Approximate Primal Solutions and Rate Analysis for Dual Subgradient Methods

Angelia Nedich
Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign
117 Transportation Building
104 S. Mathews Avenue
Urbana, IL 61853
E-mail: angelia@uiuc.edu
Phone: 217 244-1756
Fax: 217 244-5705

Asuman Ozdaglar
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
77 Massachusetts Avenue, 32-D630
Cambridge, MA 02139
E-mail: asuman@mit.edu
Phone: 617 324-0058
Fax: 617 253-3578

Approximate Primal Solutions and Rate Analysis for Dual Subgradient Methods*

Angelia Nedić[†] and Asuman Ozdaglar[‡]

November 13, 2007

Abstract

We study primal solutions obtained as a by-product of subgradient methods when solving the Lagrangian dual of a primal convex constrained optimization problem (possibly nonsmooth). The existing literature on the use of subgradient methods for generating primal optimal solutions is limited to the methods producing such solutions only asymptotically (i.e., in the limit as the number of subgradient iterations increases to infinity). Furthermore, no convergence rate results are known for these algorithms.

In this paper, we propose and analyze dual subgradient methods using averaging to generate approximate primal optimal solutions. These algorithms use a constant stepsize as opposed to a diminishing stepsize which is dominantly used in the existing primal recovery schemes. We provide estimates on the convergence rate of the primal sequences. In particular, we provide bounds on the amount of feasibility violation of the generated approximate primal solutions. We also provide upper and lower bounds on the primal function values at the approximate solutions. The feasibility violation and primal value estimates are given per iteration, thus providing practical stopping criteria. Our analysis relies on the Slater condition and the inherited boundedness properties of the dual problem under this condition.

Keywords: subgradient methods, averaging, approximate primal solutions, convergence rate.

*We would like to thank Robert Freund and Pablo Parrilo for useful comments and discussions.

†Department of Industrial and Enterprise Systems Engineering, University of Illinois, angelia@uiuc.edu

‡Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, asuman@mit.edu

1 Introduction

Lagrangian relaxation and duality have been effective tools for solving large-scale convex optimization problems and for systematically providing lower bounds on the optimal value of nonconvex (continuous and discrete) optimization problems. Subgradient methods have played a key role in this framework providing computationally efficient means to obtain near-optimal dual solutions and bounds on the optimal value of the original optimization problem. Most remarkably, in networking applications, over the last few years, subgradient methods have been used with great success in developing decentralized cross-layer resource allocation mechanisms (see Low and Lapsley [15], and Srikant [27] for more on this subject).

The subgradient methods for solving dual problems have been extensively studied starting with Polyak [21] and Ermoliev [7]. Their convergence properties under various stepsize rules have been long established and can be found, for example, in Shor [26], Demyanov and Vasilev [6], Polyak [22], Hiriart-Urruty and Lemaréchal [8], Bertsekas [3], and Bertsekas, Nedić, and Ozdaglar [4]. Numerous extensions and implementations including parallel and incremental versions have been proposed and analyzed (for example, see Kiwiel and Lindberg [10], Zhao, Luh and Wang [29], Ben-Tal, Margalit and Nemirovski [2], Nedić and Bertsekas [16], [17], Nedić, Bertsekas and Borkar [18]).

Despite widespread use of the subgradient methods for solving dual (nondifferentiable) problems, there are some aspects of subgradient methods that have not been fully studied. In particular, in practical applications, the main interest is in solving the primal problem. In this case, the question arises whether we can use the subgradient method in dual space and exploit the subgradient information to produce primal near-feasible and near-optimal solutions. This is the issue that we pursue in this paper.

The first primal-dual subgradient scheme using primal-averaging has been investigated in a paper by Nemirovski and Judin [19]. Subsequently, a related primal-averaging scheme based on subgradient information generated by a dual subgradient method have been proposed for linear (primal) problems by Shor [26]. Shor's ideas have been further developed and computationally tested by Larsson and Liu [11] for linear problems. Sherali and Choi [25] have focused on linear optimization problems and extended these results to allow for more general averaging schemes (i.e., more general choices of the weights for convex combinations) and a wider class of stepsize choices. More recently, Larsson, Patriksson, and Strömberg generalized these results in a series of papers (see [12], [13], [14]) to convex constrained optimization problems and demonstrated promising applications of these schemes in the context of traffic equilibrium and road pricing. Sen and Sherali [24] have studied a more complex scheme combining a subgradient method and an auxiliary penalty problem to recover primal solutions. A dual subgradient method producing primal solutions, the volume algorithm, for linear problems have been proposed by Barahona and Anbil [1]. They have reported experimental results for several linear problems including set partitioning, set covering, and max-cut, but have not analyzed convergence properties of the algorithm. Kiwiel, Larsson and Lindberg [9] have studied the convergence of primal-averaging in dual subgradient methods using a target-level based stepsize. Recently, Nesterov [20] has proposed a subgradient algorithm using averaging and provided convergence rate analysis assuming the availability of a bound on the Euclidean norm of an optimal solution. Nesterov's algorithm generates a solution to a convex minimization problem, and it is not a primal-recovery scheme. More recently, Ruszczynski [23] has proposed a new subgradient method that uses averaging to identify both an optimal solution of a convex minimization problem and a subgradient that appears in the optimality condition.

Our work here is related to the primal-recovery algorithms of Shor [26], Sen and Sherali [24], Sherali and Choi [25], Larsson *et. al* [12], [13], [14], and Kiwiel *et. al* [9]. The previous work has several common characteristics: First, the focus has been only on the asymptotic behavior of the primal sequence, i.e., the convergence properties in the limit as the number of subgradient iterations increases to infinity. Second, the convergence analysis has been almost exclusively limited to diminishing stepsize rules (with divergent sum). The exception is the paper [9] where a target-level based stepsize (i.e., a modification of Polyak's stepsize [21]) has been considered. Third, there are no convergence rate results for primal-recovery methods. All of this motivates our work here.

