirror descent in [3]. For a general reference on subgradient methods, see the notes of Boyd [5].

1.1.2 Approximate Primal Solutions

Finally, why do we use an *average sequence*?

In spite of its widespread popularity, the dual subgradient method may not necessarily solve the primal problem, which is of our main concern. Some reasons are stated as the following:

- The exact dual solution may not be obtained with a constant stepsize (recall that the dual problem converges within some error level), not to mention the convergence in the primal space.
- When the original problem is nonconvex, the duality gap usually exists and to solve the dual is no longer equivalent to the primal, which may make it difficult to set termination criteria to generate near-feasible and near-optimal primal solutions.
- In general, the dual cannot solve linear programs, since the solutions to the Lagrangian relaxed problems are normally infeasible in the original linear program (see Larsson and Liu [16]).
- The dual function is typically nonsmooth at an optimal dual solution, therefore an optimal primal solution is usually a nontrivial convex combination of the extreme subproblem solutions.

In order to efficiently tackle these problems and directly recover the primal solutions from the subgradient information generated in dual subgradient methods, a primal-averaging scheme is firstly proposed for linear primal problems by Shor [24].

Subsequently, Larsson and Liu [16] implement and numerically test two averaging schemes for the linear minimum cost multicommodity network flow problem. These results are generalized to a more flexible choice of step lengths and extended to the deflected subgradient method by Sherali and Choi [26] for linear problems. This utilization of ergodic sequence for primal convergence is explored in the case of general convex programs by Larsson, Patriksson and A. Strömberg [17], and the two schemes proposed (with different combinations of step lengths and weights) are investigated in an application to traffic equilibrium assignment under road pricing, which shows promising results.

Unlike these previous work which almost exclusively focus on the asymptotic behavior of the primal sequence under diminishing stepsize rules, Nedić and Ozdaglar [22] recently propose and analyze the dual subgradient methods with averaging under the constant stepsize rule, so as to generate approximate primal solutions for convex optimization problems. Moreover, in their work, the convergence properties per iteration, including bounds for feasibility violation and estimate of primal cost, are obtained under the Slater condition. To complement these theoretical analyses, we numerically test and evaluate the proposed scheme in the application of network utility maximization (NUM) problems, which we make a brief review below.

1.1.3 Network Applications

The framework of NUM proposed by Kelly, Maulloo and Tan [14] has many applications, for instance, network rate allocation, user behavior models, and the Internet congestion control in communication networks. Due to the special structure of this model, dual subgradient methods have been used with great success in developing decentralized cross-layer resource allocation mechanisms (see Chiang et al. [8]).

In modeling NUM problems, the objective is to maximize utility functions which represent the level of user satisfaction or quality of service (QoS) at the the allocated rate. For the elastic traffic flows where the users can adjust their transmission rate gradually, concave utility functions are used. Some practical examples include traditional data services, such as file transfer and e-mail. Concavity is also guaranteed

when the source rate is large enough, due to the law of diminishing marginal utility. This class of concave NUM problem can be efficiently solved by the canonical distributed algorithm proposed by Low and Lapsley [19].

However, even though concavity is widely assumed in the NUM problem, nonconcave utilities that model inelastic flows have become increasingly important. These models have found applications in delay and rate sensitive services, such as voice service and streaming video. Unlike the concave case, inducing the primal convergence for the nonconcave NUM is a difficult nonconvex optimization problem and is not widely studied. Currently, there are three different approaches in previous work to solve the nonconcave NUM.

Approach 1 (Rate-allocation scheme with self-regulation property) This approach enforces certain regulating properties locally without central admission control to provide distributed heuristics. However, the solution is suboptimal and the algorithm works only for a specific class of sigmoidal utilities; particularly, sigmoidal with only one inflexion point (see Lee et al. [18]).

Approach 2 (Sum-of-squares (SOS) method) This method uses SOS to efficiently compute global optimum for objectives that can be transformed into polynomial utilities. However, it is centralized and only numerically tested for some functions, without being theoretically proved (see Fazel et al. [11]).

Approach 3 (Optimality conditions) This approach determines the optimality conditions in which the canonical distributed algorithm for convex case is still applicable for the nonconvex problem. However, this condition cannot be satisfied in many cases (see Chiang et al. [9]).

Due to the limitations of these recent approaches and the unexplored aspects in existing literatures, our work is motivated.

1.2 Scope of the Thesis

In this thesis, our goal is to study primal solutions recovered in the dual subgradient with primal-averaging scheme, for general optimization problems. We numerically evaluate performance of the averaging scheme in the Network Utility Maximization (NUM) framework. We examine the established theoretical analysis in the convex case and develop an effective way to reduce infeasibility of the approximate primal solutions. We extend the insights from behaviors of the convex implementation to the nonconvex case, and further explore convergence properties of the approximate primal sequence when the duality gap is possibly positive. We compare the averaging scheme with other models in solving nonconcave NUM problems and computationally demonstrate the advantages of our scheme.

1.3 Thesis Outline

The organization of the thesis is as follows.

In Chapter 2, we present the primal optimization problem and its dual counterpart. We introduce the dual subgradient method and primal-averaging scheme. Then we present a distributive algorithm in the specific framework of NUM, which serves as the model for our numerical experiments.

In Chapter 3, we study convergence properties of the averaged primal sequence in terms of feasibility violation and primal cost per iteration, and further explore its ability to reduce infeasibility by controlling initial parameters. We numerically investigate the performance of the scheme in NUM framework with concave utilities; moreover, we discuss the effects of system parameters (i.e. the value of constant stepsize and initial dual iterate) upon convergence properties of the system.

In Chapter 4, we extend the convergence analysis to the nonconvex case and provide modified bounds on the amount of primal cost. We implement and compare the averaging scheme with other models.

Finally, in Chapter 5, we summarize our work and discuss future developments.

Chapter 2

Model Formulation

In this chapter, we present the basic model and framework that are used for further development. In Section 2.1, we start by introducing the notations that will be used throughout the thesis. In Section 2.2, we formulate the dual subgradient scheme applied to the general optimization problem. In Section 2.3, we introduce the primal-averaging method. Based on the general formulation, we use the dual decomposition method to build a distributive model for the specific NUM application in Section 2.4. Assumptions for further theoretical analysis are also stated.

2.1 Notation and Preliminaries

For any vector $x \in \Re^n$, we use x_i to indicate its ith component, and we write $x = (x_1, \cdots, x_n)$. We view x as the column vector and its transpose x^T as a row vector with the same components. The inner product of the vectors x and y is defined as $x^T y = \sum_{i=1}^n x_i y_i$. In some situations, we may also use $x_k \in \mathbb{R}^n$ to denote a *vector* in the sequence $\{x_k\}$, according to the context.

Unless otherwise stated, we use $\|x\|$ to denote the L-2 norm, which is $\|x\| = \sqrt{x^T x}$. L-1 norm and L-∞ norm are also used occasionally, i.e. $\|x\|_1 = \sum_i |x_i|$ and $\|x\|_\infty = \max_{1 \le i \le n} |x_i|$, respectively. Note that $\|x\|_\infty \le \|x\| \le \|x\|_1$.

For any $\mu \in \mathbb{R}^m$, we define its projection on the nonnegative orthant $\mathbb{R}^m_+$ as $[\mu]^+$,

which is the component-wise maximum of vectors μ and 0, namely

$$[\mu]^+ = [\max(0, \mu_1), \cdots, \max(0, \mu_m)] .$$

In this case, we can see that $[\mu]^+ \geq \mu$.

The function of a single variable defined on the interval I is *strictly concave* if, for all $a, b \in I$ with $a \neq b$ and all $\lambda \in (0, 1)$, we have

$$f[(1-\lambda)a + \lambda b] > (1-\lambda)f(a) + \lambda f(b).$$

The function is *concave* if we also allow equality in the above expression. Strictly concave function has no linear parts and has a unique global maximum.

Given a concave function $q(\mu) : \mathbb{R}^m \rightarrow \mathbb{R}$, we denote a vector $d \in \mathbb{R}^m$ as the *subgradient* of q at the point $\bar{u} \in \mathbb{R}^m$ if

$$q(\bar{u}) + d^T(u - \bar{u}) \geq q(u) \qquad \forall \mu \in \mathbb{R}^m.$$

The set of all subgradients of q at $\bar{u} \in \mathbb{R}^m$ is called the *subdifferential* and is denoted by $\partial q(\bar{u})$.

2.2 General Model

2.2.1 Primal and Dual Formulation

Consider the following generic optimization problem, which is referred to as a *primal problem* (P):

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq 0 \\ & x \in X, \end{array} \tag{2.1}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the objective function, $g(x) = (g_1, g_2, \cdots, g_n)$ and each $g_j(x) : \mathbb{R}^n \rightarrow \mathbb{R}$, $j = 1, \cdots, m$ is one of the m inequality constraint functions.

$X \subset \mathbb{R}^n$ is a nonempty closed convex set. A point x in the domain of the problem is *feasible* if it satisfies the constraints $g(x) \leq 0$. The *primal optimal value* is denoted as f^*, which is assumed to be finite and is achieved at an optimal (feasible) solution x^*, i.e. $f^* = f(x^*)$.

If both the objective and constraint functions are convex, the problem is a *convex optimization problem.* Otherwise, the problem is *nonconvex.* In the following chapters, we assume that the convexity of $g(x)$ always holds, and (P) is convex (nonconvex) if $f(x)$ is convex (nonconvex).

To recover the primal optimal solution, we use the Lagrange duality theory to relax the primal problem of Equation (2.1) by its constraints, and the *Lagrangian function* is

$$L(x, \mu) = f(x) + \mu^T g(x).$$

The *dual function* is then defined as

$$\begin{aligned} q(\mu) &= \min_{x \in X} L(x, \mu) \\ &= \min_{x \in X} \{f(x) + \mu^T g(x)\}, \end{aligned} \tag{2.2}$$

where $\mu \in \mathbb{R}^m_+$ is the *Lagrange multiplier.* Equation (2.2) is concave regardless of the convexity of the primal problem, as it is a piecewise linear function of μ. Thus, we can maximize the dual function to obtain a tightest lower bound of the optimal primal optimal cost f^*. This is defined as a *dual problem*, which is always a convex optimization problem (D):

$$\begin{aligned} &\text{maximize} && q(\mu) \\ &\text{subject to} && \mu \in \mathbb{R}^m_+. \end{aligned} \tag{2.3}$$

The dual optimal solution is denoted as μ^*, at which the dual optimal value (also the tightest lower bound) is obtained, i.e. $q^* = q(\mu^*)$. This is in fact the *Weak Duality Theorem*, i.e. $q^* \leq f^*$. Note that when $q^* = f^*$, *duality gap* is zero.

2.2.2 Dual Subgradient Method

We denote the dual feasible set as $M = \{\mu \,|\, \mu \geq 0, -\infty < q(\mu) < \infty\}$, the dual optimal value as q^* and the dual optimal set as M^*. For every fixed $\mu \in M$, we have the solution set for the dual function of Equation (2.2):

$$X(\mu) = \arg\min_{x \in X}\{f(x) + \mu^T g(x)\}. \tag{2.4}$$

Assumption 2.1 *The convex set X is compact (i.e. closed and bounded).*

According to Weierstrass Theorem (see Bertsekas [2]), if Assumption 2.1 and the continuity of $f(x)$ and $g_j(x)$ hold, it follows that $X(\mu)$ is nonempty, i.e., there exists at least one optimal solution $x_\mu \in X(\mu)$. Note that the vector x_μ is unique if $f(x)$ is strictly convex, otherwise there may be multiple solutions.

The *subgradient method* is used to solve the dual problem. From the definition of q, we can see that $g(x_\mu)$ is readily obtained as a subgradient of q at μ. As such, a sequence of dual feasible points can be generated iteratively by

$$\mu_{k+1} = [\mu_k + \alpha g_k]^+ \qquad \text{for } k = 0, 1, \cdots, \tag{2.5}$$

where $\alpha > 0$ is a constant stepsize and $\mu_0 \geq 0$ is an initial dual iterate. The subgradient iterate g_k is

$$g_k = g(x_k), \tag{2.6}$$

where the corresponding primal feasible iterate is given by any solution of the set,

$$x_k = x_{\mu_k} \in \arg\min_{x \in X}\{f(x) + \mu_k^T g(x)\}. \tag{2.7}$$

As the subgradient method can usually generate a good estimation of the dual optimal solutions within a number of iterations, approximate primal solutions are obtained accordingly. The constant stepsize α is a simple parameter to control, and by choosing an appropriate value of α, the algorithm can approach the optimal value arbitrarily close within a finite number of steps.

For further development, we state the following assumptions which hold throughout the analysis.

Assumption 2.2 (Slater Condition) *There exists* $\bar{x} \in \mathbb{R}^n$, *such that*

$$g_j(\bar{x}) < 0 \qquad j = 1, \cdots, m.$$

A point satisfying Assumption 2.2 is called a *Slater point* or an inner point.

Assumption 2.3 (Bounded Subgradients) *There exists* $L > 0$, $L \in \mathbb{R}$, *such that*

$$\|g_k\| < L \qquad \forall k \geq 0.$$

Under Assumption 2.1, if we let $L = \max_{x \in X} \|g(x)\|$, Assumption 2.3 can be satisfied.

The dual subgradient schemes can be applied efficiently to the network optimization problems, due to their decomposable structures. In Section 2.4, we define this general model in a specific network setting, where the objective is to maximize utility functions [cf. *min* $f(x)$ of Equation (2.1)].

2.3 The Averaging Scheme

Although the dual subgradient method can generate near-optimal dual solutions with a sufficiently small stepsize and a large number of iterations (which is hard to be satisfied), it does not directly provide primal solutions which are of our interest. In some cases, it may even fail to produce any useful information (see reasons discussed in Section 1.1.2).

Thus, it motivates us to apply an averaging scheme to the primal sequence $\{x_k\}$ for approximate primal optimal solutions. In particular, for the sequence of primal vec-

tors $\{x_k\} \subset X$ generated in the dual subgradient scheme, we define another sequence $\{\hat{x}_k\}$ as the averages of the previous vectors $x_0, \cdots, x_{k-1}$, namely,

$$\hat{x}_k = \frac{1}{k} \sum_{i=0}^{k-1} x_i \in X \qquad \forall k \geq 1. \tag{2.8}$$

The average sequence $\{\hat{x}_k\}$ provides approximate primal solutions, whose properties are explored both theoretically and numerically throughout the thesis.

2.4 Network Utility Maximization Framework

Network Utility Maximization (NUM) problem has recently been used extensively to analyze and design distributive resource allocations in networks. We formulate the NUM model with the dual subgradient method, so as to set a platform to test the performance of the averaging scheme and verify our theoretical analysis.

Consider a network with L undirectional links of capacity $l \in L$ and S sources of transmitting rate x_s, where c_l and $s \in S$. We have

$$S(l) = \{s \in S \mid l \in L(s)\}, \text{ for each link } l$$

and

$$L(s) = \{l \in L \mid s \in S(l)\}, \text{ for each source } s,$$

where $S(l)$ denotes the set of sources that are on link l, and $L(s)$ denotes the set of links that source s uses. The connection of l and s can also be expressed compactly as an incident matrix $A \in \mathbb{R}^{L \times S}$, where each component is

$$A_{ls} = \begin{cases} 1, & \text{if } l \in L(s) \text{ or } s \in S(l) \\ 0, & \text{otherwise.} \end{cases}$$

The general problem of NUM is formulated as

$$\begin{array}{lll} \text{maximize} & \sum_s U_s(x_s) & \\ \text{subject to} & \sum_{s \in S(l)} x_s \leq c_l & \forall l = 1, \cdots, L \\ & 0 \leq x_s \leq M_s & \forall s = 1, \cdots, S, \end{array} \tag{2.9}$$

where $x \in \mathbb{R}^S, c \in \mathbb{R}^L$ and $M_s < \infty$ is the maximum transmission rate. The constraint set can be expressed as a matrix form of $Ax \leq c$. As x_s is bounded, $x \in X$ is a compact set (cf. Assumption 2.1), which also guarantees the existence of $X(\mu)$ in Equation (2.4). Here, if M_s is not specified in the problem, we can set it as

$$M_s = M = \max_{1 \leq l \leq L} \{c_l\}.$$

The following example demonstrates topologies which are used in our numerical experiments.

Example 2.1 (Network Topology)

Let $L = 2$, $S = 3$, we have a simple 2-link and 3-user network with link capacity $c = (1, 2)$, as shown in Figure 2-1.

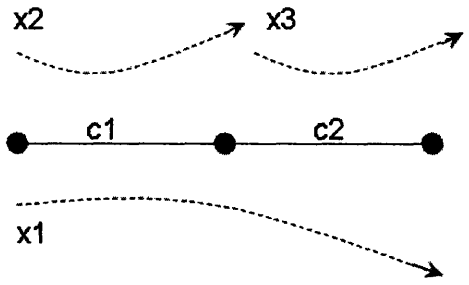


Figure 2-1: A simple line topology with 2 links and 3 users.

Besides the line topology, we consider also a tree network topology with 4 links and 4 users as in Figure 2-2.

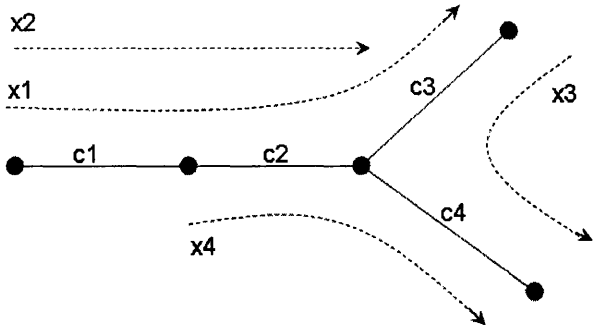


Figure 2-2: A tree topology with 4 links and 4 users.

The primal objective function in NUM is called the *utility function*, which represents the degree of satisfaction of the user when it acquires a certain amount of the resource (see Lee et al. [18]). We make some basic assumptions on the utility functions, though there are many other forms depending on the different network systems they model.

Assumption 2.4 (Utility Functions)

1. $U_s(0) = 0$.
2. *Each* $U_s(x_s)$ *is a monotonically increasing function.*
3. $U(x)$ *is the sum of* $U_s(x_s)$ *(additivity and separability), where* U_s *may or may not be the same for each of the* S *sources.*

Therefore, the primal problem is to maximize the total utility over a compact set, subject to the coupling linear constraints [cf. (P) where we minimize $f(x)$]. If the utility functions are concave, it is a convex optimization problem with zero duality gap. In a more realistic setting, the utility functions can be nonconcave and then the NUM problem becomes nonconvex. Note that in NUM, we consider convex (nonconvex) optimization and maximization of concave (nonconcave) utility functions interchangeably. We will discuss both the convex and nonconvex cases in details in

the numerical experiments.

The linear constraints in Equation (2.9) can be decoupled as

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

where $\mu \in \mathbb{R}^L$ is the multiplier or link price in the network. We denote the aggregate link price of source s as $\mu^s = \sum_{l \in L(s)} \mu_l$. The Lagrange dual function is then

$$\begin{aligned} q(\mu) &= \max_x \; L(x,\mu) \\ &= \sum_s [\max_{x_s} \; L_s(x_s, \mu^s)] + \sum_l c_l \mu_l, \end{aligned}$$

which is equivalent to solving a subproblem at each source s distributively:

$$q_s(\mu) = \max_{0 \le x_s \le M_s} \{U_s(x_s) - \mu^s x_s\}. \tag{2.10}$$

At the higher level, the master dual problem is

$$\begin{aligned} \text{maximize} \quad & q(\mu) \\ \text{subject to} \quad & \mu \ge 0 \\ \text{where} \quad & q(\mu) = \textstyle\sum_s q_s(\mu) + c^T \mu. \end{aligned}$$

The dual variable μ is updated iteratively at each iteration k by the subgradient method introduced in Equation (2.5), except that it can be solved locally at each link l, i.e.

$$\mu_l(k+1) = [\mu_l(k) + \alpha\{x^l[\mu(k)] - c_l\}]^+ \qquad \forall l, \quad \forall k = 0, 1, \cdots,$$

where $x^l = \sum_{s \in S(l)} x_s$ is the aggregate rate at link l. At each iteration k, the primal iterate $x_s(k) = x_s[\mu^s(k)]$ can be obtained parallelly at all sources by

$$x_s[\mu^s(k)] = \arg \max_{0 \le x_s \le M_s} [U_s(x_s) - \mu^s(k) x_s] \qquad \forall s,$$

which is the solution of the subproblem in Equation (2.10). Collectively, we have $x(k) = (x_1, \cdots, x_S)$ and $\mu(k) = (\mu_1, \cdots, \mu_L)$ at iteration k.

The NUM problem in Equation (2.9) is a special case of the general optimization problem in Equation (2.1). To avoid confusion, we convert the NUM problem into the more general form of a minimization problem, by setting

$$\begin{aligned} \max_{0 \le x_s \le M_s} \sum_s U_s(x_s) &= -\min_{0 \le x_s \le M_s} \sum_s [-U_s(x_s)] \\ &= -\min_{x_s} \sum_s f_s(x_s). \end{aligned} \tag{2.11}$$

Now we deal with a convex function $f(x) = \sum_s f_s(x_s)$. The (linear) constraints $g(x)$ remain unchanged. As maximizing a concave function is equivalent to minimizing a convex function, we have the same iterate pair $\{x_k, \mu_k\}$ (so is $\{\hat{x}_k\}$) in the two formulations, and the objective values are just negative of each other. We use this converted form in the numerical experiments.

Chapter 3

Approximate Primal Sequence in the Convex Case

In the previous chapter, we formulated the basic primal and dual problems of interest, and introduced the averaging scheme in dual subgradient methods as a means to recover near-feasible and near-optimal primal solutions.

Based on the model, we start with the convex optimization problems in this chapter. In Section 3.1, we study the convergence properties of the averaged primal solutions. We further investigate its ability to reduce the amount of feasibility violation by varying the value of system parameter. In Section 3.2, we conduct numerical experiments in the NUM framework to test the theoretical results. We explicitly compare the performance of the proposed averaging scheme with ordinary dual subgradient scheme, in terms of feasibility and optimality. From the theoretical insights and computational results, we also show the effects of system parameters, namely, initial dual iterate and stepsize on the quality of approximate primal solutions.

3.1 Convergence Analysis

In the following analysis, we let Assumptions (1) to (4) defined in Chapter 2 hold and focus on the cases in which both $f(x)$ and $g(x)$ are convex.

3.1.1 Properties of Approximate Primal Sequence

Intuitively, the averaged primal solution $\hat{x}_k$ generated in Equation (2.8) is expected to have some convergence properties the original sequence lacks. It turns out there are many far-reaching benefits introduced by the averaging scheme.

The following proposition by Nedić and Ozdaglar [22] provides bounds on the feasibility violation, as well as the primal cost of the running averages, which is the basis for our further investigation. Note that, the bounds are given per iteration, which can provide a practical stopping criteria for the iterations. We include the proof for completeness.

Proposition 3.1 (Nedić and Ozdaglar [22]) *Consider the convex problems defined in the form of (P). Let the dual sequence $\{\mu_k\}$ be generated by the subgradient method of Equation (2.5). Let the vectors $\hat{x}_k$ for $k \geq 1$ be the averages in Equation (2.7). Under the assumptions that $f(x)$ and $g(x)$ are convex and X is a convex compact set, we have the following bounds, for all $k \geq 1$:*

(a) An upper bound on the amount of constraint violation of the vector $\hat{x}_k$,

$$\|g(\hat{x}_k)^+\| \leq \frac{\|\mu_k\|}{k\alpha}. \tag{3.1}$$

(b) An upper bound on the primal cost of $\hat{x}_k$,

$$f(\hat{x}_k) \leq f^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha}{2k}\sum_{i=0}^{k-1}\|g(x_i)\|^2. \tag{3.2}$$

If we have $\|g_k\| < L$, $\forall k \geq 0$ (Assumption 2.3), we can further have

$$f(\hat{x}_k) \leq f^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha L^2}{2}. \tag{3.3}$$

(c) and a lower bound on the primal cost of $\hat{x}_k$,

$$f(\hat{x}_k) \geq f^* - \|\mu^*\|\|g(\hat{x}_k)^+\|. \tag{3.4}$$

where μ^* is a dual optimal solution, and f^* is the primal optimal value.

Proof. (a) (Feasibility) According to the dual iteration in Equation (2.5) of the subgradient method, for all $k \geq 0$, we have

$$\begin{aligned} \mu_{k+1} &= [\mu_k + \alpha g_k]^+ \\ &\geq \mu_k + \alpha g_k. \end{aligned} \tag{3.5}$$

it follows that

$$\begin{aligned} \alpha g(x_k) &= \alpha g_k \\ &\geq \mu_{k+1} - \mu_k. \end{aligned}$$

We have $x_k \in X$ for all k according to Equation (2.7). It follows that $\hat{x}_k \in X$ since the set X is convex and averaging is a convex operation.

As the constraint function $g(x)$ is convex, we have

$$\begin{aligned} g(\hat{x}_k) &\leq \frac{1}{k}\sum_{i=0}^{k-1} g(x_i) \\ &= \frac{1}{k\alpha}\sum_{i=0}^{k-1} \alpha g(x_i) \\ &\leq \frac{1}{k\alpha}(\mu_k - \mu_0) \\ &\leq \frac{\mu_k}{k\alpha} \qquad \forall k \geq 1. \end{aligned}$$

Since all the μ_k is nonnegative, we have $g(\hat{x}_k)^+ \leq \frac{\mu_k}{k\alpha}$. Thus the constraint violation is bounded:

$$\|g(\hat{x}_k)^+\| \leq \frac{\|\mu_k\|}{k\alpha} \qquad \forall k \geq 1.$$

(b) (Optimality - upper bound) By the definitions of dual function in Equation (2.5)

and primal iterate in Equation (2.7), we have

$$f(x_i) + \mu_i^T g(x_i) = q(\mu_i) \leq q^*,$$

where q^* is a dual optimal solution. Under the assumption of convexity and Slater condition, we have $q^* = f^*$ (see Bertsekas [2]).

Together with the condition that $f(x)$ is convex and $x_k \in X$, we can obtain

$$\begin{aligned} f(\hat{x}_k) &\leq \frac{1}{k}\sum_{i=0}^{k-1} f(x_i) \\ &= \frac{1}{k}\sum_{i=0}^{k-1}\{f(x_i) + \mu_i^T g(x_i)\} - \frac{1}{k}\sum_{i=0}^{k-1}\mu_i^T g(x_i) \\ &= \frac{1}{k}\sum_{i=0}^{k-1} q(\mu_i) - \frac{1}{k}\sum_{i=0}^{k-1}\mu_i^T g(x_i) \\ &\leq q^* - \frac{1}{k}\sum_{i=0}^{k-1}\mu_i^T g(x_i). \end{aligned}$$

By applying the nonexpansiveness property of projection to the dual iterate, we have the following relation for every $i \geq 1$,

$$\begin{aligned} \|\mu_{i+1}\|^2 &= \|[\mu_i + \alpha g_i]^+ - [0]^+\|^2 \\ &\leq \|\mu_i + \alpha g_i\|^2 \\ &\leq \|\mu_i\|^2 + 2\alpha\mu_i^T g(x_i) + \alpha^2\|g(x_i)\|^2. \end{aligned}$$

Thus, by plugging in $-\mu_i^T g(x_i)$ and summing over $0, \cdots, k-1$, we can obtain an upper bound for the primal cost:

$$\begin{aligned} f(\hat{x}_k) &\leq q^* + \frac{1}{k}\sum_{i=0}^{k-1}\frac{\|\mu_i\|^2 - \|\mu_{i+1}\|^2 + \alpha^2\|g(x_k)\|^2}{2\alpha} \\ &= q^* + \frac{\|\mu_0\|^2 - \|\mu_k\|^2}{2k\alpha} + \frac{\alpha}{2k}\sum_{i=0}^{k-1}\|g(x_i)\|^2 \\ &= f^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha L^2}{2}. \end{aligned}$$

(c) (Optimality - lower bound) By the definition of the primal and dual optimal values, for any $x \in X$, we have

$$\begin{aligned} f(x) + (\mu^*)^T g(x) &\geq f(x^*) + (\mu^*)^T g(x^*) \\ &= q(\mu^*). \end{aligned}$$

As $\hat{x}_k \in X$ for all $k \geq 1$, we can use the above relation to get a lower bound for each iterate $\hat{x}_k$:

$$\begin{aligned} f(\hat{x}_k) &= f(\hat{x}_k) + (\mu^*)^T g(\hat{x}_k) - (\mu^*)^T g(\hat{x}_k) \\ &\geq q(\mu^*) - (\mu^*)^T g(\hat{x}_k) \\ &\geq q(\mu^*) - (\mu^*)^T g(\hat{x}_k)^+ \\ &\geq f^* - \|\mu^*\| \|g(\hat{x}_k)^+\|. \end{aligned}$$

■

As the duality gap is zero in the convex optimization problems, we can interchange q^* with f^* in the above results.

These bounds on the feasibility and optimality measurements of the approximate primal solution $\hat{x}_k$ can be easily obtained if we have bounds on the norms of dual iterate $\|\mu_k\|$, dual optimal $\|\mu^*\|$ and subgradient $\|g(x_k)\|$, which are available from the simulations. Thus, these bounds can provide us with practical information to estimate optimal solutions, which the original primal sequence without averaging lacks.

Finally, we note that the upper bound for the constraint violation $\|g(\hat{x}_k)^+\|$ suggests that the infeasibility of the averaged primal sequence is always bounded and it can be reduced in the order of $1/k$ as the iteration continues. This observation implies that, by increasing the number of iterations, we can reduce the amount of feasibility violation to a certain level accepted by the system. As congestion control is one of the most important requirements for rate allocation in network applications, the ability of the averaging scheme to reduce infeasibility is of practical significance.

In fact, further investigations reveal a more powerful means to directly control the amount of feasibility violation of the approximate primal vectors, i.e., the infeasibility can be reduced by increasing the initial dual iterate μ_0, which is discussed next.

3.1.2 Infeasibility Reduction

One of the drawbacks of the dual subgradient scheme is that it usually yields infeasible primal solutions (see reviews in Section 1.1.2), and we try to remedy this. We show below that the averaging scheme can effectively suppress and control the feasibility violation in the initialization of system.

Proposition 3.2 *Let $\bar{x}$ be a primal vector satisfying the Slater condition in Assumption 2.1. Let L be the bound of subgradient as in Assumption 3. For any stepsize $\alpha > 0$, if we set the initial dual iterate μ_0 sufficiently large, in particular,*

$$(\mu_0)_j \geq \frac{f(\bar{x}) - q^* + \alpha L^2/2}{\min_{1\leq j\leq m}\{-g_j(\bar{x})\}} \qquad \forall j,$$

we can eliminate the infeasibility of approximate primal solutions $\hat{x}_k$ from the beginning of iterations, i.e. $\|g(\hat{x}_k)^+\| = 0$ for all $k \geq 1$.

Proof. By the nonexpansiveness of projection and the definition of dual iteration in Equation (2.5), for all $k \geq 1$, we have

$$\begin{aligned}\|\mu_{k+1} - \mu^*\|^2 &= \|[\mu_k + \alpha g_k]^+ - [\mu^*]\|^2 \\ &\leq \|\mu_k + \alpha g_k - \mu^*\|^2 \\ &= \|\mu_k - \mu^*\|^2 + 2\alpha g_k^T(\mu_k - \mu^*) + \alpha^2\|g_k\|^2.\end{aligned}$$

Using the definition of subgradient, we have

$$g_k^T(\mu_k - \mu^*) \leq q(\mu_k) - q^*.$$

By the boundedness of subgradient, we can obtain

$$\|\mu_{k+1} - \mu^*\|^2 \leq \|\mu_k - \mu^*\|^2 - 2\alpha(q^* - q(\mu_k)) + \alpha^2 L^2. \tag{3.6}$$

Now we consider the following two cases for any constant stepsize α,

Case 1:

$$0 < \alpha < \frac{2(q^* - q(\mu_k))}{L^2}.$$

Under this stepsize (sufficiently small), we can see that the sum of the last two terms in Equation (3.6) is negative, and it follows that

$$\|\mu_{k+1} - \mu^*\| < \|\mu_k - \mu^*\|. \tag{3.7}$$

It implies that, if α is sufficiently small, the distance of dual iterates to the optimal value μ^* is reduced at each iteration. Thus, μ_k can eventually converge to the level set arbitrarily which is close to q^*, i.e. $\{\mu \geq 0 \,|\, q(\mu) < q^* - \frac{\alpha L^2}{2}\}$.

Case 2:

$$\alpha \geq \frac{2(q^* - q(\mu_k))}{L^2}.$$

This is the case when α is relatively large and μ_k may typically experience an oscillatory behavior without converging. Equivalently, we can express the expression of α above in terms of a level set, i.e.

$$\mu_k \in Q_\alpha = \{\mu \geq 0 \,|\, q(\mu) \geq q^* - \frac{\alpha L^2}{2}\} \qquad \forall k.$$

According to the definition of dual function $q(\mu)$, for any nonnegative vector μ in this level set, we have

$$\begin{aligned} q^* - \frac{\alpha L^2}{2} &\leq q(\mu) \\ &= \inf_{x \in X}\{f(x) + \mu^T g(x)\} \\ &= f(x_\mu) + \mu^T g(x_\mu) \\ &\leq f(\bar{x}) + \mu^T g(\bar{x}). \end{aligned}$$

It follows that

$$\begin{aligned} -\mu^T g(\bar{x}) &= -\sum_{j=1}^{m} \mu_j g_j(\bar{x}) \\ &\leq f(\bar{x}) - q^* + \frac{\alpha L^2}{2}. \end{aligned}$$

As $g_j(\bar{x}) < 0$ for all j, we have

$$\begin{aligned} \min_{1\le j\le m}\{-g_j(\bar{x})\}\left(\sum_{j=1}^{m}\mu_j\right) &\le -\mu^T g(\bar{x}) \\ &\le f(\bar{x}) - q^* + \frac{\alpha L^2}{2}. \end{aligned}$$

Recall that $\|\mu\|_\infty = \max_{1\le j\le m}\mu_j$, $\|\mu\|_1 = \sum_{j=1}^{m}\mu_j$ and $\|\mu\|_\infty \le \|\mu\|_1$, we can have

$$\max_{1\le j\le m}\mu_j \le \frac{f(\bar{x}) - q^* + \alpha L^2/2}{\min_{1\le j\le m}\{-g_j(\bar{x})\}}. \tag{3.8}$$

Since we have each dual iterate $\mu_k \in Q_\alpha$ for all k, every component of the vector μ_k can satisfy the upper bound of Equation (3.8).

Thus, for any given stepsize α, if we set each component of the initial dual sufficiently large, in particular,

$$(\mu_0)_j \ge \frac{f(\bar{x}) - q^* + \alpha L^2/2}{\min_{1\le j\le m}\{-g_j(\bar{x})\}} \qquad \forall j,$$

then we have $(\mu_0)_j \ge \max_{1\le i\le m}(\mu_k)_i$ for each component of the dual vector, and it follows that

$$\mu_0 \ge \mu_k \qquad \forall k = 0, 1, \cdots$$

Now, if the constraint functions are convex, as assumed throughout the thesis, we have

$$\begin{aligned} g(\hat{x}_k)^+ &\le \frac{1}{k}\sum_{i=0}^{k-1} g(x_i) \\ &= \frac{1}{k\alpha}\sum_{i=0}^{k-1}\alpha g(x_i) \\ &\le \frac{1}{k\alpha}\sum_{i=0}^{k-1}(\mu_{i+1} - \mu_i) \\ &= \frac{1}{k\alpha}(\mu_k - \mu_0) \\ &\le 0. \end{aligned}$$

Therefore, we can obtain $\|g(\hat{x}_k)^+\| = 0$ for all the iterations k.

■

Further Observation and Discussion

The above result suggests that, at a given stepsize α, if we set μ_0 sufficiently large at the initialization, the infeasibility of $\{\hat{x}_k\}$ in the averaging scheme can be reduced to zero even from the beginning of iterations (with the possible loss of accuracy for the approximate primal cost). It gives us the idea of reducing feasibility violation of the averaged primal solutions by increasing the parameter μ_0, which is a nice property that the original primal sequence $\{x_k\}$ does not posses. In effect, in the case that $\|g(\hat{x}_k)^+\| = 0$ for all $k \geq 0$, we can further deduce the following properties of $\{\hat{x}_k\}$:

1. $\{f(\hat{x}_k)\}$ converges downward from above, i.e., there exists a $k' > 0$, such that $f(\hat{x}_{k+k'}) < f(\hat{x}_k)$ for all k.

2. The upper bound of $f(\hat{x}_k)$ becomes looser as μ_0 increases.

3. The lower bound becomes $f(\hat{x}_k) \geq q^*$ if $\|g(\hat{x}_k)^+\| = 0$ for all $k \geq 0$.

4. Each component of $\{\hat{x}_k\}$ (cf. data flow in the NUM framework) converges upward from below.

An intuitive explanation can be, when $\mu_0 > \mu_k$ for all k (it is practically unlikely that $\mu_0 = \mu_k$ for all k), $x(\mu)$ is a decreasing function throughout the domain (see Lee, Mazumdar and Shroff [18], and simulation results in Section 4.2.3). It follows that $f(x_0) > f(x_k)$ if $f(x)$ is monotonically decreasing (this is often the case in network problems). Therefore, no matter how oscillatory the original primal iterates are, the running averages of the previous iterates and the primal cost tend to have certain trends.

In contrast, by reducing μ_0, we can expect different trends of the sequences generated in the averaging scheme.

In the NUM problem, if we set $\mu_0 = 0$, we can observe the following properties of the averaged sequences generated in the simulation:

1. The feasibility violation is bounded by $\|g(\hat{x}_k)^+\| \approx \frac{\|\mu_k\|}{k\alpha}$, and is reduced at the rate of 1/k approximately.

2. The upper bound of primal cost becomes $f(\hat{x}_k) \leq q^* + \frac{\alpha}{2k}\sum_{i=0}^{k-1}\|g(x_i)\|^2$.

3. $\{f(\hat{x}_k)\}$ converges upward from below, i.e., there exists $k' > 0$, such that $f(\hat{x}_{k+k'}) > f(\hat{x}_k)$ for all k.

Similarly, we can make some intuitive explanations for the above claims. If each constraint function $g_j(x)$ is affine (or linear) in x, which is usually the case in network flow problems, we can have

$$\begin{aligned} g(\hat{x}_k) &= g(\frac{1}{k}\sum_{i=0}^{k-1} x_i) \\ &= \frac{1}{k}\sum_{i=0}^{k-1} g(x_i). \end{aligned}$$

Since $\mu_k + \alpha g_k \leq [\mu_k + \alpha g_k]^+ = \mu_{k+1}$, it follows that

$$\begin{aligned} g(\hat{x}_k) &= \frac{1}{k\alpha}\sum_{i=0}^{k-1} \alpha g(x_i) \\ &\leq \frac{1}{k\alpha}[(\mu_k - \mu_{k-1}) + (\mu_{k-1} - \mu_{k-2}) + \cdots + (\mu_1 - \mu_0)] \\ &= \frac{\mu_k}{k\alpha}. \end{aligned}$$

As $g(\hat{x}_k)^+$ essentially maintains the nonnegative components while increasing its negative components to zero, and the relation $\mu_k + \alpha g_k = \mu_{k+1}$ holds only for the nonnegative components of g_k, we can have $g(\hat{x}_k)^+ \approx \frac{\mu_k}{k\alpha} \geq 0$ if most components of $g(\hat{x}_k)$ are nonnegative.

If $\mu_0 = 0$, it follows that $\mu_k > \mu_0$ for all $k \geq 1$, since the nontrivial dual vector μ_k must have some positive components for $k \geq 1$, i.e. $(\mu_k)_j > 0$ for some j. Thus, we

have $\hat{x}_{\max} = \hat{x}(\mu_0)$, where $x(\mu)$ is a nonincreasing function. As every $g_j(x)$ is linear, it implies that $\|g(\hat{x}_{\max})^+\|$ has the largest amount of infeasibility. Thus, under dual subgradient methods, $\|g(\hat{x}_k)^+\|$ decreases as the iterations carry on. It is clear that the upper bound is tightest when we set $\mu_0 = 0$.