Specifically, our interest is in solving the primal problem approximately by using a simple averaging of primal vectors obtained as a by-product of a dual subgradient method. In this paper, we deal with "approximate primal solutions" as opposed to asymptotically "optimal solutions" studied in the existing literature. We are interested in the constant stepsize rule for dual subgradient algorithms, mainly because of its practical importance and simplicity for implementations. We show how approximate primal solutions can be generated for general (possibly nonsmooth) convex constrained optimization problems under the Slater constraint qualification. We first show that the sequence of dual solutions generated by the subgradient method is bounded under the Slater condition. We use this result in estimating the approximation error of the solutions both in terms of primal feasibility and primal optimality. In particular, we show that the amount of constraint violation for the average primal sequence goes to zero at the rate of $1/k$ with the number of iterations k for the ordinary subgradient method. We also provide per-iterate estimates for the violation of the constraints, and upper and lower bounds on the objective function value of the average primal solution.

We next consider an alternative subgradient method under the Slater condition. This method exploits the boundedness of the dual optimal set by projecting the dual iterates to a bounded superset of the dual optimal set. Thus, the method does not permit the dual iterates to wander too far from the optimal set, which has potential of resulting in fast convergence rate of the method. In this method, also, we use an averaging scheme to generate approximate primal solutions and provide error estimates on the amount of constraint violation and the cost of the average primal solution. We compare the error estimates of the two proposed methods to illustrate the potential advantages.

In summary, the contributions of this paper include:

- The development and analysis of new algorithms producing approximate primal feasible and primal optimal solutions. Unlike the existing literature on primal recovery in dual methods, here, the focus is on the constant stepsize rule in view of its simplicity and practical significance.

- The convergence and convergence rate analysis of the methods under the Slater constraint qualification. The error estimates of the approximate primal solutions are derived, including estimates of the amount of feasibility violation, and upper and lower bounds for the primal objective function. These estimates are per iteration, and can be used as a stopping criteria based on a user specified accuracy.

- The insights into the tradeoffs involved in the selection of the stepsize value. In particular, our convergence rate analysis explicitly illustrates the tradeoffs between the solution accuracy and computational complexity in selecting the stepsize value.

The paper is organized as follows: In Section 2, we define the primal and dual problems, and provide an explicit bound on the level sets of the dual function under Slater condition. In Section 3, we consider a subgradient method with a constant stepsize and study its properties under Slater. In Section 4, we introduce approximate primal solutions generated through averaging and provide bounds on their feasibility violation and primal cost values. In Section 5, we consider an alternative to the basic subgradient method based on the boundedness properties of the dual function under the Slater condition, and we provide error estimates for the generated approximate primal solutions. We conclude in Section 6 by summarizing our work and providing some comments.

2 Primal and Dual Problems

In this section, we formulate the primal and dual problems of interest. We, also, give some preliminary results that we use in the subsequent development. We start by introducing the notation and the basic terminology that we use throughout the paper.

2.1 Notation and Terminology

We consider the n-dimensional vector space $\mathbb{R}^n$ and the m-dimensional vector space $\mathbb{R}^m$. We view a vector as a column vector, and we denote by $x'y$ the inner product of two vectors x and y. We use $\|y\|$ to denote the standard Euclidean norm, $\|y\| = \sqrt{y'y}$. Occasionally, we also use the standard 1-norm and ∞-norm denoted respectively by $\|y\|_1$ and $\|y\|_\infty$, i.e., $\|y\|_1 = \sum_i |y_i|$ and $\|y\|_\infty = \max_i |y_i|$. We write $dist(\bar{y}, Y)$ to denote the standard Euclidean distance of a vector $\bar{y}$ from a set Y, i.e.,

$$dist(\bar{y}, Y) = \inf_{y \in Y} \|\bar{y} - y\|.$$

For a vector $u \in \mathbb{R}^m$, we write u^+ to denote the projection of u on the nonnegative orthant in $\mathbb{R}^m$, i.e., u^+ is the component-wise maximum of the vector u and the zero vector:

$$u^+ = (\max\{0, u_1\}, \cdots, \max\{0, u_m\})' \quad \text{for } u = (u_1, \cdots, u_m)'.$$

For a concave function $q : \mathbb{R}^m \mapsto [-\infty, \infty]$, we denote the domain of q by $\mathrm{dom}(q)$, where

$$\mathrm{dom}(q) = \{\mu \in \mathbb{R}^m \mid q(\mu) > -\infty\}.$$

We use the notion of a subgradient of a concave function $q(\mu)$. In particular, a subgradient $s_{\bar{\mu}}$ of a concave function $q(\mu)$ at a given vector $\bar{\mu} \in \mathrm{dom}(q)$ provides a linear overestimate of the function $q(\mu)$ for all $\mu \in \mathrm{dom}(q)$. We use this as the subgradient defining property: *$s_{\bar{\mu}} \in \mathbb{R}^m$ is a subgradient of a concave function $q(\mu)$ at a given vector $\bar{\mu} \in \mathrm{dom}(q)$* if the following relation holds:

$$q(\bar{\mu}) + s_{\bar{\mu}}'(\mu - \bar{\mu}) \geq q(\mu) \qquad \text{for all } \mu \in \mathrm{dom}(q). \tag{1}$$

The set of all subgradients of q at $\bar{\mu}$ is denoted by $\partial q(\bar{\mu})$.

In this paper, we focus on the following constrained optimization problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq 0 \\ & x \in X, \end{array} \tag{2}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a convex function, $g = (g_1, \ldots, g_m)'$ and each $g_j : \mathbb{R}^n \mapsto \mathbb{R}$ is a convex function, and $X \subset \mathbb{R}^n$ is a nonempty closed convex set. We refer to this as the *primal problem*. We denote the primal optimal value by f^*, and throughout this paper, *we assume that the value f^* is finite.*

To generate approximate solutions to the primal problem of Eq. (2), we consider approximate solutions to its dual problem. Here, the *dual problem* is the one arising from Lagrangian relaxation of the inequality constraints $g(x) \leq 0$, and it is given by

$$\begin{array}{ll} \text{maximize} & q(\mu) \\ \text{subject to} & \mu \geq 0 \\ & \mu \in \mathbb{R}^m, \end{array} \tag{3}$$

where q is the dual function defined by

$$q(\mu) = \inf_{x \in X} \{f(x) + \mu' g(x)\}. \tag{4}$$

We often refer to a vector $\mu \in \mathbb{R}^m$ with $\mu \geq 0$ as a *multiplier*. We denote the dual optimal value by q^* and the dual optimal set by M^*. We say that there is *zero duality gap* if the optimal values of the primal and the dual problems are equal, i.e., $f^* = q^*$.