The explanation for the trend of $f(\hat{x}_k)$ is similar to the previous case. Note that as the upper bound $f(\hat{x}_k) \leq q^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha}{2k}\sum_{i=0}^{k-1}\|g(x_i)\|^2$ can suggest, the value of $\mu_0 = 0$ also affects the estimate of primal cost. Therefore, there is always a trade-off at the selection of μ_0.

These trends of different sequences are to be illustrated in the plots of the numerical experiments.

3.2 Applications to Concave NUM

To test the theoretical results and demonstrate the behaviors of the averaging scheme, we investigate its performance in the framework of Network Utility Maximization (NUM). With the distributive model built specifically for NUM in Chapter 2.4, we can test the scheme with instances of different concave utility functions and topologies. We also compare it with the ordinary dual subgradient system without averaging.

3.2.1 Concave Utility Functions

In NUM, the network flow is *elastic* if the sources can adjust their transmission data rate gradually in response to congestion levels within the network. It can tolerate packet delays and losses gracefully (see Shenker in [25]).

We can model this class of services with concave utility functions, which turns out to be convex optimization problems. Moreover, the application of Lagrangian dual method to the network flow problems with separable utility functions gives rise to highly parallelizable algorithms. Thus, we can focus on individual utility functions (which are scalar functions) and combine them for homogenous or heterogenous

systems. We illustrate different concave utilities below.

Example 3.1 We plot the functions that are typical concave utilities in NUM to model elastic flows, as shown in Figure 3-1. Note that functions **(F1)** to **(F3)** are strictly convex.

(F1) $U(x) = 1 - e^{-kx}, \quad k > 0.$

(F2) $U(x) = x^k, \quad 0 < k < 1.$

(F3) $U(x) = log(kx + 1), \quad k > 0.$

(F4) $U(x) = kx, \quad k > 0.$

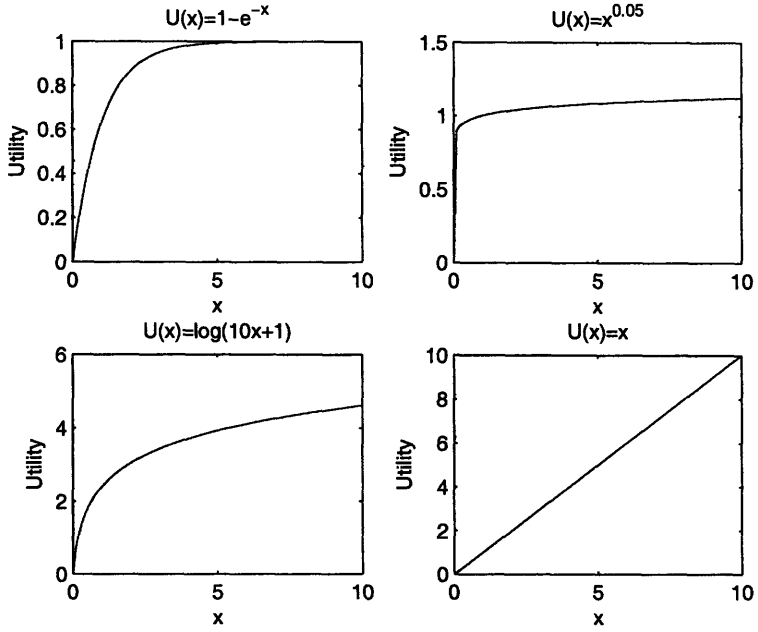


Figure 3-1: Examples of concave utility functions.

3.2.2 Models and Algorithms

We implement and compare two (distributive) schemes: a basic dual subgradient scheme (abbreviated as DS) and a modified scheme with the averaging techniques

(abbreviated as DSA). Note that DSA is essentially a dual subgradient scheme with primal-averaging method and we summarize its algorithm below.

Algorithm for DSA

1. *Initialize the system by setting a nonnegative* μ_0 *and an appropriate stepsize* α.

2. *For each user* s, *compute the primal iterate at iteration* k:
   $x_s(\mu_k) = \arg\min_{x \in X}\{f(x) + \mu_k^T g(x)\}$, *where* $\mu_k = \mu(k)$.

3. *For each link* l, *update the dual iterate at iteration* k:
   $\mu_l(k+1) = \max[0, \mu_l(k) + \alpha x^l(\mu_k) - c_l]$.

4. *Repeat steps 2 to 3 if* N *iterations have not been completed.*

5. *Otherwise, compute* $\hat{x}_k$ *using Equation (2.8).*

6. *Evaluate the performance parameters [i.e.* $g(x_k)$, $f(x_k)$, $g(\hat{x}_k)$, $f(\hat{x}_k)$*], at the master level.*

To conform to the theoretical analysis, we focus on the numerical measurements per iteration instead of the asymptotical convergence. Thus the maximum number of iterations is preset to N in the experiment. Essentially, when the utility function of each user is concave, the aggregate objective function is concave.

3.2.3 Numerical Results

In this section, we provide computational results to demonstrate the performance of the DSA systems. In the experiment, we first convert the original NUM to a convex minimization problem as in Equation (2.11). Thus the figures are generated for the convex version, i.e. $f(x) = -U(x)$, while the data in tables is converted back to the original utility maximization problem. Typically, we run the simulations in a fairly small number of iterations, e.g. $N = 40$, and investigate properties of the averaged

primal sequence per iteration. The following parts discuss different aspects of the scheme.

Varying α – Performance Comparison of DS vs. DSA

Consider the 3-user, 2-link line topology in Figure 2-1, with link capacity $c = (1, 2)$. In a *homogeneous system* system, we assume that each user is modeled with the same utility function (F2). This optimization problem is formulated as:

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{s=1}^{3} U_s(x_s) = \sum_{s=1}^{3} x_s^{0.05} \\ \text{subject to} \quad & x_1 + x_2 \leq 1 \\ & x_1 + x_3 \leq 2 \\ & 0 \leq x_1, x_2, x_3 \leq 2. \end{aligned}$$

we can also express the constraints in the matric form as

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} x \leq \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Due to the small number of flows and links, we use an exhaustive search to obtain the global optimum which can be used to check against the simulated solution: $U^* = 2.9538$, at $x^* = (0.4145, 0.5855, 1.5855)$.

To compare the performance of the two schemes with and without averaging, we first set the initial dual iterate $\mu_0 = 0$ and vary the stepsize from 0.03 to 2. At each test instance, we evaluate the amount of constraint violation and value of primal cost in the two schemes, as recorded in Table 3.1. The quantity a/b denotes the two extreme values between which the parameter oscillates, when the system becomes unstable as k approaches N. A single quantity (i.e. value of the last iterate at N) in the table implies that the corresponding sequence converges smoothly. We also compare the estimated cost value in each scheme with the optimal utility (in the last two columns).

Figures 3-2 to 3-5 illustrate the convergence properties of difference sequences

α	$\|g^+\|$	$\|\hat{g}^+\|$	U	$\hat{U}$	U/U^*	$\hat{U}/U^*$
0.03	0	0.0743	2.9538	2.9604	1.0000	1.0022
0.04	0/0.4511	0.0550	2.9336/2.9703	2.9589	0.9932/1.0056	1.0017
0.05	0/0.4750	0.0435	2.9310/2.9724	2.9579	0.9923/1.0063	1.0014
0.06	0/0.5065	0.0358	2.9275/2.9750	2.9573	0.9911/1.0072	1.0012
0.07	0/0.5503	0.0305	2.9227/2.9783	2.9568	0.9895/1.0083	1.0010
0.08	0/0.6109	0.0266	2.9151/2.9825	2.9561	0.9869/1.0097	1.0008
0.09	0/0.6394	0.0230	2.9026/2.9855	2.9536	0.9827/1.0107	0.9999
0.1	0/0.6730	0.0200	2.8898/2.9888	2.9516	0.9783/1.0119	0.9993
0.2	0/1.1742	0.0072	2.6905/3.0231	2.9460	0.9109/1.0235	0.9974
1	0.04/0	0	2.4145/3.1058	2.9453	0.8174/1.0515	0.9971
2	0.0199/0	0	2.3278/3.1058	2.9446	0.7881/1.0515	0.9969

Table 3.1: Performance comparisons of DS and DSA schemes towards $k = N$, with $\mu_0 = 0$ and α varying from 0.03 to 2.

in the two schemes. From the simulation plots, we can seen that the DS scheme converges smoothly only at a sufficiently small α (e.g. 0.03) and sufficiently large N. A little perturbation of stepsize (even as close as $\alpha = 0.04$) results in oscillations of $\{x_k\}$. The estimated primal cost $f(x_k)$ (i.e. U in the table)and constraint infeasibility $\|g(x_k)^+\|$ (i.e. $\|g^+\|$ in the table) have the same behaviors. As the stepsize becomes larger, the fluctuation is more severe and it is harder to retrieve useful information from the simulation results. Figure 3-4 shows that the aggregate data rates of users in DS is prone to exceed the capacity of the link (it oscillates between being feasible and infeasible), and there is no tendency of the scheme to decrease the amount of congestion when iterations continues.

In contrast, the averaged primal sequence $\{\hat{x}_k\}$ in DSA is robust to the change of stepsize. As α increases, $f(\hat{x}_k)$ (corresponding to $\hat{U}$ in the table) still converges within an error level of f^* [cf. Equation (3.2)]. Even at a large stepsize $\alpha = 2$, the estimated $\hat{U}$ is within 0.31% error of the optimal value. Also, the infeasibility $\|g(\hat{x}_k)^+\|$ (i.e. $\|\hat{g}^+\|$ in the table) in DSA can always diminish to zero as k increases to infinity. Note that when we set $\mu_0 = 0$, the constraint violation is closest to the upper bound in Proposition 3.1(a).

Varying μ_0 – Infeasibility Reduction in DSA

In this part, we fix the stepsize at $\alpha = 0.03$ and vary μ_0, in order to test its effect on the infeasibility of $\hat{x}_k$. Note that $(\mu_0)_l$ denotes the value of each component of the initial dual vector.

Numerical results in Table 3.2 show the reduction in the infeasibility of averaged primal sequence in DSA when μ_0 is increased. In fact, the congestion can be eliminated at each averaged primal iterate $\hat{x}_k$ provided that μ_0 is large enough, as shown in Proposition 3.2. As q^* is not readily available before the simulation, we can start with $\mu = 0$ and slowly increase it to reduce the amount of constraint violation.

Figures 3-6 to 3-7 illustrate the trends of convergence for different sequences in DSA, which agree with our discussion in Section 3.1.2.

Other testing instances (e.g. at a larger α) also show that, unlike DSA, increasing μ_0 in DS cannot improve the feasibility of the ordinary primal sequence $\{x_k\}$. Thus, by adjusting the value of initial dual vector, we can have control over the amount of congestion in the DSA scheme, which is another advantage the averaging technique introduces to the recovery scheme for primal optimal solutions.

$(\mu_0)_l$	$\|\hat{g}^+\|$	$\hat{U}$	$\hat{U}/U^*$
0	0.0743	2.9604	1.0022
0.02	0.0536	2.9585	1.0016
0.04	0.0360	2.9566	1.0010
0.06	0.0193	2.9547	1.0003
0.08	0.0026	2.9527	0.9996
0.1	0	2.9508	0.9993

Table 3.2: Performance measurement of DSA scheme towards $k = N$, with $\alpha = 0.03$ and μ_0 varying from 0 to 0.1. Note that there is no relation between μ_0 and convergence of x_k in DS.

Simulation Results versus Theoretical Bounds

In this part, we check how the theoretical bounds in Proposition 3.1 predict the performance of DSA with different parameters.

Figure 3-8 compares the simulation results of the amount of feasibility violation against its theoretical bound, at different combinations of μ_0 and α. It shows that if $\mu_0 = 0$, the two curves almost coincide and infeasibility decreases approximately at the rate of $1/k$. The increase of μ_0 can accelerate the decrease of feasibility violation. In all of the test instances, the amount of infeasibility is well bounded by the trendline with a rate of $1/k$.

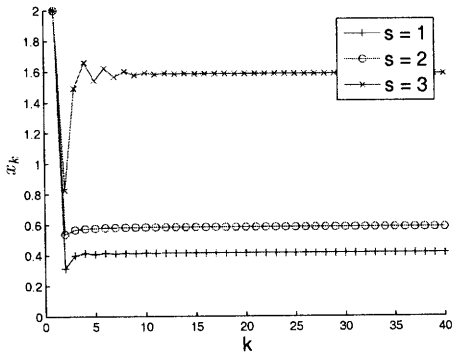
Figure 3-9 compares the experimental value of the primal cost per iteration with its theoretical bounds. The lower bound is always tight if we use μ^* in the calculation; in practice, it can be replaced with bounds on μ_k (see Nedić and Ozdaglar [22]). If μ_0 is not zero, the upper bound of $f(\hat{x}_k)$ is looser. Thus, there is always a tradeoff between reducing infeasibility $\|g(\hat{x}_k)^+\|$ and estimating the optimal value $f(\hat{x}_k)$ at the choice of μ_0. We can also view tradeoffs between accuracy of approximate solutions and number of iterations at the selection of stepsize.

From the observations in the numerical experiments, we can see that the averaging scheme can considerably improve the performance of approximate primal solutions. Its robustness to a large range of stepsizes, as well as its ability to control the amount of infeasibility (cf. congestion in NUM), are advantageous in the practical applications.

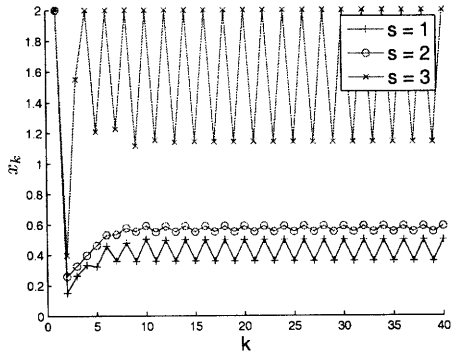
3.3 Summary

In this chapter, we have studied the convergence properties of approximate primal solutions, which are generated in the dual subgradient with averaging scheme (DSA). We have further discovered its effectiveness in reducing infeasibility of the system, by controlling the value of initial dual vector. Finally, we have seen from the numerical experiments conducted in the concave NUM framework, that the DSA scheme performs better than the DS approach in most scenarios.

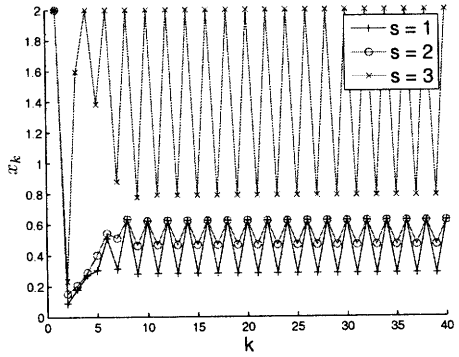
Therefore, we can conclude that the DSA scheme can provide a consistently good estimate of the primal solution for the convex optimization problems.



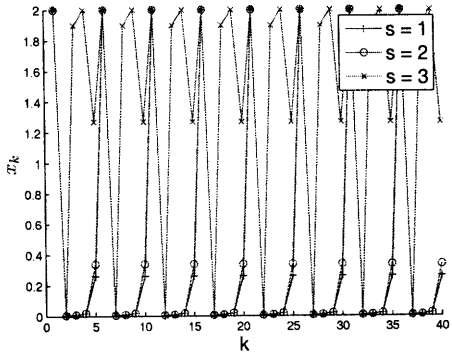
(a) $\alpha = 0.03$, $\mu_0 = 0$.



(b) $\alpha = 0.06$, $\mu_0 = 0$.

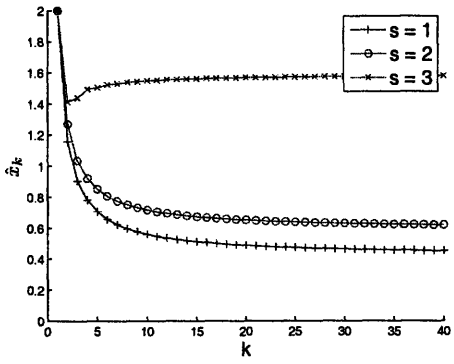


(c) $\alpha = 0.1$, $\mu_0 = 0$.

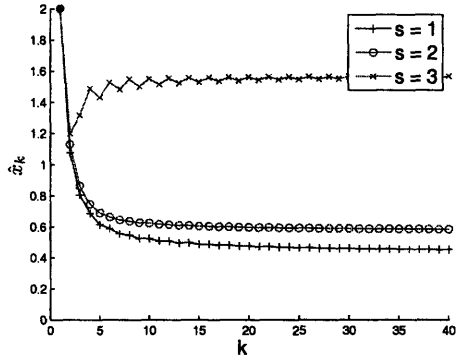


(d) $\alpha = 2$, $\mu_0 = 0$.

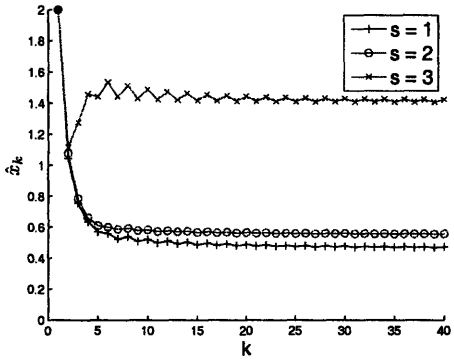
Figure 3-2: Three data flows $\{x_k\}$ of DS scheme without averaging, with α increasing from 0.03 to 2 and $\mu_0 = 0$. Oscillation is present when stepsize is slightly larger than 0.03 and becomes more severe as it grows.



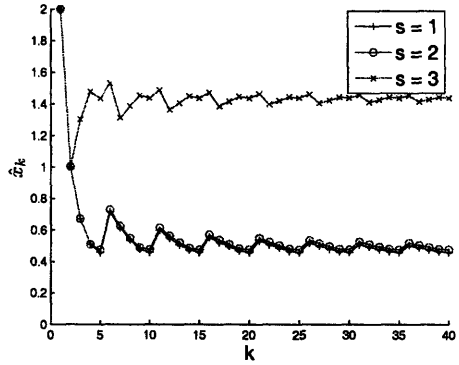
(a) $\alpha = 0.03$, $\mu_0 = 0$.



(b) $\alpha = 0.06$, $\mu_0 = 0$.

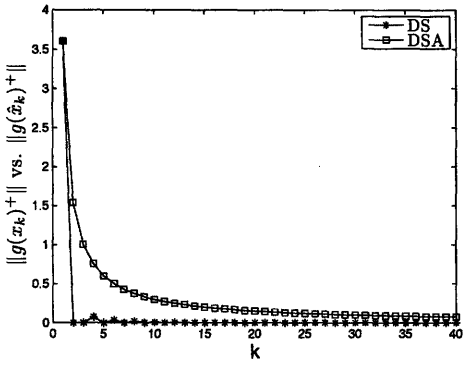


(c) $\alpha = 0.1$, $\mu_0 = 0$.

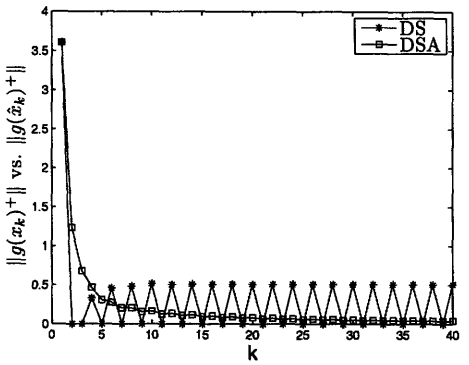


(d) $\alpha = 2$, $\mu_0 = 0$.

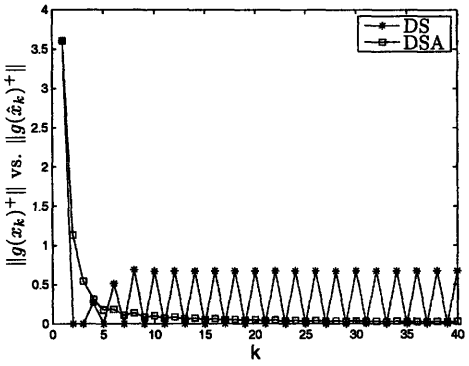
Figure 3-3: Three data flows $\{\hat{x}_k\}$ of DSA scheme with averaging, with α increasing from 0.03 to 2 and $\mu_0 = 0$. Convergence is guaranteed even when the stepsize is relatively large.