We assume that the minimization problem associated with the evaluation of the dual function $q(\mu)$ has a solution for every $\mu \geq 0$. This is the case, for instance, when the set X is compact (since f and $g_j's$ are continuous due to being convex over $\mathbb{R}^n$). Furthermore, we assume that the minimization problem in Eq. (4) is simple enough so that it can be solved efficiently. For example, this is the case when the functions f and g_j's are affine or affine plus norm-square term [i.e., $c\|x\|^2 + a'x + b$], and the set X is the nonnegative orthant in $\mathbb{R}^n$. Many practical problems of interest, such as those arising in network optimization, often have this structure.

In our subsequent development, we consider subgradient methods as applied to the dual problem given by Eqs. (3) and (4). Due to the form of the dual function q, the subgradients of q at a vector μ are related to the primal vectors x_μ attaining the minimum in Eq. (4). Specifically, the set $\partial q(\mu)$ of subgradients of q at a given $\mu \geq 0$ is given by

$$\partial q(\mu) = \mathrm{conv}\left(\{g(x_\mu) \mid x_\mu \in X_\mu\}\right), \qquad X_\mu = \{x_\mu \in X \mid q(\mu) = f(x_\mu) + \mu' g(x_\mu)\}, \tag{5}$$

where $\mathrm{conv}(Y)$ denotes the convex hull of a set Y.

2.2 Slater Condition and Boundedness of the Multiplier Sets

In this section, we consider sets of the form $\{\mu \geq 0 \mid q(\mu) \geq q(\bar{\mu})\}$ for a fixed $\bar{\mu} \geq 0$, which are obtained by intersecting the nonnegative orthant in $\mathbb{R}^m$ and (upper) level sets of the concave dual function q. We show that these sets are bounded when the primal problem satisfies the standard Slater constraint qualification, formally given in the following.

Assumption 1 *(Slater Condition)* There exists a vector $\bar{x} \in \mathbb{R}^n$ such that

$$g_j(\bar{x}) < 0 \qquad \text{for all } j = 1, \dots, m.$$

We refer to a vector $\bar{x}$ satisfying the Slater condition as *a Slater vector*.

Under the assumption that f^* is finite, it is well-known that the Slater condition is sufficient for a zero duality gap as well as for the existence of a dual optimal solution (see for example Bertsekas [3] or Bertsekas, Nedić, and Ozdaglar [4]). Furthermore, the dual optimal set is bounded (see Hiriart-Urruty and Lemaréchal [8]). This property of the dual optimal set under the Slater condition, has been observed and used as early as in Uzawa's analysis of Arrow-Hurwicz gradient method in [28]. Interestingly, most work on subgradient methods has not made use of this powerful result, which is a key in our analysis.

The following proposition extends the result on the optimal dual set boundedness under the Slater condition. In particular, it shows that the Slater condition also guarantees the boundedness of the (level) sets $\{\mu \geq 0 \mid q(\mu) \geq q(\bar{\mu})\}$.

Lemma 1 Let the Slater condition hold [cf. Assumption 1] . Let $\bar{\mu} \geq 0$ be a vector such that the set $Q_{\bar{\mu}} = \{\mu \geq 0 \mid q(\mu) \geq q(\bar{\mu})\}$ is nonempty. Then, the set $Q_{\bar{\mu}}$ is bounded and, in particular, we have

$$\max_{\mu \in Q_{\bar{\mu}}} \|\mu\| \leq \frac{1}{\gamma} \left(f(\bar{x}) - q(\bar{\mu}) \right),$$

where $\gamma = \min_{1 \leq j \leq m}\{-g_j(\bar{x})\}$ and $\bar{x}$ is a vector satisfying the Slater condition.

Proof. Let $\mu \in Q_{\bar{\mu}}$ be arbitrary. By the definition of the set $Q_{\bar{\mu}}$, we have for any $\mu \in Q_{\bar{\mu}}$,

$$q(\bar{\mu}) \leq q(\mu) = \inf_{x \in X} \{f(x) + \mu' g(x)\} \leq f(\bar{x}) + \mu' g(\bar{x}) = f(\bar{x}) + \sum_{j=1}^{m} \mu_j g_j(\bar{x}),$$

implying that

$$-\sum_{j=1}^{m} \mu_j g_j(\bar{x}) \leq f(\bar{x}) - q(\bar{\mu}).$$

Because $g_j(\bar{x}) < 0$ and $\mu_j \geq 0$ for all j, it follows that

$$\min_{1\leq j\leq m}\{-g_j(\bar{x})\}\sum_{j=1}^{m}\mu_j \leq -\sum_{j=1}^{m}\mu_j g_j(\bar{x}) \leq f(\bar{x}) - q(\bar{\mu}).$$

Therefore,

$$\sum_{j=1}^{m}\mu_j \leq \frac{f(\bar{x}) - q(\bar{\mu})}{\min_{1\leq j\leq m}\{-g_j(\bar{x})\}}.$$

Since $\mu \geq 0$, we have $\|\mu\| \leq \sum_{j=1}^{m}\mu_j$ and the estimate follows. ■

It follows from the preceding Lemma that under the Slater condition, the dual optimal set M^* is nonempty. In particular, by noting that $M^* = \{\mu \geq 0 \mid q(\mu) \geq q^*\}$ and by using Lemma 1, we see that

$$\max_{\mu^*\in M^*}\|\mu^*\| \leq \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right), \tag{6}$$

with $\gamma = \min_{1\leq j\leq m}\{-g_j(\bar{x})\}$.