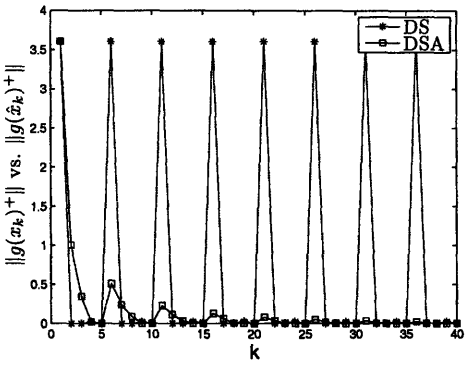
(a) Feasibility violation at $\alpha = 0.03$.



(b) Feasibility violation at $\alpha = 0.06$.

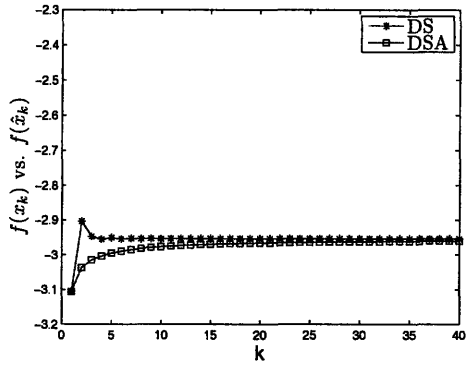


(c) Feasibility violation at $\alpha = 0.1$.

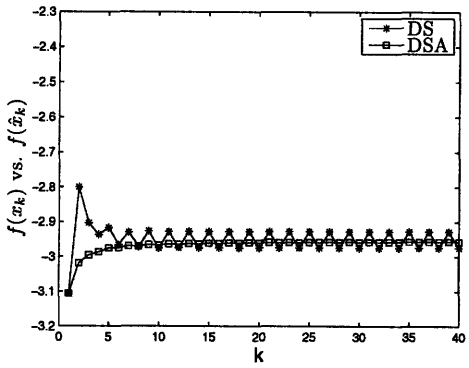


(d) Feasibility violation at $\alpha = 2$.

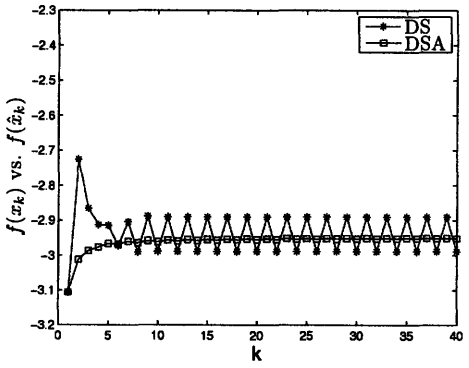
Figure 3-4: Comparison of feasibility violations in DS and DSA schemes, with α increasing from 0.03 to 2 and $\mu_0 = 0$. As α increases, $\|g(\hat{x}_k)^+\|$ consistently decreases to zero while $\|g(x_k)^+\|$ oscillates more severely.



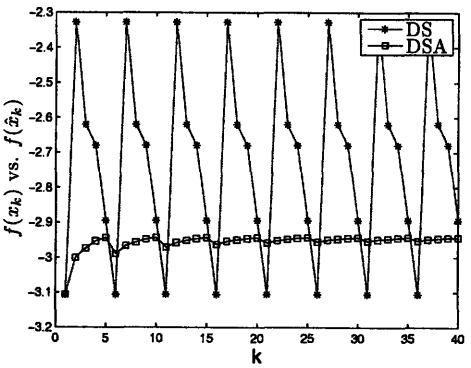
(a) Value of primal cost at $\alpha = 0.03$.



(b) Value of primal cost at $\alpha = 0.06$.

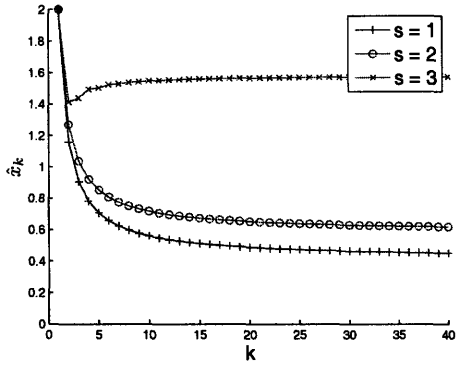


(c) Value of primal cost at $\alpha = 0.1$.

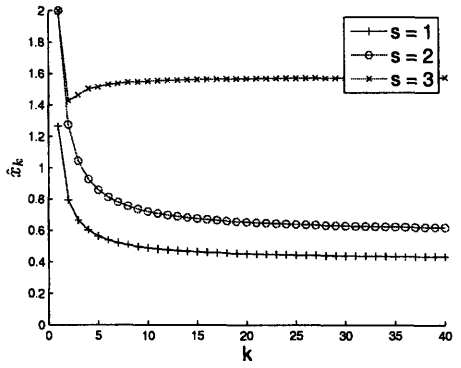


(d) Value of primal cost at $\alpha = 2$.

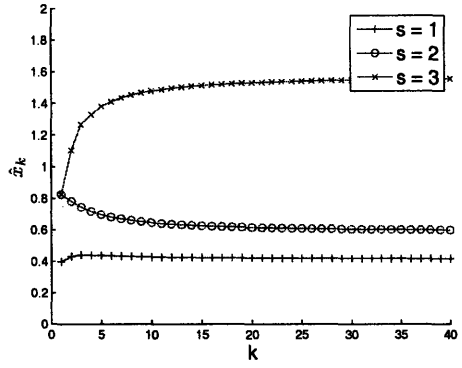
Figure 3-5: Comparison of primal costs in DS and DSA schemes, with α increasing from 0.03 to 2 and $\mu_0 = 0$. As α increases, $f(\hat{x}_k)$ consistently converges to a near-optimal solution while $f(x_k)$ oscillates more severely.



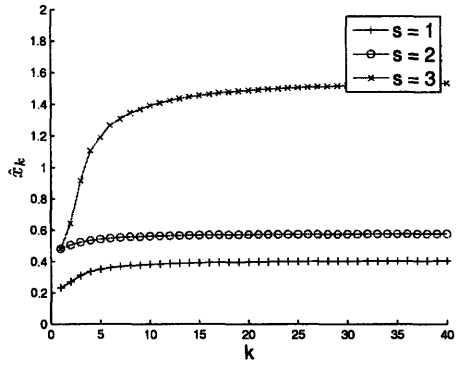
(a) $\mu_0 = 0$, $\alpha = 0.03$.



(b) $\mu_0 = 0.02$, $\alpha = 0.03$.

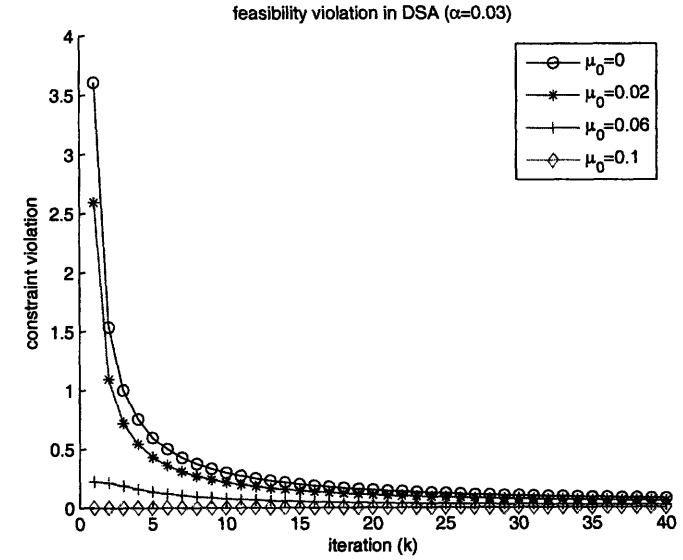


(c) $\mu_0 = 0.06$, $\alpha = 0.03$.

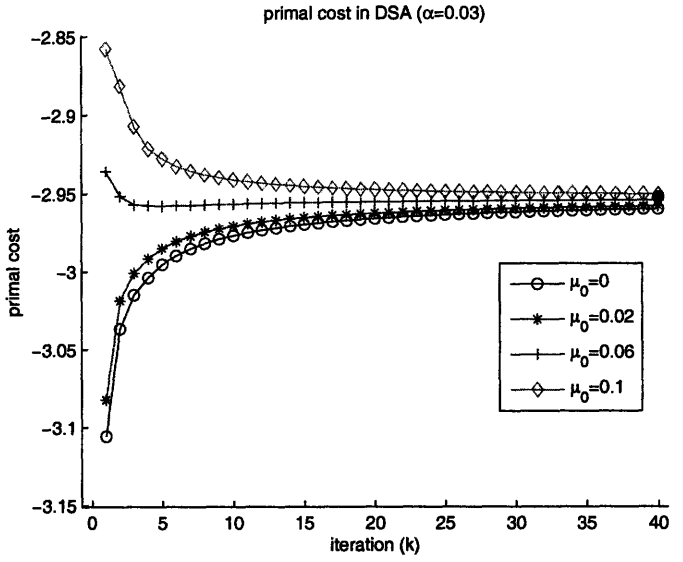


(d) $\mu_0 = 0.1$, $\alpha = 0.03$.

Figure 3-6: Infeasibility reduction for rate allocations of the DSA scheme, with μ_0 increasing from 0 to 0.01 and $\alpha = 0.03$. Note that the trends of convergence for the approximate primal sequence are different when μ_0 changes.

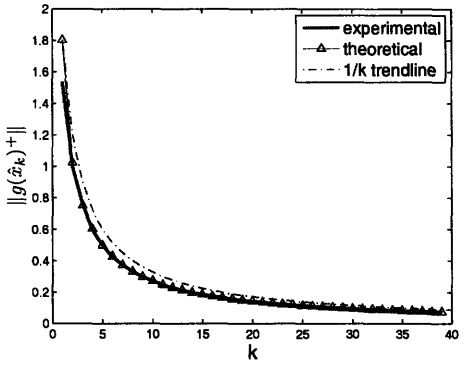


(a) $\|g(\hat{x}_k)^+\|$ at different μ_0.

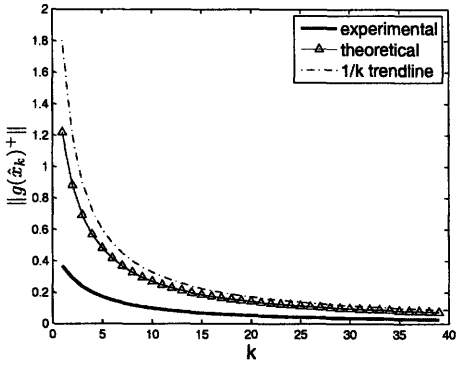


(b) $f(\hat{x}_k)$ at different μ_0.

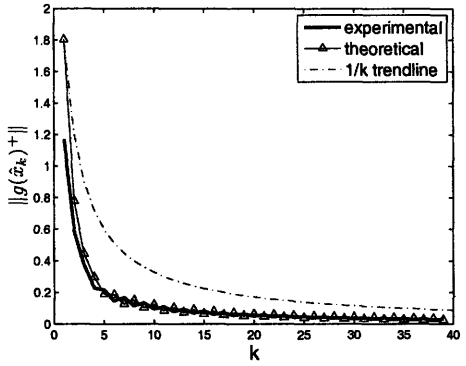
Figure 3-7: Trends of convergence for the feasibility violation and primal cost of $\hat{x}_k$ in DSA, with fixed $\alpha = 0.03$ and increasing μ_0.



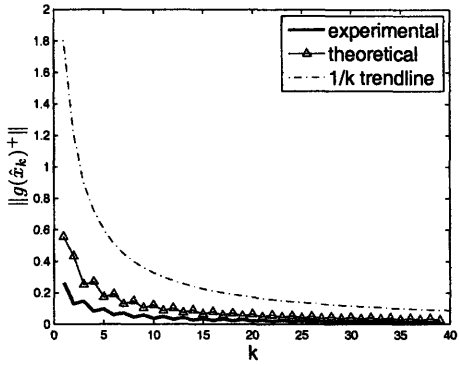
(a) $\mu_0 = 0$, $\alpha = 0.03$.



(b) $\mu_0 = 0.05$, $\alpha = 0.03$.

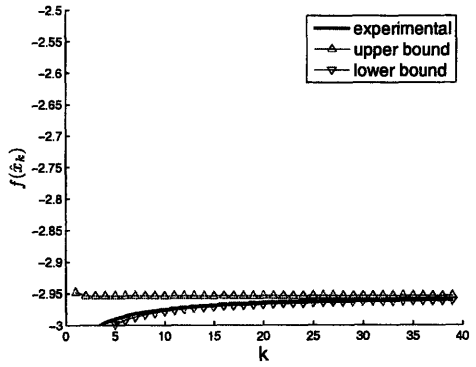


(c) $\mu_0 = 0$, $\alpha = 0.08$.

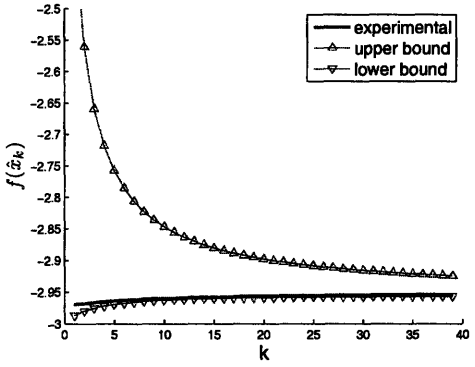


(d) $\mu_0 = 0.05$, $\alpha = 0.08$.

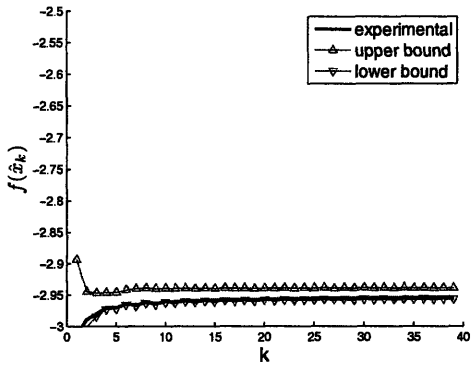
Figure 3-8: Comparison of experimental result with the theoretical bound for the feasibility violation in DSA, with different combinations of α and μ_0. The trendline with rate $1/k$ is also plotted.



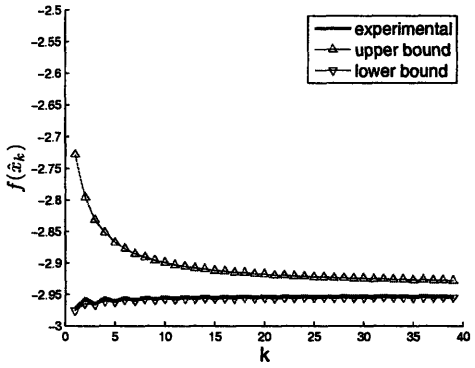
(a) $\mu_0 = 0$, $\alpha = 0.03$.



(b) $\mu_0 = 0.05$, $\alpha = 0.03$.



(c) $\mu_0 = 0$, $\alpha = 0.08$.



(d) $\mu_0 = 0.05$, $\alpha = 0.08$.

Figure 3-9: Comparison of experimental result with theoretical bounds for the value of primal cost in DSA, with different combinations of α and μ_0.

Chapter 4

Approximate Primal Sequence in the Nonconvex Case

In this chapter, we extend the results in convex minimization problems to general nonlinear optimization problems, which can be nonconvex. In Section 4.1, we first prove convergence of the averaged primal sequence generated in dual subgradient methods, and provide bounds for the value of primal cost in the existence of a duality gap. In Section 4.2, we investigate performance of the averaging scheme in the framework of NUM with nonconcave utilities. We show that in the nonconvex setting, the averaging scheme can still effectively reduce the amount of infeasibility; we also check its ability to recover the primal optimal solutions approximately. We compare its performance with other approaches in solving nonconcave NUM problems, and show that there are considerable benefits by employing the averaging scheme.

4.1 Convergence Analysis

In previous analysis, we make the assumption that both the objective function $f(x)$ and the constraint set $g(x)$ are convex, which results in a convex minimization problem. However, in a general setting, this assumption may not hold true, and we are faced with nonconvex optimization problems. It is intrinsically difficult to solve this class of problems, since we can hardly use the standard analyzing methods such as

Karush-Kuhn-Tucker conditions or duality theorem to identify the optimal primal solutions as in the convex case.

Depending on the convexity of objective and constraint functions, there are several types of nonconvex problems which have found different practical applications. Here, we focus on the category of *minimizing nonconvex objective with a convex constraint set.* An example is the congestion control of inelastic traffic such as the Internet. In the following sections, we analyze the properties of approximate primal solutions for this class of nonconvex problems, and carry out numerical experiments to verify our results and gain deeper insights of the system.

4.1.1 Convergence of Approximate Primal Sequence

Recall that the averaging scheme is built upon the dual subgradient method with a constant stepsize. We apply it to a general nonlinear optimization problem which can be either convex or nonconvex, and analyze the convergence properties of the approximate primal solution in the nonconvex setting.

For the averaged primal sequence $\{\hat{x}_k\}$ of Equation (2.8), we can show that it always converges under the condition that X is compact (cf. Assumption 2.1), regardless of the convexity of primal problem and properties of dual iterates. The following proposition proves this result.

Proposition 4.1 *Let X be a compact set (cf. Assumption 2.1). Let the approximate primal sequence $\{\hat{x}_k\}$ be the running averages of the primal iterates given in Equation (2.8). Then $\{\hat{x}_k\}$ can converge to its limit $\hat{x}^*$.*

Proof. To prove the convergence, we first show that $\{\hat{x}_k\}$ is a Cauchy sequence, i.e., for any $\varepsilon > 0$, there is a $K \in \mathbb{N}$ such that $\|\hat{x}_{k'} - \hat{x}_k\| < \varepsilon$, for any $k, k' \geq K$.

Given that $\hat{x}_k = \dfrac{1}{k}\sum_{i=0}^{k-1} x_i$ for all $k \geq 1$, we have

$$\begin{aligned}
\hat{x}_{k+1} &= \frac{1}{k+1}\sum_{i=0}^{k} x_i \\
&= \frac{k}{k+1}\left(\frac{1}{k}\sum_{i=0}^{k} x_i\right) \\
&= \frac{k}{k+1}\left(\frac{1}{k}\sum_{i=0}^{k-1} x_i + \frac{1}{k}x_k\right) \\
&= \frac{k}{k+1}\hat{x}_k + \frac{1}{k+1}x_k \\
&= \hat{x}_k - \frac{1}{k+1}\hat{x}_k + \frac{1}{k+1}x_k.
\end{aligned}$$

Hence, $\hat{x}_{k+1} - \hat{x}_k = \dfrac{x_k - \hat{x}_k}{k+1}$.

Without loss of generality, we assume that $k' > k$ (the result immediately follows when $k' = k$). As X is a compact convex set, we have $x_k, \hat{x}_k \in X$, and $\|x_k\|, \|\hat{x}_k\| \leq M$ under the boundedness assumption of the sequences, where $M \geq 0$. By applying the above expression iteratively, we have

$$\begin{aligned}
\|\hat{x}_{k'} - \hat{x}_k\| &= \|\hat{x}_{k'} - \hat{x}_{k'-1} + \cdots + \hat{x}_{k+1} - \hat{x}_k\| \\
&= \|\frac{x_{k'-1} - \hat{x}_{k'-1}}{k'} + \cdots + \frac{x_k - \hat{x}_k}{k+1}\| \\
&\leq \frac{\|x_{k'-1}\| + \|\hat{x}_{k'-1}\|}{k'} + \cdots + \frac{\|x_k\| + \|\hat{x}_k\|}{k+1} \\
&\leq \frac{2M(k'-k)}{k+1}.
\end{aligned}$$

Thus, for any arbitrary $\varepsilon > 0$, we can let $\dfrac{2M(k'-k)}{k+1} < \varepsilon$. In other words, for any k satisfying $k = k(\varepsilon) \geq K = \lceil \dfrac{2M(k'-k)}{\varepsilon} - 1 \rceil$, it follows that

$$\|\hat{x}_{k'} - \hat{x}_k\| < \varepsilon \qquad \forall k,\ k' \geq K,$$

where $\lceil x \rceil$ is defined as the *ceiling function* which returns the smallest integer larger or equal to x. Note that $(k' - k)$ is the relative index number, which is independent of k and can take any non-negative integer number. Thus, K is well defined and $\{\hat{x}_k\}$

is a Cauchy sequence.