In practice, the dual optimal value q^* is not readily available. However, having a dual function value $q(\tilde{\mu})$ for some $\tilde{\mu} \geq 0$, we can still provide a bound on the norm of the dual optimal solutions. In particular, since $q^* \geq q(\tilde{\mu})$, from relation (6) we obtain the following bound:

$$\max_{\mu^*\in M^*}\|\mu^*\| \leq \frac{1}{\gamma}\left(f(\bar{x}) - q(\tilde{\mu})\right).$$

Furthermore, having any multiplier sequence $\{\mu_k\}$, we can use the dual function values $q(\mu_k)$ to generate a sequence of (possibly improving) upper bounds on the dual optimal solution norms $\|\mu^*\|$. Formally, since $q^* \geq \max_{0\leq i\leq k} q(\mu_i)$, from relation (6) we have

$$\max_{\mu^*\in M^*}\|\mu^*\| \leq \frac{1}{\gamma}\left(f(\bar{x}) - \max_{0\leq i\leq k} q(\mu_i)\right) \qquad \text{for all } k \geq 0.$$

Note that these bounds are nonincreasing in k. These bounds have far reaching consequences for they allow us to "locate dual optimal solutions" by using only a Slater vector $\bar{x}$ and a multiplier sequence $\{\mu_k\}$ generated by a subgradient method. All of this is of practical significance.

Such bounds play a key role in our subsequent development. In particular, we use these bounds to provide error estimates of our approximate solutions as well as to design a dual algorithm that projects on a set containing the dual optimal solution.

3 Subgradient Method

To solve the dual problem, we consider the classical subgradient algorithm with a constant stepsize:

$$\mu_{k+1} = [\mu_k + \alpha g_k]^+ \qquad \text{for } k = 0, 1, \ldots, \tag{7}$$

where the vector $\mu_0 \geq 0$ is an initial iterate and the scalar $\alpha > 0$ is a stepsize. The vector g_k is a subgradient of q at μ_k given by

$$g_k = g(x_k), \qquad x_k \in \operatorname*{argmin}_{x \in X}\{f(x) + \mu_k' g(x)\} \qquad \text{for all } k \geq 0 \tag{8}$$

[see Eq. (5)].

One may consider other stepsize rules for the subgradient method. Our choice of the constant stepsize is primarily motivated by its practical importance and in particular, because in practice the stepsize typically stays bounded away from zero. Furthermore, the convergence rate estimates for this stepsize can be explicitly written in terms of the problem parameters that are often available. Also, when implementing a subgradient method with a constant stepsize rule, the stepsize length α is the only parameter that a user has to select, which is often preferred to more complex stepsize choices involving several stepsize parameters without a good guidance on their selection.

3.1 Basic Relations

In this section, we establish some basic relations that hold for a sequence $\{\mu_k\}$ obtained by the subgradient algorithm of Eq. (7). These properties are important in our construction of approximate primal solutions, and in particular, in our analysis of the error estimates of these solutions.

We start with a lemma providing some basic relations that hold under minimal assumptions. The relations given in this lemma have been known and used in various ways to analyze subgradient approaches (for example, see Shor [26], Polyak [22], Demyanov and Vasilev [6], Correa and Lemaréchal [5], Nedić and Bertsekas [16], [17]). The proofs are provided here for completeness.

Lemma 2 *(Basic Iterate Relation)* Let the sequence $\{\mu_k\}$ be generated by the subgradient algorithm (7). We then have:

(a) For any $\mu \geq 0$,

$$\|\mu_{k+1} - \mu\|^2 \leq \|\mu_k - \mu\|^2 - 2\alpha\,(q(\mu) - q(\mu_k)) + \alpha^2\|g_k\|^2 \qquad \text{for all } k \geq 0.$$

(b) When the optimal solution set M^* is nonempty, there holds

$$dist^2(\mu_{k+1}, M^*) \leq dist^2(\mu_k, M^*) - 2\alpha\,(q^* - q(\mu_k)) + \alpha^2\|g_k\|^2 \qquad \text{for all } k \geq 0,$$

where $dist(y, Y)$ denotes the Euclidean distance from a vector y to a set Y.

Proof.

(a) By using the nonexpansive property of the projection operation, from relation (7) we obtain for any $\mu \geq 0$ and all k,

$$\|\mu_{k+1} - \mu\|^2 = \left\|[\mu_k + \alpha g_k]^+ - \mu\right\|^2 \leq \|\mu_k + \alpha g_k - \mu\|^2.$$

Therefore,

$$\|\mu_{k+1} - \mu\|^2 \le \|\mu_k - \mu\|^2 + 2\alpha g_k'(\mu_k - \mu) + \alpha^2 \|g_k\|^2 \qquad \text{for all } k.$$

Since g_k is a subgradient of q at μ_k [cf. Eq. (1)], we have

$$g_k'(\mu - \mu_k) \ge q(\mu) - q(\mu_k),$$

implying that

$$g_k'(\mu_k - \mu) \le -\left(q(\mu) - q(\mu_k)\right).$$

Hence, for any $\mu \ge 0$,

$$\|\mu_{k+1} - \mu\|^2 \le \|\mu_k - \mu\|^2 - 2\alpha\left(q(\mu) - q(\mu_k)\right) + \alpha^2 \|g_k\|^2 \qquad \text{for all } k.$$

(b) By using the preceding relation with $\mu = \mu^*$ for any optimal solution μ^*, we obtain

$$\|\mu_{k+1} - \mu^*\|^2 \le \|\mu_k - \mu^*\|^2 - 2\alpha\left(q^* - q(\mu_k)\right) + \alpha^2 \|g_k\|^2 \qquad \text{for all } k \ge 0.$$

The desired relation follows by taking the infimum over all $\mu^* \in M^*$ in both sides of the preceding relation.

■

3.2 Bounded Multipliers

Here, we show that the multiplier sequence $\{\mu_k\}$ produced by the subgradient algorithm is bounded under the Slater condition and the bounded subgradient assumption. We formally state the latter requirement in the following.

Assumption 2 *(Bounded Subgradients)* The subgradient sequence $\{g_k\}$ is bounded, i.e., there exists a scalar $L > 0$ such that

$$\|g_k\| \le L \qquad \text{for all } k \ge 0.$$

This assumption is satisfied, for example, when the primal constraint set X is compact. In this case, due to the convexity of the constraint functions g_j over $\mathbb{R}^n$, each g_j is continuous over $\mathbb{R}^n$. Thus, $\max_{x \in X} \|g(x)\|$ is finite and provides an upper bound on the norms of the subgradients g_k, and hence, we can let

$$L = \max_{x \in X} \|g(x)\| \qquad \text{or} \qquad L = \max_{1 \le j \le m} \max_{x \in X} |g_j(x)|.$$

In the following lemma, we establish the boundedness of the multiplier sequence. In this, we use the boundedness of the dual sets $\{\mu \ge 0 \mid q(\mu) \ge q(\bar{\mu})\}$ [cf. Lemma 1] and the basic relation for the sequence $\{\mu_k\}$ of Lemma 2(a).

Lemma 3 *(Bounded Multipliers)* Let the multiplier sequence $\{\mu_k\}$ be generated by the subgradient algorithm of Eq. (7). Also, let the Slater condition and the bounded subgradient assumption hold [cf. Assumptions 1 and 2]. Then, the sequence $\{\mu_k\}$ is bounded and, in particular, we have

$$\|\mu_k\| \leq \frac{2}{\gamma}\left(f(\bar{x}) - q^*\right) + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\},$$

where $\gamma = \min_{1\leq j\leq m}\{-g_j(\bar{x})\}$, L is the subgradient norm bound of Assumption 2, $\bar{x}$ is a vector that satisfies the Slater condition, and $\alpha > 0$ is the stepsize.

Proof. Under the Slater condition the optimal dual set M^* is nonempty. Consider the set Q_α defined by

$$Q_\alpha = \left\{\mu \geq 0 \mid q(\mu) \geq q^* - \frac{\alpha L^2}{2}\right\},$$

which is nonempty in view of $M^* \subset Q_\alpha$. We fix an arbitrary $\mu^* \in M^*$ and we first prove that for all $k \geq 0$,

$$\|\mu_k - \mu^*\| \leq \max\left\{\|\mu_0 - \mu^*\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \|\mu^*\| + \alpha L\right\}, \tag{9}$$

where $\gamma = \min_{1\leq j\leq m}\{-g_j(\bar{x})\}$ and L is the bound on the subgradient norms $\|g_k\|$. Then, we use Lemma 1 to prove the desired estimate.

We show that relation (9) holds by induction on k. Note that the relation holds for $k = 0$. Assume now that it holds for some $k > 0$, i.e.,

$$\|\mu_k - \mu^*\| \leq \max\left\{\|\mu_0 - \mu^*\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \|\mu^*\| + \alpha L\right\} \text{ for some } k > 0. \tag{10}$$

We now consider two cases: $q(\mu_k) \geq q^* - \alpha L^2/2$ and $q(\mu_k) < q^* - \alpha L^2/2$.

Case 1: $q(\mu_k) \geq q^* - \alpha L^2/2$. By using the definition of the iterate μ_{k+1} in Eq. (7) and the subgradient boundedness, we obtain

$$\|\mu_{k+1} - \mu^*\| \leq \|\mu_k + \alpha g_k - \mu^*\| \leq \|\mu_k\| + \|\mu^*\| + \alpha L.$$

Since $q(\mu_k) \geq q^* - \alpha L^2/2$, it follows that $\mu_k \in Q_\alpha$. According to Lemma 1, the set Q_α is bounded and, in particular, $\|\mu\| \leq \frac{1}{\gamma}\left(f(\bar{x}) - q^* + \alpha L^2/2\right)$ for all $\mu \in Q_\alpha$. Therefore

$$\|\mu_k\| \leq \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma}.$$

By combining the preceding two relations, we obtain

$$\|\mu_{k+1} - \mu^*\| \leq \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \|\mu^*\| + \alpha L,$$

thus showing that the estimate in Eq. (9) holds for $k + 1$.

Case 2: $q(\mu_k) < q^* - \alpha L^2/2$. By using Lemma 2(a) with $\mu = \mu^*$, we obtain

$$\|\mu_{k+1} - \mu^*\|^2 \leq \|\mu_k - \mu^*\|^2 - 2\alpha\left(q^* - q(\mu_k)\right) + \alpha^2\|g_k\|^2.$$

By using the subgradient boundedness, we further obtain

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

Since $q(\mu_k) < q^* - \alpha L^2/2$, it follows that $q^* - q(\mu_k) - \alpha L^2/2 > 0$, which when combined with the preceding relation yields

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

By the induction hypothesis [cf. Eq. (10)], it follows that the estimate in Eq. (9) holds for $k+1$ in this case, too. Hence, the estimate in Eq. (9) holds for all $k \geq 0$.

From Eq. (9) we obtain for all $k \geq 0$,

$$\|\mu_k\| \leq \|\mu_k - \mu^*\| + \|\mu^*\| \leq \max\left\{\|\mu_0 - \mu^*\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \|\mu^*\| + \alpha L\right\} + \|\mu^*\|.$$

By using $\|\mu_0 - \mu^*\| \leq \|\mu_0\| + \|\mu^*\|$, we further have for all $k \geq 0$,

$$\begin{aligned} \|\mu_k\| &\leq \max\left\{\|\mu_0\| + \|\mu^*\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \|\mu^*\| + \alpha L\right\} + \|\mu^*\| \\ &= 2\|\mu^*\| + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\}. \end{aligned}$$

Since $M^* = \{\mu \geq 0 \mid q(\mu) \geq q^*\}$, according to Lemma 1, we have the following bound on the dual optimal solutions

$$\max_{\mu^* \in M^*} \|\mu^*\| \leq \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right),$$

implying that for all $k \geq 0$,

$$\|\mu_k\| \leq \frac{2}{\gamma}\left(f(\bar{x}) - q^*\right) + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\}.$$

■

The bound of Lemma 3 depends explicitly on the dual optimal value q^*. In practice, the value q^* is not readily available. However, since $q^* \geq q(\mu_0)$, by replacing q^* with $q(\mu_0)$, we have obtain the following norm bound for the multiplier sequence:

$$\|\mu_k\| \leq \frac{2}{\gamma}\left(f(\bar{x}) - q(\mu_0)\right) + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\left(f(\bar{x}) - q(\mu_0)\right) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\},$$

where $\gamma = \min_{1 \leq j \leq m}\{-g_j(\bar{x})\}$. Note that this bound depends on the algorithm parameters and problem data only. Specifically, it involves the initial iterate μ_0 of the subgradient method, the stepsize α, the vector $\bar{x}$ satisfying the Slater condition, and the subgradient norm bound L. In some practical applications, such as those in network optimization, such data is readily available. One may think of optimizing this bound with respect to the Slater vector $\bar{x}$. This might be an interesting and challenging problem on its own. However, this is outside the scope of our paper.