Since the Cauchy sequence $\{\hat{x}_k\}$ is also bounded, we say that it has a subsequence b_n converging to L (to be defined later). For any $\varepsilon > 0$, we choose $K \in \mathbb{N}$ so that there exist $n, m \geq K$ satisfying $\|\hat{x}_n - \hat{x}_m\| < \varepsilon/2$. Thus, there is a $b_k = \hat{x}_{m_k}$, such that $m_k \geq K$ and $\|b_{m_k} - L\| < \varepsilon/2$. We have

$$\begin{aligned}
\|\hat{x}_n - L\| &= \|\hat{x}_n - b_k + b_k - L\| \\
&\leq \|\hat{x}_n - b_k\| + \|b_k - L\| \\
&< \|\hat{x}_n - \hat{x}_m\| + \varepsilon/2 \\
&< \varepsilon/2 + \varepsilon/2 \\
&= \varepsilon.
\end{aligned}$$

Since ε can be arbitrarily small, the sequence $\{\hat{x}_k\}$ converges to L asymptotically. Thus, we can denote L as the limit of the sequence $\{\hat{x}_k\}$, i.e. $L = \hat{x}^*$.

■

Thus, the approximate primal sequence in the dual subgradient scheme with primal-averaging always converges. The distance of $\hat{x}_k$ to $\hat{x}^*$ can be controlled by the number of iterations.

4.1.2 Properties of Nonconvex Optimization

It follows next to check whether such an approximate primal vector in the converging sequence is near-feasible and near-optimal to the global optimal. Note that under the condition that $g(x)$ is convex, Proposition 3.1(a) and Proposition 3.2 hold for any nonlinear objective function $f(x)$. However, the bounds of primal cost for the vector $\hat{x}_k$ need to be modified.

As it is intrinsically difficult to trace the properties of general nonconvex optimization problems, we first make the following assumption for the group of problems we focus on.

Assumption 4.1 (Nonconvex Problem)

1. *The objective function is separable, i.e.* $f(x) = \sum_{i=1}^{n} f_i(x_i)$.
2. *The constraint* $g(x)$ *is linear; in the matrix form, we can have* $Ax \leq c$.
3. *There are* $0 < r \leq n$ *nonconvex functions in the* n *local objective functions of the form* $f_i(x_i) : \mathbb{R} \to \mathbb{R}$, *thus the objective function is nonconvex.*

Clearly, nonconvex optimization problems under this assumption are decomposable. Recall that NUM problem has a structure satisfying these assumptions. Thus, in the following analysis, we consider $min\, f(x)$ interchangeable with $max\, U(x)$. We discuss the conditions under which the previous convergence results under convexity still hold, and next we show the new bounds when the duality gap is strictly positive.

Optimality Condition for Zero Duality Gap

It is important to note that even for nonconvex problems, the duality gap can be zero. In this case, the results in Proposition 3.1 still hold, since we have $f^* = q^*$. Chiang et al. [9] discuss the conditions for global convergence of NUM. We present it below (with slight modifications) and it can be readily extended to optimization problems with the structure under Assumption 4.1.

Proposition 4.2 (Chiang et al. [6]) *For problems with the structure of NUM framework, if the primal solution* $x(\mu)$ *generated in the dual subgradient method is continuous at the dual optimal solution* μ^*, *the duality gap is zero.*

Proof. In NUM, the dual function is convex (possibly nondifferentiable). At the dual optimal solution $\mu^* = (\mu_1^*, \cdots, \mu_L^*)$, we have the optimality condition:

$$[\partial q(\mu^*)]^T(\mu - \mu^*) \leq 0 \qquad \forall \mu \geq 0.$$

As the problem is dual decomposable, we consider each component of the dual solution by letting $\overline{\mu_l^*} = \mu_l^* + \delta$ and $\underline{\mu_l^*} = \mu_l^* + \delta$, where $\delta > 0$ is arbitrarily small.

Now we consider two cases for each component of the nonnegative vector μ.

Case 1: $\mu_l > 0$.

As $\overline{\mu}_l^* > \mu_l^*$, the subgradient at $\overline{\mu}_l^*$ is negative and $\sum_{s \in S(l)} x_s^*(\overline{\mu}^{s*}) \leq c_l$. Similarly, as $\underline{\mu}_l^* < \mu_l^*$, the subgradient at $\underline{\mu}_l^*$ is positive and $\sum_{s \in S(l)} x_s^*(\underline{\mu}^{s*}) \geq c_l$.

Given that $x(\mu)$ is continuous at μ^* and δ is sufficiently small, we have

$$\sum_{s \in S(l)} x_s^*(\mu^{s*}) = c_l.$$

Case 2: $\mu_l = 0$.

Now consider $\overline{\mu}_l$ only, we can have $\sum_{s \in S(l)} x_s^*(\overline{\mu}^{s*}) \leq c_l$ from previous discussion. Under continuity, it follows that

$$\sum_{s \in S(l)} x_s^*(\mu^{s*}) \leq c_l.$$

The complementary slackness is hence satisfied,

$$\left[\sum_{s \in S(l)} x_s^*(\mu^{s*}) \right] \mu_l = 0 \qquad \forall l.$$

Together with the definitions of dual function, primal optimal and weak duality, we have

$$\begin{aligned} q^* &= q(\mu^*) \\ &= \max_x L[x(\mu^*), \mu^*] \\ &= \textstyle\sum_s U_s[x_s^*(\mu^{s*})] + \sum_l \mu_l [c_l - \sum_{s \in S(l)} x_s^*(\mu^{s*})] \\ &= \textstyle\sum_s U_s[x_s^*(\mu^{s*})] \\ &\leq U^* \\ &\leq q^*. \end{aligned}$$

Therefore, it follows that $U^* = q^*$ and there is no duality gap.

■

This result suggests that continuity of $x(\mu)$ at μ^* is a sufficient condition for optimality. As discussed in Chiang et al. [9], if we can bound the range of μ^* so as to exclude the discontinuous points, we can guarantee the continuity of $x(\mu)$ and thus eliminate the duality gap. However, we are not going into details of this method here, as it is problem-specific to check the continuity condition. In our approach, we simply apply the averaging scheme to any problem (either convex or nonconvex), and we are interested in evaluating the performance of approximate primal solutions. Specifically, if the continuity condition is met, Proposition 3.1 holds and the bounds is illustrated in Figure 4-1. In next part, we discuss the situation when the duality gap is positive.

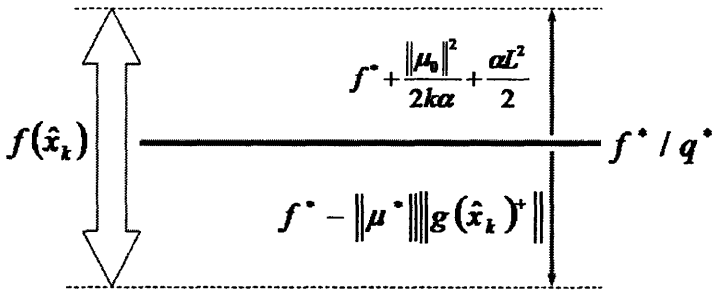


Figure 4-1: Illustration of bounds for the value of objective function at an approximate primal solution, in the nonconvex case under the continuity condition which implies zero duality gap.

Bounds with Duality Gap

When the condition for zero duality gap is no longer satisfied (cf. Proposition 4.2), the previous algorithms solving the dual problem may produce infeasible primal solutions and the system may not converge at all, since solving the dual problem is not equivalent to solving the primal problem any more.

Now we consider the case when the duality gap is positive. We first present some definitions for further analysis.

Definition 4.1. *Let $X \subset \mathbb{R}^n$ be convex and compact, and let $f : X \to \mathbb{R}$ be lower semicontinuous on X. A function $\tilde{f} : X \to \mathbb{R}$ is called the* ***convex envelope*** *of f on X if it satisfies*

1. *$\tilde{f}(x)$ is convex on X.*
2. *$\tilde{f}(x) \leq f(x)$ for all $x \in X$.*
3. *There is no function $\hat{f}(x) : X \to \mathbb{R}$, satisfying the first two conditions and $\tilde{f}(\bar{x}) \leq \hat{f}(\bar{x})$ for some point $\bar{x} \in X$.*

In other words, we have $\tilde{f}(x) \geq \hat{f}(x)$ for all $x \in X$. Thus, $\tilde{f}(x)$ is the tightest convex relaxation for the original function $f(x)$.

Definition 4.2. *Let f be a real-valued function defined on some convex set X. We can measure its* ***lack of convexity*** *by the number*

$$\rho(f) = \sup\{f(\sum \alpha_i x_i) - \sum \alpha_i f(x_i)\},$$

where $\alpha_i \in [0,1]$ with $\sum \alpha_i = 1$, and $x_i \in X$. It can be seen that $\rho(f) = 0$ if and only if f is convex.

We prove the following Lemmas using these definitions.

Lemma 4.1. *For any function $f(x)$, its convex envelope and lack of convexity are related by*

$$0 \leq f(x) - \tilde{f}(x) \leq \rho(f) \qquad \forall x \in X.$$

Proof. For any function $\bar{f}(x)$ satisfying Definition 4.1(1) and (2), we have $\bar{f}(x) \leq f(x)$ and the following holds due to convexity,

$$\begin{aligned} \bar{f}(\sum \alpha_i x_i) &\leq \sum \alpha_i \bar{f}(x_i) \\ &\leq \sum \alpha_i f(x_i) \qquad \forall x \in X. \end{aligned}$$

By Definition 4.1(3), we have $\tilde{f}(x) \geq \bar{f}(x)$, where $\tilde{f}(x)$ is the tightest underestimation of $f(x)$. Thus, at some $\bar{x}_i$ where $f(\sum \alpha_i \bar{x}_i) > \sum \alpha_i f(\bar{x}_i)$ [i.e. the nonconvex portion of $f(x)$], we have $\tilde{f}(\sum \alpha_i \bar{x}_i) = \sum \alpha_i f(\bar{x}_i)$.

As we can write $x = \sum \alpha_i x_i$ for any $x \in X$, it follows that

$$\begin{aligned} \sup(f(x) - \tilde{f}(x)) &= \sup\{f(\textstyle\sum \alpha_i x_i) - \tilde{f}(\sum \alpha_i x_i)\} \\ &= \sup\{f(\textstyle\sum \alpha_i x_i) - \sum \alpha_i f(x_i)\} \\ &= \rho(f). \end{aligned}$$

Thus, we can have $0 \leq f(x) - \tilde{f}(x) \leq \rho(f)$, for all $x \in X$.

■

Lemma 4.2. *Let Assumption 4.1(1) hold [i.e. $f(x) = \sum_{i=1}^n f_i(x_i)$]. Then, we have $\tilde{f}(x) = \sum_{i=1}^n \tilde{f}_i(x_i)$, and moreover, $f(x) - \tilde{f}(x) \leq \sum_{i=1}^n \rho(f_i)$.*

Proof. See Theorem IV.8, Horst and Tuy [13], for the proof that $\tilde{f}(x) = \sum_{i=1}^n \tilde{f}_i(x_i)$. It then follows that

$$\begin{aligned} \sup\{f(x) - \tilde{f}(x)\} &= \sup\{\textstyle\sum_{i=1}^n f_i(x_i) - \sum_{i=1}^n \tilde{f}_i(x_i)\} \\ &= \sup\{\textstyle\sum_{i=1}^n [f_i(x_i) - \tilde{f}_i(x_i)]\} \\ &\leq \textstyle\sum_{i=1}^n \sup\{f_i(x_i) - \tilde{f}_i(x_i)\} \\ &\leq \textstyle\sum_{i=1}^n \rho(f_i). \end{aligned}$$

Thus, $f(x) - \tilde{f}(x) \leq \sup\{f(x) - \tilde{f}(x)\} \leq \sum_{i=1}^n \rho(f_i)$.

■

Lemma 4.3 *For any function $f(x)$ with an optimum, we have $f^*(x) \geq q^*(\mu) \geq \tilde{q}^*(\mu) = \tilde{f}^*$. Moreover, if Assumption 4.1(2) holds [i.e., $g(x)$ is linear], we have*

$$f^*(x) \geq q^*(\mu) = \tilde{q}^*(\mu) = \tilde{f}^*.$$

Proof. We have $\tilde{q}(\mu)$ as the dual of the convex envelope $\tilde{f}(x)$, thus $\tilde{q}^* = \tilde{f}^*$. By the

definitions of dual function and convex envelope, we have

$$\begin{aligned} q(\mu) &= \max_x\{f(x) + \mu^T g(x)\} \\ &\geq \max_x\{\tilde{f}(x) + \mu^T g(x)\} \\ &= \tilde{q}(\mu). \end{aligned}$$

Under the linearity assumption of $g(x)$, it is proved in Horst and Tuy [13] that $\tilde{q}(\mu) = q(\mu)$ for any μ. By applying Weak Duality Theorem to $f(x)$, we can obtain the above expression.

■

Note that the duality gap is defined as $\eta = f^* - q^*$, and for any general function, we have $f^* - q^* \leq f^* - \tilde{f}^* \leq \rho(f)$. Under the condition of linear constraints, we can have $\eta = f^* - q^* = f^* - \tilde{f}^* \leq \rho(f)$. In this case, duality gap is the same as the difference between original function and the convex envelope at optimal.

In the next proposition, we provide per-iterate bounds for the amount of constraint violation and value of primal cost of the average vectors $\hat{x}_k$, to the nonconvex optimization problems where $f(x)$ is nonconvex, $g(x)$ is convex and X is a nonempty compact convex set. Note that this is a more general setting than problems under Assumption 4.1. Relation of the bounds is illustrated in Figure 4-2.

Proposition 4.3 *Let the bounded subgradient assumption hold. Let the dual sequence $\{\mu_k\}$ be generated by the subgradient method of Equation (2.5). For the averaged primal sequence defined in Equation (2.7), we have the following results for all $k \geq 1$:*

(a) An upper bound on the amount of constraint violation of the vector $\hat{x}_k$ as

$$\|g(\hat{x}_k)^+\| \leq \frac{\|\mu_k\|}{k\alpha}.$$

(b) An upper bound on the primal cost of $\hat{x}_k$ as

$$f(\hat{x}_k) \leq q^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha L^2}{2} + \rho(f). \tag{4.1}$$

(c) and a lower bound on the primal cost of $\hat{x}_k$ as

$$f(\hat{x}_k) \geq q^* - \|\mu^*\| \|g(\hat{x}_k)^+\|,$$

where $\rho(f)$ is the lack of convexity for $f(x)$.

Proof.

(a) The proof follows the same argument as in the proof of Proposition 3.1(a).

(b) Let $\tilde{f}(x)$ be the convex envelope of $f(x)$. Since $\hat{x}_k \in X$ and $\tilde{q}^* \leq q^*$, by Lemma 4.1 and Proposition 3.1(b), we have

$$\begin{aligned} f(\hat{x}_k) &\leq \tilde{f}(\hat{x}_k) + \rho(f) \\ &\leq q^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha L^2}{2} + \rho(f). \end{aligned}$$

(c) The proof follows the same argument as in Proposition 3.1(c), except that we cannot replace q^* with f^*, since $q^* \leq f^*$ in the nonconvex case.

■

Note that if $f(x)$ is separable, we have $\rho(f) = \sum_{i=1}^n \rho(f_i) = r \max_i \rho(f_i)$ by Lemma 4.2, where $r \leq n$ is the number of nonconvex functions $f_i(x_i)$. We can further estimate the duality gap for individual functions with special structures, which is discussed in the following section.

4.2 Applications to Nonconcave NUM

In this section, we test the theoretical results in the NUM framework with nonconcave utility functions. We observe how the DSA scheme performs in various test instances and contrast it to other approaches in solving nonconcave NUM problems.

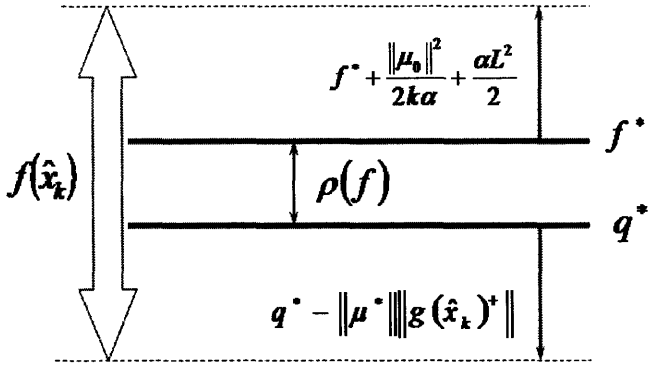


Figure 4-2: Illustration of bounds for the value of objective function at an approximate primal solution, in the nonconvex case with positive duality gap.

4.2.1 Nonconcave Utility Functions

In addition to elastic flows which can adjust the transmission data rates gradually in response to congestion levels of the network, there exist other types of traffic which are not elastic. Some examples include voice, streaming video and audio services. Based on the specific service requirements, we can take different forms of the utility functions. In particular, we use smooth and concave functions to model elastic traffic, and nonsmooth or nonconcave utilities for inelastic flows. It is also common to have a mixture of inelastic and elastic data flows.

In last chapter, we have discussed performance of the averaging scheme in elastic application. In the inelastic traffic, however, the nonconcavity of utility functions results in nonconvex optimization problems which are significantly more difficult to solve. Moreover, of the three approaches tackling nonconcave NUM reviewed in Chapter 1, only the rate control scheme with self-regulation provides a distributive algorithm for suboptimal solutions. There is a scarcity of methods to solve such problems.

Considering its robustness in introducing convergence properties, we evaluate the averaging scheme in the nonconcave NUM. We focus on some typical nonconcave utilities, in particular, sigmoidal and quadratic functions.

Example 4.1 (Nonconcave Utilities)

Sigmoidal utility function is typically used in nonconcave NUM, with the general form:

$$U(x) = c\{\frac{1}{1+e^{-a(x-b)}} - d\}. \tag{4.2}$$

The parameters a and b control the shape and position of the sigmoid functions. We set $d = \frac{1}{1+e^{ab}}$ to ensure $U(0) = 0$, and $c = \frac{1+e^{ab}}{e^{ab}}$ such that $U(\infty) = 1$. If normalization is not necessary, we can let $c = 1$. We illustrate sigmoidal utilities with different parameters in Figure 4-3.

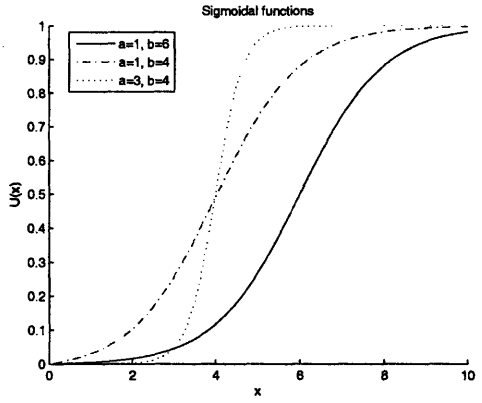
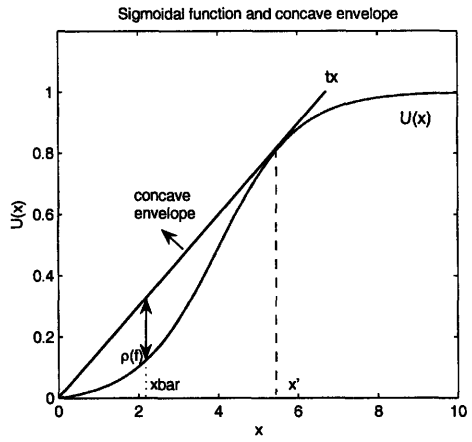


Figure 4-3: Sigmoidal functions with different parameters.

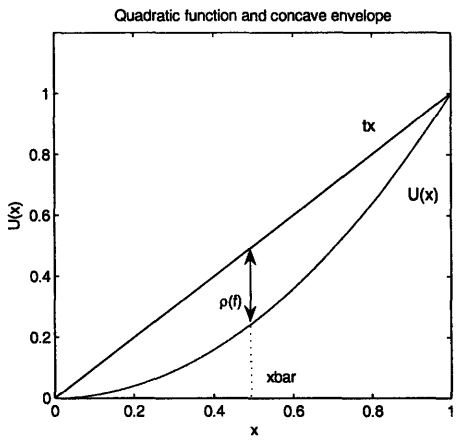
Note that the sigmoidal function has an *inflection point* x_o, at which $\frac{d^2f(x)}{dx^2} = 0$. When $x < x_o$, $\frac{d^2f(x)}{dx^2} > 0$, it is on the convex portion; when $x > x_o$, $\frac{d^2f(x)}{dx^2} < 0$, it is on the concave portion. Also note that, parameter a controls the shape of the sigmoidal function while b determines the position of the inflection point.

The simple quadratic function $U(x) = x^2$ can model nonconcave utilities as well. It is useful when the bottleneck link capacity limits sources to their convex region of a sigmoidal utility.

Based on the formulations, we can estimate the duality gap for specific nonconcave utility functions. The following propositions demonstrate the procedures to obtain the estimation.



(a) Sigmoidal function and its concave envelope.



(b) Quadratic function and its concave envelope.

Figure 4-4: Illustration of lack of concavity $\rho(f)$ for two typical nonconcave utility functions.

Proposition 4.4 *Let $U(x)$ be the normalized sigmoidal function of Equation (4.2). Let $\rho(f)$ be the lack of concavity in Definition 4.2. If we let $a = 1$, the duality gap is not greater than $1/2$.*

Proof. First, we can define the concave envelope of $U(x)$ [cf. Figure 4-4(a)] as

$$\tilde{U}(x) = \begin{cases} tx & 0 \leq x \leq x' \\ U(x) & x \geq x', \end{cases}$$

where t is the slope of tangent of the straight line from origin with the sigmoidal curve, and x' is the coordinate on x-axis of the intersection point. We can calculate x' by solving $U(x) = xU'(x)$. It follows that $t = U'(x')$. Note that while the gradient t can be obtained by solving the nonlinear problem with computational tools such as Newton's method, it is very difficult to get a theoretical formulation.

Second, we calculate the lack of concavity by

$$\rho(f) = \max_{0 \leq x \leq x'} \{tx - U(x)\},$$

where we denote the solution as $\bar{x}$.

By setting its derivative to zero, we have

$$t = U'(x) = \frac{(ac)y}{(1+y)^2} \qquad \text{where } y = e^{-a(x-b)}.$$

After rearranging, we get a quadratic function $y^2 + (2 - ac/t)y + 1 = 0$. Note that in the sigmoidal function, $U(b) \approx 1/2$, and $\bar{x}$ must fall in the range of $[0, b)$ [cf. Figure 4-4(a)]. Thus, we must have $y > 1$. In particular, the solution is

$$y = \frac{1}{2t}\left[ac - 2t + \sqrt{(ac)(ac-4t)}\right].$$

$\bar{x}$ can be obtained once y is known,

$$\begin{aligned}\bar{x} &= -\frac{\ln y}{a} + b \\ &= -\frac{1}{a}\ln\left[\frac{ac - 2t + \sqrt{(ac)(ac-4t)}}{2t}\right] + b.\end{aligned}$$

Third, we can obtain the lack of convexity by evaluating $\rho(f)$ at $\bar{x}$,

$$\begin{aligned}\rho(f) &= t\bar{x} - U(\bar{x}) \\ &= tb + cd + \frac{t}{a}\ln\left[\frac{2t}{ac - 2t + \sqrt{(ac)(ac-4t)}}\right] - \frac{2ct}{ac + \sqrt{(ac)(ac-4t)}}.\end{aligned}$$

Note that the last two terms of the above expression are negative. If ab is relatively large, we have $c = \frac{1+e^{ab}}{e^{ab}} \to 1$ and $d = \frac{1}{1+e^{ab}} \to 0$. In fact, when $ab = 5$, $|c-1| < 1\%$ and when $ab = 7$, $|c-1| < 0.1\%$. Thus, e^{ab} grows very fast and we can approximate $c \approx 1$ and $d \approx 0$. Even though we cannot get a closed-form solution for t, we can try to estimate the lack of concavity (an upper bound of the duality gap) as below.

Assume that $a = 1$, which holds for our testing examples. At $x = b$, the function value is $\frac{1}{2} - \frac{1}{2e^{ab}}$, which is slightly below half of the maximum function value. Note that $x' > b$, which falls in the concave region of $U(x)$. If b is relatively large, the shape of the sigmoidal is "flat" and the gradient t at x' gets close to $\frac{1}{2b}$. Therefore, we can estimate $t \approx \frac{1}{2b}$.

As mentioned before, $c \approx 1$ and $d \approx 0$ in most cases. Substituting these values into the expression of $\rho(f)$, we can obtain $\rho(f) \leq \frac{1}{2} - \frac{1/b}{1+\sqrt{1-2/b}} < \frac{1}{2}$.

Even though the result is obtained by estimation, it agrees with the experimental results in our experience, i.e., the duality gap in the sigmoidal function is always (much) less than half of its maximum value.

■

Proposition 4.5 *Let the normalized quadratic function be* $U(x) = x^2, 0 \leq x \leq 1$. *Let* $\rho(f)$ *be the lack of concavity of Definition 4.2. Then the duality gap is not greater than 1/4.*

Proof. From the illustration in Figure 4-4(b), it is straightforward to see that the concave envelope is the line relaxation connecting the two end points, i.e. $\tilde{U}(x) = x$. Thus we have

$$\rho(f) = \max_{0 \leq x \leq 1} \{x - x^2\}.$$

By setting $\partial\rho(f) = 1 - 2x = 0$, we have $\bar{x} = 1/2$ and thus $\rho(f) = 1/4$. Thus, the duality gap for a quadratic function is at most one quarter of its function value.

■

Example 4.2 (Price-based Data Rate Allocation)

To show that different types of utility functions respond to the network price and adjust the data rate differently, we plot the price-based rate allocation $x(\mu)$ for: (1) logarithmic; (2) sigmoidal; and (3) quadratic functions, as shown in Figure 4-5.

These curves imply elasticity level of the data flows. We can see that for the concave function, $x(\mu)$ has a continuously differentiable curve and it decreases to zero smoothly throughout its domain; for the sigmoidal function, the primal vector can be adjusted gradually according to the dual feedback only up to a certain level, and there is a discontinuity at the μ_{max}; and for the convex function, the allocated rate is either at the maximum value or zero. The discontinuity in the nonconcave case implies that it is where the instability and excessive congestion of the system originate from.

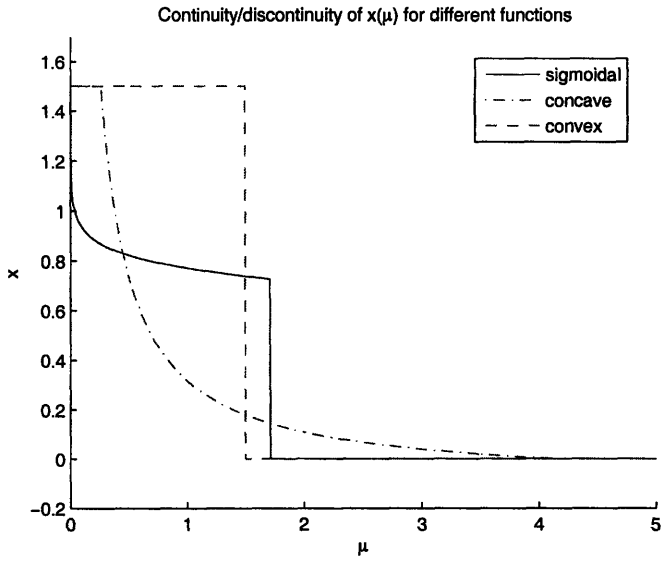


Figure 4-5: Illustration of price-based rate allocation $x(\mu)$, for different types of utility functions. Discontinuity can be observed in the nonconcave case.

4.2.2 Models and Algorithms

We compare the averaging scheme to the rate control algorithm with a self-regulation heuristic, which takes a different approach in tackling the issue of instability from nonconcavity of utility functions. We denote the models tested in the numerical experiment with the following abbreviations.

DS canonical distributed algorithm in NUM using the standard dual subgradient method (see Low and Lapsley in [19]).

DSA distributive dual subgradient methods with the proposed averaging scheme (see Chapter 2).

SR rate control algorithm with self-regulation property (see below for a brief description and Lee et al. [18] for details).

SRA modified self-regulation scheme combined with the averaging method.

We first present the "self-regulating" properties in the SR scheme.

Properties of SR

1. Each user satisfies the "self-regulating" policy.

2. Each user s has thresholds of tolerance th_s and δ_s, such that if it consecutively receives net utility less than or equal to δ_s by transmitting data for th_s iterations, it stops sending data. Note that if we consider a homogenous system, we can define the same th and δ for all users.

3. Link l allocates a received rate (we can consider it as the final output data rate) to each user $s \in S(l)$ by

$$x'_s = \begin{cases} x_s & \text{if } \sum_{i \in S(l)} x_i \leq c_l \\ f_{sl}(\bar{x}_l) & \text{if } \sum_{i \in S(l)} x_i > c_l, \end{cases}$$

   where x_s is the transmission data rate of user s (we can consider it as the primal iterate before its feasibility is checked), $\bar{x}_l$ is a vector for the transmission rate of users in $S(l)$, and we can choose f_{sl} as the continuous function

$$f_{sl}(\bar{x}_l) = \frac{x_s}{\sum_{i \in S(l)} x_i} c_l.$$

We can see that the SR scheme is essentially built upon the DS algorithm, with the additional properties to (locally) turn off users which cause system congestions due to the discontinuity in rate allocation. As such, if the continuity condition for zero duality gap is satisfied, SR reduces to standard DS, since the self-regulation policy is not executed; otherwise, it converts the original nonconvex problem with positive duality gap into a smaller (fewer users) problem with zero duality gap (the continuity condition of Proposition 4.2 is satisfied in the new problem). Essentially, the SR algorithm generates a suboptimal solution to the original nonconvex problem, which is optimal for the new problem.

Note that congestion control is the first criteria in the SR algorithm (a practical example is the multi-class services in the Internet); thus, property 3 is to externally

impose congestion control if the estimated primal solution is still infeasible. All the rate allocations at output are guaranteed to be feasible under the self-regulation.

In the experiment, we implement SR to compare with the performance of proposed DSA scheme. To alleviate the possible system oscillations, we also consider applying the averaging method to SR as an additional technique to impose convergence properties to the system. Thus, the combined scheme SRA is simply to compute the by-product $\hat{x}_k$ during the process of self-regulation, as described below.

Algorithms for SR and SRA

1. *Set the thresholds th_s and δ_s.*
2. *Initialize μ_0 and α.*
3. *Compute $x_s[\mu^s(k)]$ and its net utility $L_s(x_s, \mu^s)$, for each user s; if it receives net utility less than or equal to δ for th_s iterations consecutively, this user s is turned off locally, i.e., we assign zero data rate to it afterwards and continue the algorithm with the remaining users.*
4. *Update $\mu_l(k)$ for each link l.*
5. *Repeat steps 3 to 4 if N iterations have not completed; otherwise, get the temporary sequence of $\{x_k\}$ without applying property 3 of self-regulating.*
6. *(for SR) Check the congestion level at each link l and impose feasibility to the data rate by applying property 3.*
7. *(for SRA) Compute $\hat{x}_k$ by running averages of the temporary primal iterates $\{x_k\}$; and apply property 3 to the averaged sequence to ensure feasibility.*
8. *Evaluate the performance parameters for $\{x_k\}$ ($\hat{x}_k$ in SRA) at the master level.*

4.2.3 Numerical Results

In the experiment, we consider the 3-user 2-link line topology and 4-user 4-link tree topology introduced in Example 2.1. Our test instances include homogeneous (same

source utilities) and heterogeneous (different source utilities) systems. Note that the objective function is concave if any of the source utility is concave.

We summarize the nonconcave utilities used in the experiment in Table 4.1, and plot the price-based rate allocations in Figure 4-6.

Type	Notation	Function	M	μ_{max}	$x(\mu_{max})$
Sigmoidal	S1	$U(x) = 1/[1+e^{-(x-5)}] - 1/(1+e^5)$	15	0.22	$0-5.7$
	S2	$U(x) = 1/[1+e^{-(x-1)}] - 1/(1+e^1)$	15	0.25	$0-1.21$
	S3	$U(x) = 1/[1+e^{-(x-8)}] - 1/(1+e^8)$	15	0.2	$0-9$
Quadratic	Q	$U(x) = x^2$	2	2	$0-2$

Table 4.1: Nonconcave utilities and the corresponding range of discontinuity for price-based rate allocation $x(\mu)$, at $\mu_{\max}$.

Similar to the convex case, we first convert the maximization problems to the general form of minimizing the negative of utilities. In the tables, the estimated cost is changed back to the positive value [i.e. $U(x) = -f(x)$]. Typically, we run the simulation for $N = 50$ times and evaluate the performance of approximate solutions at each iteration.

Example 4.3 (Homogeneous system; line topology; sigmoidal utilities)

Consider a homogeneous system with the sigmoidal function $S1$, which is tested with three sets of link capacity,

$$\begin{aligned} \max \quad & \textstyle\sum_{s=1}^{3} S1(x_s) \\ s.t. \quad & x_1 + x_2 \le c_1 \\ & x_1 + x_3 \le c_2 \\ & 0 \le x_1, x_2, x_3 \le \max(c_1, c_2). \end{aligned}$$

(a) $c = [4; 8]$. The global optimal is

$$x^* = [0; 4; 8], \qquad U^* = 1.2081.$$

In the simulation, we vary the parameters for the two pairs of systems (namely,

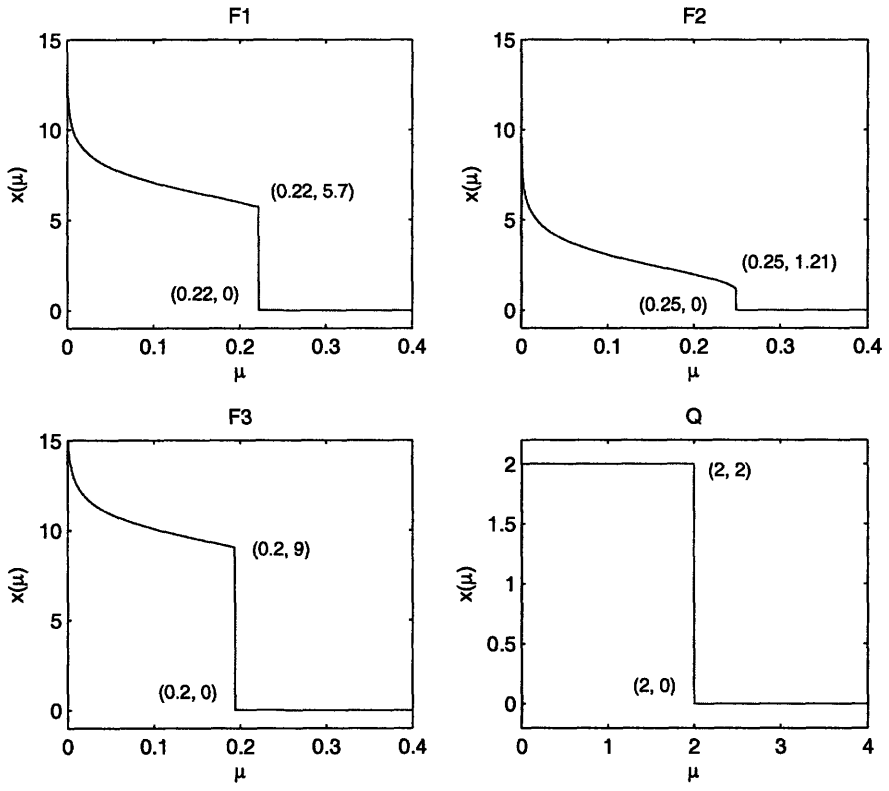


Figure 4-6: Illustration of dual-based primal $x(\mu)$ for corresponding function.

DS/DSA and SR/SRA) and compare the performance of approximate primal solutions.

Table 4.2 shows the numerical results in DS $[\|g(x_k)^+\|, f(x_k)]$ and DSA $[\|g(\hat{x}_k)^+\|, f(\hat{x}_k)]$ at different μ and α. The quantity $(\mu_0)_l$ denotes the value of each component of the initial dual vector.

Figure 4-7 shows the infeasibility reduction in DSA and the trend of convergence for $\{\hat{x}_k\}$, under an increasing μ_0 and a fixed stepsize. We can see that when $\mu_0 = 0$, the congestion level is relatively high in the nonconcave system (it is still 0.3521 at N) and it can be effectively reduced to zero when μ_0 is increased to 0.2.

We make the following observations on different parameters in DS/DSA schemes:

- α affects the quality of primal cost $f(x)$. If the stepsize is too small, it takes a longer time to converge to a certain level of the optimal solution; if it is too

$(\mu_0)_l$	α	$\|g^+\|$	$\|\hat{g}^+\|$	$U(x_k)$	$U(\hat{x}_k)$
0	**0.01**	**0/2.4749**	**0.3227**	**0.9459/1.7530**	**1.2455**
	0.02	0/10.981	0.2611	0.7686/2.5696	1.0410
	0.03	0/11.2701	0.2046	0/2.5942	0.8203
0.1	**0.01**	**0/2.4740**	**0.1194**	**0.9459/1.7529**	**1.2366**
	0.02	0/10.5865	0.0375	0.7488/2.5223	1.0142
	0.03	0/11.855	0.0467	0/2.6672	0.8108
0.17	**0.01**	**0/2.4763**	**0**	**0.9459/1.7532**	**1.2004**
	0.02	0/10.9809	0	0.7686/2.5696	1.0181
	0.03	0/11.8550	0	0/2.6672	0.8392
0.2	**0.01**	**0/2.4034**	**0**	**0.9459/1.7419**	**1.1953**
	0.02	0/10.5865	0	0.7488/2.5223	0.9889
	0.03	0/10.956	0	0/2.5475	0.7893

Table 4.2: Simulation results of Example 4.3(a), in the DS and DSA schemes.

large, the DS system is more likely to be unstable and the amount of oscillation is in proportion of α. As seen from the table, even at small $\alpha = 0.01$, instability exists (the quantity a/b implies it oscillates between the two values) for x_k.

- μ_0 can control the infeasibility (congestion level) of $\hat{x}_k$, as proved in Proposition 2. At the same α, increase of μ_0 can reduce the congestion level, at possible loss of accuracy in the primal cost. For instance, at $\alpha = 0.01$, when $(\mu_0)_l$ increases from 0 to 0.17, infeasibility is eliminated and the estimated primal cost 1.2004 is very close to the global optimal 1.2081. However, μ_0 has no influence on x_k.

- As number of iterations N increases, amount of infeasibility $\|g(\hat{x}_k)^+\|$ decreases, similar to concave NUM in DSA.

Table 4.3 shows the simulation results in SR/SRA. As congestion control is the first criteria to be satisfied under SR, infeasibility of the estimated primal solutions is always reduced to zero. Besides the system variables α, μ_0 and N as in DS/DSA, there are two other uncertain parameters in SR, i.e. threshold of tolerance δ and th. We set $\delta = 0$ throughout the experiment. We make the following observations on the effect of th in SR/SRA:

- th determines how many consecutive iterations of negative utility it allows before shutting down the suer. If it is too small, a user (at its transient phase) may be switched off wrongly; if it is too large, it may take very long to converge or even oscillate without converging. As the appropriate value of th is rather problem-specific, the selection of th adds more uncertainty to the system.

- In this example, if $th < 7$ (out of 50 iterations), the system first switches off user 1 and then user 2, which leads to a wrong solution; however, when $th \geq 7$, SR becomes oscillatory (x_2 alternates between two values) while SRA can stabilize the iterates. It may be deceiving to see the optimal values of x_N on the bold rows of Table 4.3. A close look on the actual iterates during simulation reveals that the system is oscillating between correct and wrong points, as shown in Figure 4-8.