4 Approximate Primal Solutions

In this section, we provide approximate primal solutions by considering the running averages of the primal sequence $\{x_k\}$ generated as a by-product of the subgradient method [cf. Eq. (8)]. Intuitively, one would expect that, by averaging, the primal cost and the amount of constraint violation of primal infeasible vectors can be reduced due to the convexity of the cost and the constraint functions. It turns out that the benefits of averaging are far more reaching than merely cost and infeasibility reduction. We show here that under the Slater condition, we can also provide upper bounds for the number of subgradient iterations needed to generate a primal solution within a given level of constraint violation. We also derive upper and lower bounds on the gap from the optimal primal value. These bounds depend on some assumptions and prior information such as a Slater vector and a bound on subgradient norms.

We now introduce the notation that we use in our averaging scheme throughout the rest of the paper. We consider the multiplier sequence $\{\mu_k\}$ generated by the subgradient algorithm of Eq. (7), and the corresponding sequence of primal vectors $\{x_k\} \subset X$ that provide the subgradients g_k in the algorithm of Eq. (7), i.e.,

$$g_k = g(x_k), \qquad x_k \in \operatorname*{argmin}_{x\in X}\{f(x) + \mu_k' g(x)\} \qquad \text{for all } k \geq 0.$$

[cf. Eq. (8)]. We define $\hat{x}_k$ as the average of the vectors $x_0, \ldots, x_{k-1}$, i.e.,

$$\hat{x}_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i \qquad \text{for all } k \geq 1. \tag{11}$$

The average vectors $\hat{x}_k$ lie in the set X because X is convex and $x_i \in X$ for all i. However, these vectors need not satisfy the primal inequality constraints $g_j(x) \leq 0$, $j = 0, \ldots, m$, and therefore, they can be primal infeasible.

In the rest of this section, we study some basic properties of the average vectors $\hat{x}_k$. Using these properties and the Slater condition, we provide estimates for the primal optimal value and the feasibility violation at each iteration of the subgradient method.

4.1 Basic Properties of the Averaged Primal Sequence

In this section, we provide upper and lower bounds on the primal cost of the running averages $\hat{x}_k$. We also provide an upper and a lower bound on the amount of feasibility violation of these vectors. These bounds are given per iteration, as seen in the following.

Proposition 1 Let the multiplier sequence $\{\mu_k\}$ be generated by the subgradient method of Eq. (7). Let the vectors $\hat{x}_k$ for $k \geq 1$ be the averages given by Eq. (11). Then, for all $k \geq 1$, the following hold:

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

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

(b) An upper bound on the primal cost of the vector $\hat{x}_k$ is given by

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

(c) A lower bound on the primal cost of the vector $\hat{x}_k$ is given by

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

where μ^* is a dual optimal solution.

Proof. (a) By using the definition of the iterate μ_{k+1} in Eq. (7), we obtain

$$\mu_k + \alpha g_k \leq [\mu_k + \alpha g_k]^+ = \mu_{k+1} \qquad \text{for all } k \geq 0.$$

Since $g_k = g(x_k)$ with $x_k \in X$, it follows that

$$\alpha g(x_k) \leq \mu_{k+1} - \mu_k \qquad \text{for all } k \geq 0.$$

Therefore,

$$\sum_{i=0}^{k-1} \alpha g(x_i) \leq \mu_k - \mu_0 \leq \mu_k \qquad \text{for all } k \geq 1,$$

where the last inequality in the preceding relation follows from $\mu_0 \geq 0$. Since $x_k \in X$ for all k, by the convexity of X, we have $\hat{x}_k \in X$ for all k. Hence, by the convexity of each of the functions g_j, it follows that

$$g(\hat{x}_k) \leq \frac{1}{k} \sum_{i=0}^{k-1} g(x_i) = \frac{1}{k\alpha} \sum_{i=0}^{k-1} \alpha g(x_i) \leq \frac{\mu_k}{k\alpha} \qquad \text{for all } k \geq 1.$$

Because $\mu_k \geq 0$ for all k, we have $g(\hat{x}_k)^+ \leq \mu_k/(k\alpha)$ for all $k \geq 1$ and, therefore,

$$\left\|g(\hat{x}_k)^+\right\| \leq \frac{\|\mu_k\|}{k\alpha} \qquad \text{for all } k \geq 1.$$

(b) By the convexity of the primal cost $f(x)$ and the definition of x_k as a minimizer of the Lagrangian function $f(x) + \mu_k' g(x)$ over $x \in X$ [cf. Eq. (8)], we have

$$f(\hat{x}_k) \leq \frac{1}{k} \sum_{i=0}^{k-1} f(x_i) = \frac{1}{k} \sum_{i=0}^{k-1} \{f(x_i) + \mu_i' g(x_i)\} - \frac{1}{k} \sum_{i=0}^{k-1} \mu_i g(x_i).$$

Since $q(\mu_i) = f(x_i) + \mu_i' g(x_i)$ and $q(\mu_i) \leq q^*$ for all i, it follows that for all $k \geq 1$,

$$f(\hat{x}_k) \leq \frac{1}{k} \sum_{i=0}^{k-1} q(\mu_i) - \frac{1}{k} \sum_{i=0}^{k-1} \mu_i' g(x_i) \leq q^* - \frac{1}{k} \sum_{i=0}^{k-1} \mu_i' g(x_i). \qquad (12)$$

From the definition of the algorithm in Eq. (7), by using the nonexpansive property of the projection, and the facts $0 \in \{\mu \in \mathbb{R}^m \mid \mu \geq 0\}$ and $g_i = g(x_i)$, we obtain

$$\|\mu_{i+1}\|^2 \leq \|\mu_i\|^2 + 2\alpha\mu_i' g(x_i) + \alpha^2\|g(x_i)\|^2 \qquad \text{for all } i \geq 0,$$

implying that

$$-\mu_i' g(x_i) \leq \frac{\|\mu_i\|^2 - \|\mu_{i+1}\|^2 + \alpha^2\|g(x_i)\|^2}{2\alpha} \qquad \text{for all } i \geq 0.$$

By summing over $i = 0, \ldots, k-1$ for $k \geq 1$, we have

$$-\frac{1}{k}\sum_{i=0}^{k-1} \mu_i' g(x_i) \leq \frac{\|\mu_0\|^2 - \|\mu_k\|^2}{2k\alpha} + \frac{\alpha}{2k}\sum_{i=0}^{k-1}\|g(x_i)\|^2 \qquad \text{for all } k \geq 1.$$

Combining the preceding relation and Eq. (12), we further have

$$f(\hat{x}_k) \leq q^* + \frac{\|\mu_0\|^2 - \|\mu_k\|^2}{2k\alpha} + \frac{\alpha}{2k}\sum_{i=0}^{k-1}\|g(x_i)\|^2 \qquad \text{for all } k \geq 1,$$

implying the desired estimate.

(c) Given a dual optimal solution μ^*, we have

$$f(\hat{x}_k) = f(\hat{x}_k) + (\mu^*)' g(\hat{x}_k) - (\mu^*)' g(\hat{x}_k) \geq q(\mu^*) - (\mu^*)' g(\hat{x}_k).$$

Because $\mu^* \geq 0$ and $g(\hat{x}_k)^+ \geq g(\hat{x}_k)$, we further have

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

From the preceding two relations and the fact $q(\mu^*) = q^*$ it follows that

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

■

An immediate consequence of Proposition 1(a) is that the maximum violation $\|g(\hat{x}_k)^+\|_\infty$ of constraints $g_j(x)$, $j = 1, \ldots, m$, at $x = \hat{x}_k$ is bounded by the same bound. In particular, we have

$$\max_{1\leq j\leq m} g_j(\hat{x}_k)^+ \leq \frac{\|\mu_k\|}{k\alpha} \qquad \text{for all } k \geq 1.$$

which follows from the proposition in view of the relation $\|y\|_\infty \leq \|y\|$ for any y.

We note that the results of Proposition 1 in parts (a) and (c) show how the amount of feasibility violation $\|g(\hat{x}_k)^+\|$ affects the lower estimate of $f(\hat{x}_k)$. Furthermore, we note that the results of Proposition 1 indicate that the bounds on the feasibility violation and the primal value $f(\hat{x}_k)$ are readily available provided that we have bounds on the multiplier norms $\|\mu_k\|$, optimal solution norms $\|\mu^*\|$, and subgradient norms $\|g(x_k)\|$. This is precisely what we use in the next section to establish our estimates.

Finally, let us note that the bounds on the primal cost of Proposition 1 in parts (b) and (c) hold for a more general subgradient algorithm than the algorithm of Eq. (7). In particular, the result in part (c) is independent of the algorithm that is used to generate the multiplier sequence $\{\mu_k\}$. The proof of the result in part (c) relies on the nonexpansive property of the projection operation and the fact that the zero vector belongs to the projection set $\{\mu \in \mathbb{R}^m \mid \mu \geq 0\}$. Therefore, the results in parts (b) and (c) hold when we use a more general subgradient algorithm of the following form:

$$\mu_{k+1} = \mathcal{P}_D[\mu_k + \alpha g_k] \qquad \text{for } k \geq 1,$$

where $D \subseteq \{\mu \in \mathbb{R}^m \mid \mu \geq 0\}$ is a closed convex set containing the zero vector. We study a subgradient algorithm of this form in Section 5 and establish similar error estimates.

4.2 Properties of the Averaged Primal Sequence under Slater

Here, we strengthen the relations of Proposition 1 under the Slater condition and the subgradient boundedness. Our main result is given in the following proposition.

Proposition 2 Let the sequence $\{\mu_k\}$ be generated by the subgradient algorithm (7). Let the Slater condition and the bounded subgradient assumption hold [cf. Assumptions 1 and 2]. Also, let

$$B^* = \frac{2}{\gamma}\left(f(\bar{x}) - q^*\right) + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\}, \tag{13}$$

where $\gamma = \min_{1\leq j\leq m}\{-g_j(\bar{x})\}$, $\bar{x}$ is a Slater vector of Assumption 1, and L is the subgradient norm bound of Assumption 2. Let the vectors $\hat{x}_k$ for $k \geq 1$ be the averages given by Eq. (11). Then, the following holds for all $k \geq 1$:

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

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

(b) An upper bound on the primal cost of the vector $\hat{x}_k$ is given by

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

(c) A lower bound on the primal cost of the vector $\hat{x}_k$ is given by

$$f(\hat{x}_k) \geq f^* - \frac{1}{\gamma}\left[\, f(\bar{x}) - q^* \,\right] \|g(\hat{x}_k)^+\|.$$

Proof.