$(\mu_0)_l$	α	th	$x_N, Oscillate?$	$\hat{x}_N$	$U(x_N)$	$U(\hat{x}_N)$
0	0.01	5	(0,0,8), –	(0.1582,1.0534,7.8418)	0.9459	0.9516
		10	(0,0,8), –	(0.1582,1.8076,7.8418)	0.9459	0.9721
	0.02	5	(0,0,7.9999), –	(0.1591,1.192,7.8409)	0.9516	0.9543
		10	**(0,4,7.9999), O**	**(0.2709,3.7281,7.7291)**	**1.2081**	**1.1463**
0.1	0.01	5	(0,0,8), –	(0,0.9019,7.8903)	0.9459	0.9503
		10	(0,0,8), –	(0,1.1614,7.8903)	0.9459	0.9679
	0.02	5	(0,0,8), –	(0.122,1.0499,7.8231)	0.9459	0.9503
		10	**(0,0,8), O**	**(0.2413,3.7587,7.7017)**	**0.9459**	**1.1498**
0.17	0.01	5	(0,0,8), –	0,0.7564,7.7503)	0.9459	0.9407
		10	(0,0,8), –	(0,1.5086,7.7503)	0.9459	0.9561
	0.02	5	(0,0,7.9999), –	(0,1.5503,7.8751)	0.9459	0.9640
		10	**(0,4,7.9999), O**	**(0.1249,3.8751,7.7496)**	**1.2081**	**1.1725**

Table 4.3: Simulation results of Example 4.3(a), in the SR and SRA schemes.

Simulation results in this example show that DS/DSA has a better performance than SR/SRA, as for the latter, the system either converges to a wrong solution (more than enough users are shut down) or oscillates without converging. It can be explained by the discontinuity of x_μ at μ_{max} of Figure 4-6. In SR, as $x_1(\mu_{max})$ violates in both constraints, the system shuts down x_1 first. However, as x can take any value

from 0 to 6 at $\mu_{max} = 0.2$, it results in oscillation for the remaining users: when th is small, x_2 may be shut down wrongly; when th is large, x_2 oscillates between 0 and 6 without converging. As such, $x_2 = 4$ can never be reached in a deterministic manner. In SRA, the estimates make use of the past information of iterates and evolve gradually, which provides a more trustful estimate in most cases.

This example reveals some advantages of employing the averaging method:

- It can stabilize the oscillations in DS and SR, and utilize the past information to generate a good estimate.
- Feasibility violation can be readily controlled by μ_0, at a possible loss of the cost approximation.
- It is robust to the change of stepsize; in other words, it works fairly well in a large range of α.
- As the update of values at each iteration is gradual, it is less prone to make mistakes than SR which may shut down a user wrongly.

(b) $c = [8; 7]$. The global optimal is

$$x^* = [0; 8; 7], \qquad U^* = 1.82.$$

Table 4.4 shows the simulation results at $\alpha = 0.01$. Due to the self-regulation property in SR, x_1 is turned off after a few iterations in order to suppress congestion. In contrast, feasibility of $\hat{x}_k$ in DSA is ensured by setting $\mu_0 = 0.2$ at the initialization of experiment. In this case, SR can provide the best approximate primal solution, while the primal sequences in DSA and SRA converge with a certain level of error.

(c) $c = [14; 15]$. The global optimal is

$$x^* = [6.7705; 7.2295; 8.2295], \qquad U^* = 2.6992.$$

The duality gap in this problem is zero even though all the local utilities are nonconcave, since the primal solution $x(\mu)$ at the optimal is continuous (cf. Figure 4-6) and it satisfies the continuity condition of Proposition 4.2. Thus, SR has the same

	DSA	SR	SRA
x_1	0.6099	0	0.3914
x_2	7.1331	7.9999	7.5971
x_3	6.2687	6.9926	6.6086
U_N	1.6668	1.8192	1.7537

Table 4.4: Simulation results of Example 4.3(b). All the approximate primal optimal solutions are feasible.

outputs as DS; and at an appropriate stepsize (e.g. $\alpha = 0.01$), DS/DSA and SR/SRA can converge to the optimum/near-optimum.

	DS	DSA	SR	SRA
$\alpha = 0.01$	–	$\mu_0 = 0.1$	$th = 8$	$\mu_0 = 0.1$
x_1	6.7705	6.7345	6.7705	6.7345
x_2	7.2295	7.2409	7.2295	7.2409
x_3	8.2295	8.1387	8.2295	8.1387
U_N	2.6992	2.6922	2.6992	2.6922

Table 4.5: Simulation results of Example 4.3(c), a nonconcave NUM with zero duality gap.

Example 4.4 (Heterogenous system; linear topology; sigmoidal and quadratic utilities)

Now we consider a nonconcave objective function combined with the sigmoidal and quadratic functions, in which the users at each link have different utilities. The problem is formulated below with link capacity $c = [1; 2]$.

$$\begin{aligned} \max \quad & S2(x_1) + Q(x_2) + Q(x_3) \\ s.t. \quad & x_1 + x_2 \leq 1 \\ & x_1 + x_3 \leq 2 \\ & 0 \leq x_1, x_2, x_3 \leq 2 \end{aligned}$$

The global optimal is

$$x^* = [0; 1; 2], \qquad U^* = 5.$$

Table 4.6 compares the performance of approximate primal solutions in different schemes. The primal iterates are illustrated in Figure 4-9 (we set $\alpha = 0.1$ in all test instances). We vary the parameters until we obtain a converged solution or a typical output. Note that the convergence of $\{\hat{x}_k\}$ is always guaranteed (cf. Proposition 4.1). We make the following observations:

- In DSA, when $\mu_0 = 0$, the amount of constraint violation is large and it seems to decrease as α increases. To eliminate congestion at any given α, we increase the initial value of dual iterate (cf. Proposition 3.2). When each component of μ_0 is 2, infeasibility is reduced to zero and the value of primal function is optimal. Thus, the DSA scheme can reach the global optimal for this nonconvex problem.

- In SR and SRA, feasibility always holds true. In SR, user 1 is first turned off, since its maximum willingness to pay is $\mu_{max} = 0.25$, at which x_2 and x_3 are 2 (see Table 4.1). As x_1 is on both links, its allocated data rate must be set to zero to avoid system congestion. On link 2, x_3 can take the rate of 2 which is the link capacity, i.e. $\mu_{l=2} = \mu_{max} = 2$. On link 1, however, it is hard to obtain $x_2 = 1$ due to the discontinuity (x can take any value of $0 - 2$ at the same price $\mu = 2$). If th is small, it may switch off user 2 completely; if th is large, the system becomes oscillatory without converging.

- SR can provide suboptimal solutions for sigmoidal utilities with only one inflection point, thus it goes wrong when a quadratic function is present. If we apply averaging to it, the system instability can be alleviated, and the approximate primal solution has a better performance in SRA.

Note that μ_0 does not have any established relation with the feasibility of SR; in contrast, if we increase $(\mu_0)_l$ in DSA or SRA, constraint violation can be decreased continuously.

As seen from these examples, the averaging method can usually provide an effective remedy for the system instability that prevails in nonconvex optimization

	DSA, $(\mu_0)_l = 0$	DSA, $(\mu_0)_l = 2$	SR, $(\mu_0)_l = 0$	SRA, $(\mu_0)_l = 0$
α	$\|\hat{g}^+\|$, $\hat{U}/U^*$	$\|\hat{g}^+\|$, $\hat{U}/U^*$	$\hat{U}/U^*$	$\hat{U}/U^*$
0.01	1.2157, 1.6082	0, 1	1/0.8	0.9768
0.03	1.0830, 1.6032	0, 1	1/0.8	0.9720
0.05	0.8010, 1.4211	0, 1	0.8/1	0.9535
0.07	0.5614, 1.2637	0, 1	0.8/1	0.9635
0.09	0.4418, 1.1936	0, 1	0.8/1	0.9623
0.1	0.4020, 1.1715	0, 1	0.8/1	0.9611
0.2	0.2040, 1.0707	0, 1	1/0.8	0.9569
0.4	0.1265, 1.0349	0, 1	0.8/1	0.9553
0.6	0.0894, 1.0179	0, 1	1/0.8	0.9535
0.8	0.0566, 1.0016	0, 1	0.8/1	0.9535

Table 4.6: Simulation results of Example 4.4. Note that $\|\hat{g}^+\|$ denotes $\|g(\hat{x}_k)^+\|$ and $\hat{U}$ denotes $U(\hat{x}_k)$.

problems.

Example 4.5 (Heterogenous system; tree topology; inelasti/elastic flows)

(a) Consider the tree topology with 4 users and 4 links in Figure 2-2. The nonconcave objective function consists of 4 inelastic flows,

$$\begin{aligned} \max \quad & S1(x_1) + S2(x_2) + S3(x_3) + S4(x_2) \\ s.t. \quad & x_1 + x_2 + x_3 \leq 5 \\ & x_1 + x_2 \leq 4 \\ & x_3 + x_4 \leq 3 \\ & x_2 + x_4 \leq 1 \\ & 0 \leq x_1, x_2, x_3, x_4 \leq 5. \end{aligned}$$

The global optimal is

$$x^* = [4; 0; 1; 1], \qquad U^* = 0.4939.$$

We test it in the DSA, SR and SRA schemes. Table 4.7 presents the numerical results from DSA scheme. Figure 4-10 illustrates the typical outputs (approximate data flows) in these schemes. We make the following observations:

N	μ_0	$\|g(\hat{x}_N)^+\|$	$U(\hat{x}_N)$	$U(\hat{x}_N)/U^*$	$\hat{x}_N$			
50	0	0.7010	0.5964	1.2075	3.7000	0.4166	0.8000	1.2746
	0.25	0	0.3695	0.7481	3.3000	0.0816	0.7000	0.8959
100	0	0.3574	0.5442	1.1018	3.8500	0.2083	0.9000	1.1443
	0.25	0	0.4267	0.8639	3.6500	0.0408	0.8500	0.9517
200	0	0.1787	0.5183	1.0494	3.9250	0.1042	0.9500	1.0721
	0.25	0	0.4588	0.9289	3.8250	0.0204	0.9250	0.9759
500	0	0.0715	0.5035	1.0194	3.9700	0.0417	0.9800	1.0289
	0.25	0	0.4795	0.9708	3.9300	0.0082	0.9700	0.9903

Table 4.7: Simulation results of Example 4.5(a) in the DSA scheme, with fix $\alpha = 0.007$.

- In DSA, constraint violation reduces as the number of iterations increases [cf. Proposition 4.1(a)], where the amount of infeasibility decreases at the order of $1/k$. When we increase μ_0 to 0.25, all the primal iterates are feasible. We observe from the table that DSA can provide a feasible and near-optimal $\hat{x}_k$ (e.g., relative error of $f(\hat{x}_k)$ is less than 3% in the last test instance).

- In this problem, oscillations always exist in the SR system, regardless of α. This can be explained by the discontinuity of $x(\mu)$ at optimal solution. In SR, it first correctly shuts down x_2; however, for the remaining three users, the optimality is achieved at $(x_1^*, x_3^*, x_4^*) = (4, 1, 1)$, with each component falling within the range of discontinuity (cf. Table 4.1). Thus, SR cannot get the exact value of x^* in the dual subgradient method, which results in system instability. SRA can lessen the oscillation but cannot guarantee a "good" approximation of primal solution.

- DSA consistently performs better than SR in such situations when any component of the optimal solution falls within the discontinuity region of $x(\mu)$.

(b) Now consider a system supporting both elastic and inelastic flows. Under the same network topology, users 1 and 3 have the sigmoidal utility functions S1 and S3, respectively. Users 2 and 4 have logarithmic functions with the form of $U(x) = c[\log(ax + b) + d]$, where $d = -\log b$ and $c = \frac{1}{log(a+b)+d}$ under normalization. We assign $(a, b) = (20, 1)$ to user 1 and $(a, b) = (10, 1)$ to user 3. The link capacity is

	DSA	SR		SRA	
α	0.01	0.01	0.001	0.01	0.001
μ_0	0.13	0.13	0	0.13	0
x_N	6.7720	7.2972/11.2598	10.9384	10.7357	10.0589
	1.7524	1.3726/2.6138	3.0560	2.6184	2.8544
	6.4589	0	0	0.5093	1.0688
	5.1993	4.0188/4.3862	3.9440	4.3528	4.1456
$U(x_N)$	3.8572	3.6264/3.8834	3.8898	3.8798	3.8851

Table 4.8: Performance comparison of Example 4.5(b), for the mixed traffic flow problem.

$c = [15; 14; 13; 7]$.

We evaluate the performance of each scheme in generating feasible and near-optimal primal solutions. In each test, we run the simulation for $N = 100$ iterations. Typical results are summarized in Table 4.8. We make the following observations:

- In DSA, at $\alpha = 0.01$ and $\mu_0 = 0$, the convergence of data flow is smooth but the amount of constraint violation is as high as 0.1497 at $N = 100$. Thus, we increase μ_0 to 0.13 while keeping the same stepsize, and infeasibility of all the primal iterates is diminished.

- In SR, at $\alpha = 0.01$ and $\mu_0 = 0.13$ in which DSA converges, it firstly turns down user 3 under self-regulation property; however, the remaining users cause system oscillation and instability. Only when the stepsize is decreased to a very small stepsize $\alpha = 0.001$ (and $\mu_0 = 0$), SR converges to a better primal cost (3.8898) than DSA (3.8572), as shown in Figure 4-11.

- In SRA, we apply the averaging method to the SR iterates and the oscillation can be smoothed out. See Figure 4-12.

From the discussions, we can say that for general nonconcave NUM problems, the averaging scheme can consistently provide good approximations of the primal solution with much less stringent requirements of system parameters than other schemes.

4.2.4 Other Issues

An interesting observation in the nonconvex problems is that there may be multiple solutions for the same suboptimal/optimal value [e.g. in Example 4.5(b), different schemes converge to different (feasible) primal solutions with very close primal cost, as seen in Table 4.8]. This leads to the issue of fairness, which is an important measurement for quality of service in multiuser environment of network utilization (see Korilis and Lazar [15] or Mazumdar et al. [20]). Some of its characteristics are:

- No user is denied access to the network.
- Any user is entitled to as much network utilization as any other user.
- On the onset of congestion, users that heavily load the network should be restricted more than the users that do not overload it.

Regarding the models in solving general optimization problems, if we take the factor of fairness into consideration, self-regulation scheme is not a good candidate in providing fair services. In SR, congestion control is achieved by shutting down users that cause instability of the system, so as to reach a suboptimal solution. By assigning zero data rate to certain users in the network, the system may achieve a higher overall utility than otherwise; however, it does not satisfy the requirement of fairness to keep every user in the network and is intuitively "unfair" to the user that is turned off.

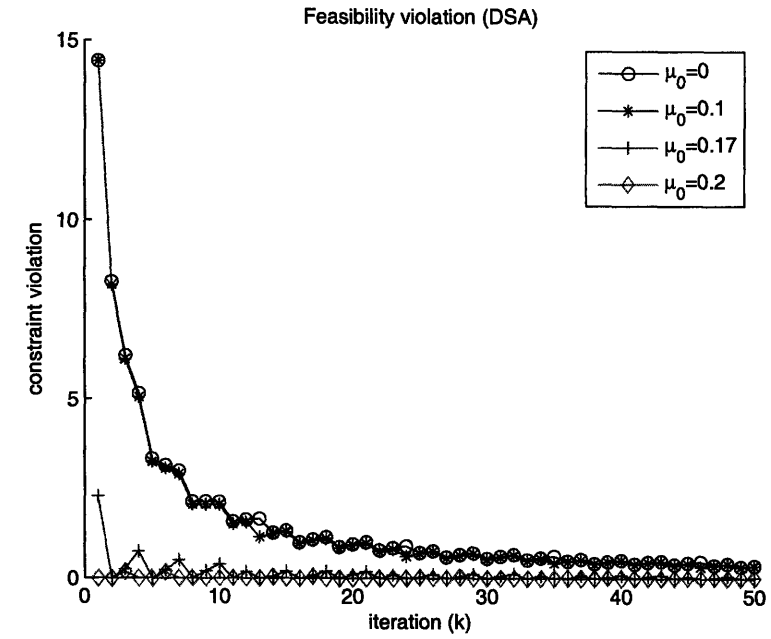
On the other hand, the averaging scheme (e.g. DSA, SRA) ensures the user with positive data allocation if it is not zero from the start of simulation. As such, the averaging scheme is "fairer" in the recovery of primal solutions.

Thus, in network flow control framework, the high-level goal can be defined as "the regulation of input traffic in a way that maintains an efficient tradeoff between high network utilization and quality of service". Nevertheless, fairness cannot be easily quantified. Hence, we will not go into details on this issue in the thesis.

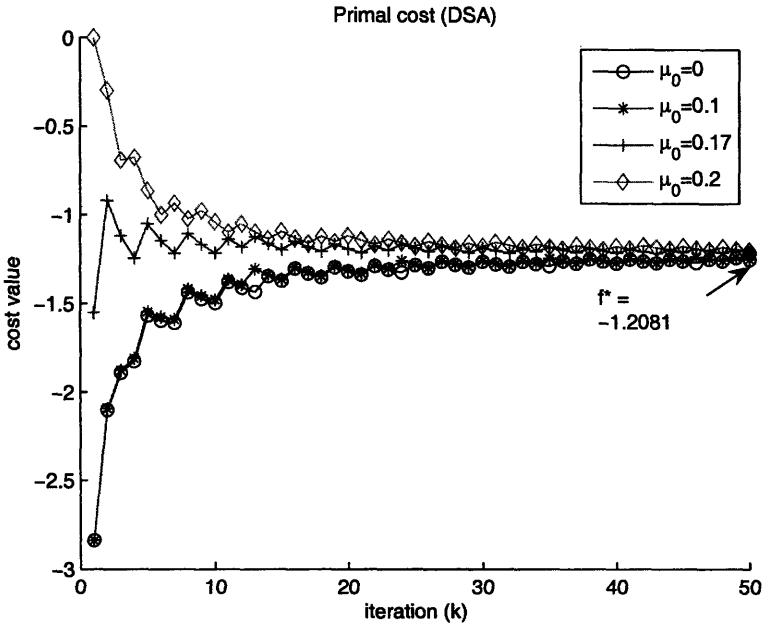
4.3 Summary

In this chapter, we have developed some convergence properties of the averaging scheme (DSA) for nonconvex optimization problems, by extending the results in the convex case. In particular, we have shown theoretically that the convergence of the approximate primal sequence is always guaranteed. We have further explored its feasibility and optimality in the presence of duality gap, and we have provided bounds for the value of primal cost.

Moreover, we have numerically tested the performance of DSA and other schemes in nonconcave NUM. Such systems with sigmoidal or quadratic utility functions typically exhibit oscillatory behavior and excessive congestion. The simulation results in various test instances show that, the averaging scheme can consistently provide near-feasible and near-optimal primal solutions while other approaches may easily fail.

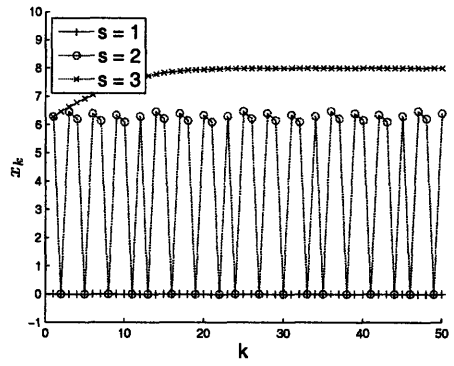


(a) $\|g(\hat{x}_k)^+\|$ as different μ_0.

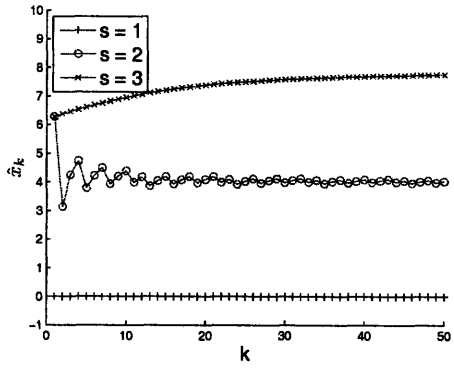


(b) $f(\hat{x}_k)$ at different μ_0.

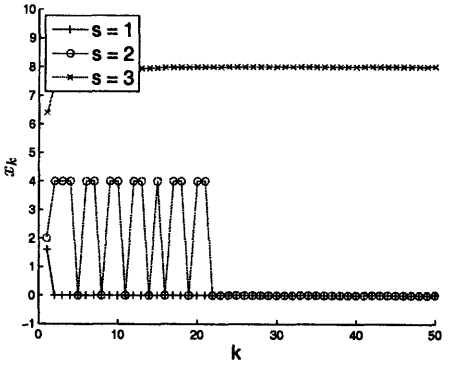
Figure 4-7: Trend of convergence for feasibility violations and primal cost of $\hat{x}_k$ in Example 4.3(a), with fixed $\alpha = 0.01$ and increasing μ_0.



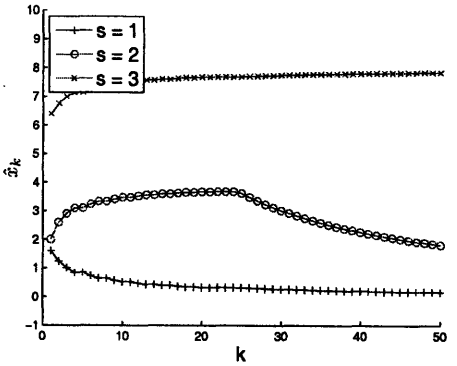
(a) (DS) $\alpha = 0.01$, $\mu_0 = 0.17$.



(b) (DSA) $\alpha = 0.01$, $\mu_0 = 0.17$.

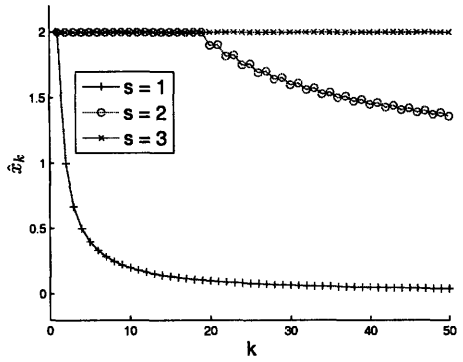


(c) (SR) $\alpha = 0.1$, $th = 10$, $\mu_0 = 0$.

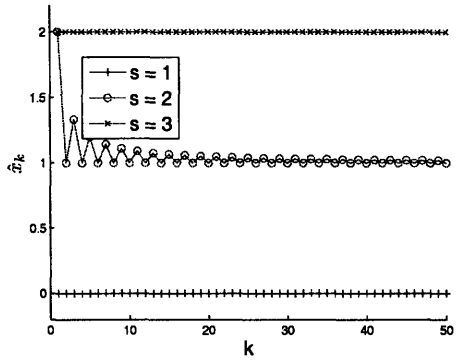


(d) (SRA) $\alpha = 0.1$, $th = 10$, $\mu_0 = 0$.

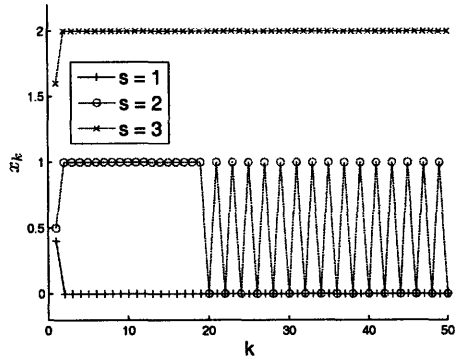
Figure 4-8: Comparison of approximate primal solutions (data flows) in DS/DSA and SR/SRA schemes, with typical values of parameters in Example 4.3(a).



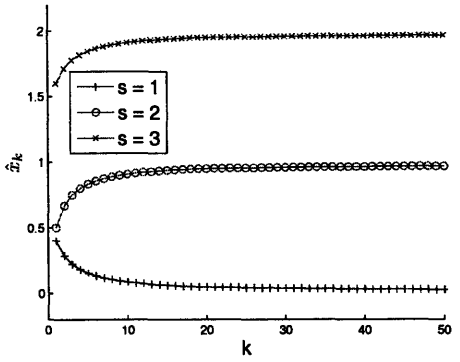
(a) (DSA) $\alpha = 0.1$, $\mu_0 = 0$. The solution is infeasible.



(b) (DSA) $\alpha = 0.1$, $\mu_0 = 2$.

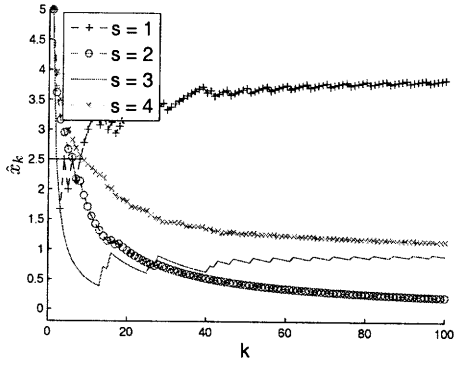


(c) (SR) $\alpha = 0.1$, $\mu_0 = 0$, $th = 5$.

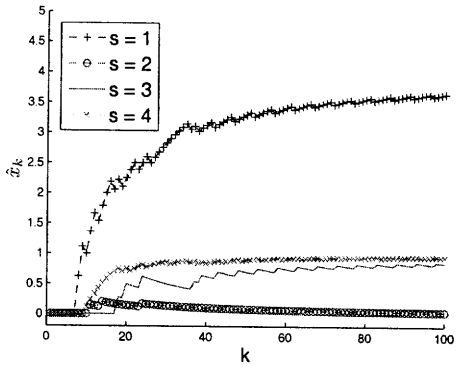


(d) (SRA) $\alpha = 0.1$, $\mu_0 = 0$, $th = 5$.

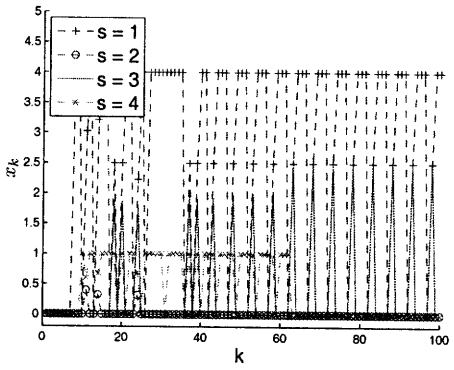
Figure 4-9: Comparison of approximate primal solutions (data flows) in DSA and SR/SRA schemes, with typical values of parameters in Example 4.4. Solutions are feasible unless otherwise stated.



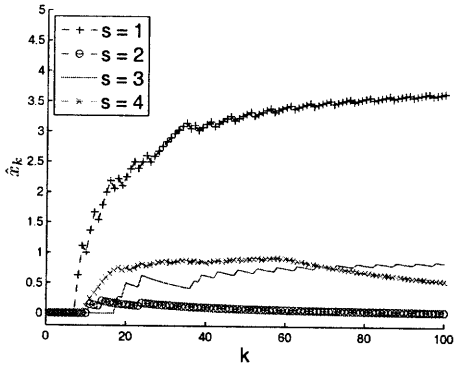
(a) (DSA) $\alpha = 0.007$, $\mu_0 = 0$. The solution is infeasible.



(b) (SA) $\alpha = 0.007$, $\mu_0 = 0.25$.

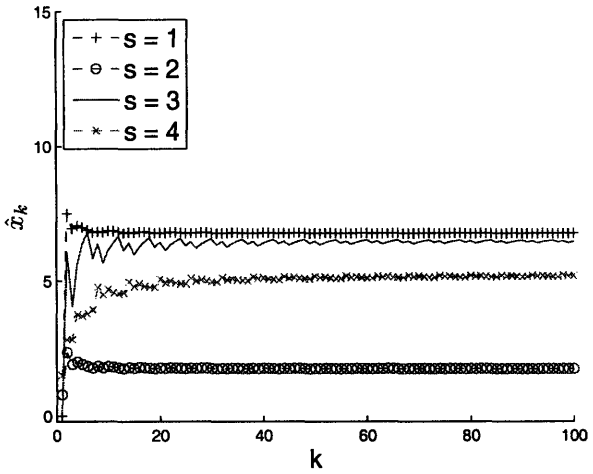


(c) (SR) $\alpha = 0.007$, $\mu_0 = 0.25$, $th = 30$.

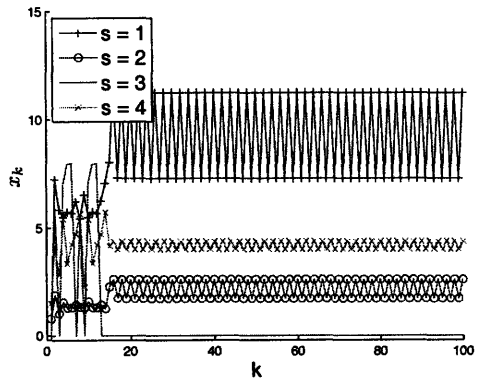


(d) (SRA) $\alpha = 0.007$, $\mu_0 = 0.25$, $th = 30$.

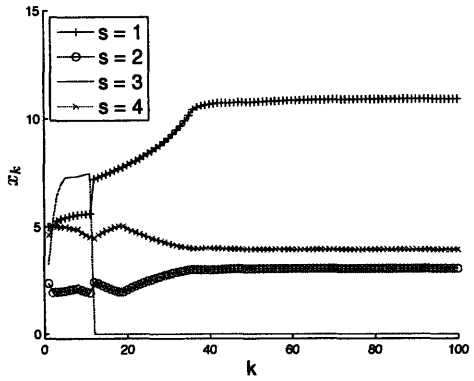
Figure 4-10: Comparison of approximate primal solutions in DSA and SR/SRA schemes, with typical values of parameters in Example 4.5(a). Solutions are feasible unless otherwise stated.



(a) (DSA) $\alpha = 0.01$, $\mu_0 = 0.13$.

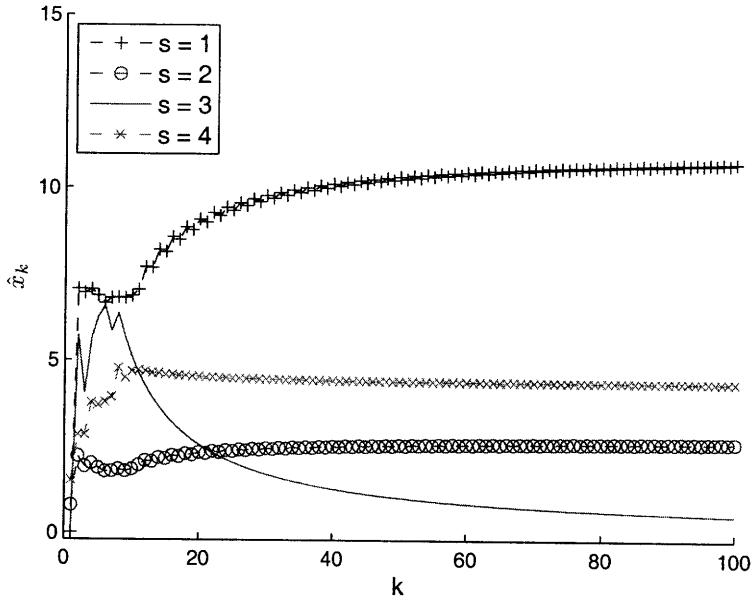


(b) (SR) $\alpha = 0.01$, $\mu_0 = 0.13$.

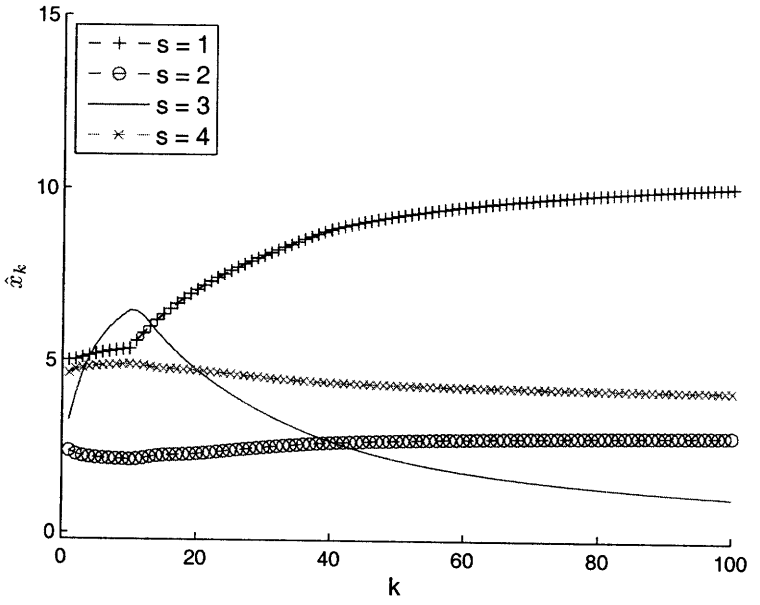


(c) (SR) $\alpha = 0.001$, $\mu_0 = 0$.

Figure 4-11: Comparison of approximate primal solutions in DSA and SR schemes, with typical values of parameters in Example 4.5(b). $th = 5$ in SR.



(a) (SRA) $\alpha = 0.01$, $\mu_0 = 0.13$.



(b) (SRA) $\alpha = 0.001$, $\mu_0 = 0$.

Figure 4-12: Approximate primal solutions the SRA scheme in Example 4.5(b). Compare with Figures 4-11(b) and 4-11(c).

Chapter 5

Conclusions

In this thesis, we have investigated a primal-averaging scheme in the application of dual subgradient methods to recover near-feasible and near-optimal primal solutions, for general optimization problems. We have chosen the constant stepsize rule in view of its simplicity and practical significance. Based on the convergence properties of approximate primal solutions that are theoretically analyzed by Nedić and Ozdaglar [22] for convex problems, we have further developed a way to effectively reduce the amount of feasibility violation in approximate primal solutions. In particular, this can be done by controlling the value of initial dual iterate, which is a nice property that the averaging scheme possesses. Moreover, to complement the theoretical analysis, we have implemented a distributive algorithm under the averaging scheme for structured network utility maximization (NUM) problems, and computationally tested the feasibility and optimality of the averaged primal solutions. We have also discussed the effects of different system parameters, namely the stepsize value and initial dual iterate, upon the tradeoffs in solution accuracy, computational complexity and convergence trends. It has shown that the averaging scheme is more robust to the change of parameters and has a better performance than the dual subgradient scheme without averaging, which easily causes system instability.

From the promising numerical results and theoretical insights in the convex case, we have extended our analysis to problems with nonconvex objective functions. We have provided modified bounds on the value of primal cost when the duality gap is

positive. In the nonconcave NUM framework where the network congestion prevails, we have specifically estimated bounds of duality gaps for sigmoidal and quadratic utilities. We have evaluated and compared the performance of averaging scheme with other models (in particular, the self-regulation heuristic) in various test instances of nonconcave NUM. From the simulation results, we have seen that the averaging scheme can consistently generate good estimations of the primal optimal solution in most situations.

In conclusion, we believe that the averaging scheme holds promise as a more robust and tractable technique to introduce convergence properties and recover primal optimal solutions compared to other existing schemes.Its ability at infeasibility reduction is particularly attractive in the nonconvex optimization, where congestion always exists due to discontinuity of primal solution at the optimal dual. For future work, we can further explore the far-reaching benefits in the averaging scheme, such as convergence analysis of the primal sequence itself, other practical applications than NUM and optimization of Slater point in the prediction of the theoretical bound.

Bibliography

[1] J.P. Aubin and I. Ekeland, *Estimates of the duality gap in nonconvex optimization*, The Institute of Management Sciences, 1976.

[2] D.P. Bertsekas, *Nonlinear programming*, Athena Scientific, Cambridge, Massachusetts, 1999.

[3] A. Ben-Tal, and A. Nemirovski, *Optimization I-II: Convex Analysis, Nonlinear Programming Theory, Standard Nonlinear Programming Algorithms*, Lecture Notes.

[4] D.P. Bertsekas, A. Nedić, and A.E. Ozdaglar, *Convex analysis and optimization*, Athena Scientific, Cambridge, Massachusetts, 2003.

[5] S. Boyd, L. Xiao and A. Mutapcic, *Subgradient Methods*, Notes for EE392o, Stanford University, Autumn 2003.

[6] L. Chen, S. Low, M. Chiang, and J. Doyle, *Optimal cross-layer congestion control, routing and scheduling design in ad hoc wireless networks*, IEEE InfoCom, 2006.

[7] M. Chiang, *Nonconvex optimization of communication systems*, Advances in Mechanics and Mathematics, Ed., D. Gao and H. Sherali, Springer, August 2007.

[8] M. Chiang, S.H. Low, A.R. Calderbank, and J.C. Doyle, *Layering as optimization decomposition: A mathematical theory of network architectures*, Proc.of the IEEE, Vol. 95, No. 1, pp. 255-312, January 2007.

[9] M. Chiang, S. Zhang and P. Hande, *Distributed rate allocation for inelastic flows: optimization frameworks, optimality conditions, and optimal algorithms*, Proc. IEEE INFOCOM, Vol.4, 2679-2690, 2005.

[10] M. Dür, *A class of problems where dual bounds beat underestimation bounds*, Journal of Global Optimization 22 (2002), Issue 1-4, 49-57.

[11] M. Fazel and M. Chiang, *Network utility maximization with nonconcave utilities using sum-of-squares method*, in Proc. IEEE Conf. Dec. Contr., 2005, 1867-1874.

[12] M. Fazel, M. Chiang, and J. Doyle, *Network rate allocation by nonconcave utility maximizaiton*, Preprint, 2005.

[13] R. Horst, and H. Tuy, *Global Optimization*, Springer, Berlin, 1996.

[14] F.P. Kelly, A. Maulloo, and D. Tan, *Rate control for communication networks: shadow prices, proportional fairness and stability*, Journal of Operations Resaerch Society, Vol.49, No.3, 237-252, 1998.

[15] Y.A. Korilis, and A.A. Lazar, *Why is Flow Control Hard: Optimality, Fairness, Partial and Delayed Information*, Proc. 2nd ORSA Telecommunications Conference, March 1992.

[16] T. Larsson and Z. Liu, *A Lagrangian relaxation scheme for structured linear programs with application to multicommodity network flows*, Optimization 40 (1997), 247-284.

[17] T. Larsson, M. Patriksson, and A. Strömberg, *Ergodic, primal convergence in dual subgradient schemes for convex programming*, Mathematical Programming 86 (1999), 283-312.

[18] J.W. Lee, R.R. Mazumdar and N.B. Shroff, *Non-convex optimization and rate control for multi-class services in the Internet*, IEEE/ACM Transactions on Networking (TON), Vol.13, Issue 4, 2005, 827-840.

[19] S. Low and D.E. Lapsley, *Optimization flow control, I: Basic algorithm and convergence*, IEEE Journal on Selected Areas in Communication, August 2006, vol.24, no.8, 1439-1451.

[20] R. Mazumdar, L.G. Mason, and C. Douligeris, *Fairness in Network Optimization Control: Optimality of Product Forms*, IEEE Transactions on Communications, Vol. 39, No. 5, pp. 775-782, May 1991.

[21] A. Nedić, D.P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, Decision and Control, Proc. of the 38th IEEE Conference, Vol. 1, pp. 907-912, 1999.

[22] A. Nedić and A.E. Ozdaglar, *Approximate Primal Solutions and Rate Analysis for Dual Subgradient Methods*, LIDS Technical Report 2753, Massachusetts Institute of Technology, Lab. for information and Decision Systems, 2007.

[23] D.P. Palomar and M. Chiang, *A Tutorial on Decomposition Methods for Network Utility Maximization*, IEEE/ACM Transactions on Networking 7 (1999), no.6, 861-874.

[24] N.Z. Shor, *Minimization methods for nondifferentiable functions*, Springer, Berlin, 1985, 116-118.

[25] S. Shenker, *Fundamental design issues for the future internet*, IEEE J.Sel.Area Comm., Vol. 13, No. 7, pp. 1176-1188, September 1995.

[26] H.D. Sherali and G. Choi, *Recovery of primal solutions when using subgradient optimization methods to solve Lagrangian duals of linear programs*, Operations Research Letters 19 (1996), 105-113.

[27] J.T. Wang, L. Li, S.H. Lo and J.C. Doyle, *Cross-layer optimization in TCP/IP networks*, IEEE/ACM Transactions on Networking (TON), Vol. 13, Iss. 3, pp. 582-595, June 2005.