(a) Under the Slater and bounded subgradient assumptions, by Lemma 3 we have

$$\|\mu_k\| \le \frac{2}{\gamma}\,(f(\bar{x}) - q^*) + \max\left\{ \|\mu_0\|, \frac{1}{\gamma}\,(f(\bar{x}) - q^*) + \frac{\alpha L^2}{2\gamma} + \alpha L \right\} \qquad \text{for all } k \ge 0.$$

By the definition of B^* in Eq. (13), the preceding relation is equivalent to

$$\|\mu_k\| \le B^* \qquad \text{for all } k \ge 0. \tag{14}$$

By using Proposition 1(a), we obtain

$$\|g(\hat{x}_k)^+\| \le \frac{\|\mu_k\|}{k\alpha} \le \frac{B^*}{k\alpha} \qquad \text{for all } k \ge 1.$$

(b) From Proposition 1(b), we obtain

$$f(\hat{x}_k) \le q^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha}{2k}\sum_{i=0}^{k-1} \|g(x_i)\|^2 \qquad \text{for all } k \ge 1.$$

Under the Slater condition, there is zero duality gap, i.e., $q^* = f^*$. Furthermore, the subgradients are bounded by a scalar L [cf. Assumption 2], so that

$$f(\hat{x}_k) \le f^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha L^2}{2} \qquad \text{for all } k \ge 1.$$

(c) Under the Slater condition, a dual optimal solution exists and there is zero duality gap, i.e., $q^* = f^*$. Thus, by Proposition 1(c), for any dual solution μ^* we have

$$f(\hat{x}_k) \ge f^* - \|\mu^*\| \|g(\hat{x}_k)^+\| \qquad \text{for all } k \ge 1.$$

By using Lemma 1 with $\bar{\mu} = \mu^*$, we see that the dual set is bounded and, in particular, $\|\mu^*\| \le \frac{1}{\gamma}(f(\bar{x}) - q^*)$ for all dual optimal vectors μ^*. Hence,

$$f(\hat{x}_k) \ge f^* - \frac{1}{\gamma}\,[\, f(\bar{x}) - q^* \,]\, \|g(\hat{x}_k)^+\| \qquad \text{for all } k \ge 1.$$

■

As indicated in Proposition 2(a), the amount of feasibility violation $\|g(\hat{x}_k)^+\|$ of the vector $\hat{x}_k$ diminishes to zero at the rate $1/k$ as the number of subgradient iterations k increases. By combining the results in (a)-(c), we see that the function values $f(\hat{x}_k)$ converge to f^* within error level $\alpha L^2/2$ with the same rate of $1/k$ as $k \to \infty$.

Let us note that a more practical bound than the bound B^* of Proposition 2 can be obtained by noting that $q^* \ge q(\mu_0)$. In this case, instead of B^* we may use the following:

$$B_0 = \frac{2}{\gamma}\,(f(\bar{x}) - q(\mu_0)) + \max\left\{ \|\mu_0\|, \frac{1}{\gamma}\,(f(\bar{x}) - q(\mu_0)) + \frac{\alpha L^2}{2\gamma} + \alpha L \right\},$$

which satisfies $B_0 \geq B^*$. Similarly, we can use $q(\mu_0) \leq q^*$ in the estimate in part (c). Thus, the bounds in parts (a) and (c) of Proposition 2 can be replaced by the following

$$\|g(\hat{x}_k)^+\| \leq \frac{B_0}{k\alpha}, \qquad f(\hat{x}_k) \geq f^* - \frac{1}{\gamma}\,[\,f(\bar{x}) - q(\mu_0)\,]\,\|g(\hat{x}_k)^+\| \qquad \text{for } k \geq 1.$$

Another possibility is to use $\max_{0\leq i\leq k} q(\mu_i)$ as an approximation of the dual optimal value q^*. In this case, in view of $q^* \geq \max_{0\leq i\leq k} q(\mu_i)$, the bounds in parts (a) and (c) of Proposition 2 may be replaced by

$$\|g(\hat{x}_k)^+\| \leq \frac{B_k}{k\alpha}, \qquad f(\hat{x}_k) \geq f^* - \frac{1}{\gamma}\left[f(\bar{x}) - \max_{0\leq i\leq k} q(\mu_i)\right]\|g(\hat{x}_k)^+\| \qquad \text{for } k \geq 1, \tag{15}$$

where B_k is given by: for all $k \geq 1$,

$$B_k = \frac{2}{\gamma}\left[f(\bar{x}) - \max_{0\leq i\leq k} q(\mu_i)\right] + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\left[f(\bar{x}) - \max_{0\leq i\leq k} q(\mu_i)\right] + \frac{\alpha L^2}{2\gamma} + \alpha L\right\}. \tag{16}$$

Finally, let us note that it seems reasonable to choose the initial iterate as $\mu_0 = 0$, as suggested by the upper bound for $f(\hat{x}_k)$ in part (b). Then, the estimate in part (b) reduces to

$$f(\hat{x}_k) \leq f^* + \frac{\alpha L^2}{2} \qquad \text{for all } k \geq 1. \tag{17}$$

Thus, all of the approximate solutions $\hat{x}_k$ have primal value below optimal f^* within the error level $\frac{\alpha L^2}{2}$ proportional to the stepsize. Furthermore, for $\mu_0 = 0$, the bounds B_k of Eq. (16) reduce to

$$B_k^0 = \frac{3}{\gamma}\left[f(\bar{x}) - \max_{0\leq i\leq k} q(\mu_i)\right] + \frac{\alpha L^2}{2\gamma} + \alpha L \qquad \text{for } k \geq 1. \tag{18}$$

For $\mu_0 = 0$, it can be seen that the feasibility violation estimate for $\hat{x}_k$ and the lower bound for $f(\hat{x}_k)$ of Eq. (15) can be replaced by the following:

$$\left\|g(\hat{x}_k)^+\right\| \leq \frac{B_k^0}{k\alpha}, \qquad f(\hat{x}_k) \geq f^* - \frac{1}{\gamma}\left[f(\bar{x}) - \max_{0\leq i\leq k} q(\mu_i)\right]\|g(\hat{x}_k)^+\| \qquad \text{for } k \geq 1. \tag{19}$$

The preceding bounds can be used for developing practical stopping criteria for the subgradient algorithm with primal averaging. In particular, a user may specify a maximum tolerable infeasibility and/or a desired level of primal optimal accuracy. These specifications combined, for example, with the estimates in Eqs. (17)–(19) can be used to choose the stepsize value α, and to analyze the trade-off between the desired accuracy and the associated computational complexity (in terms of the number of subgradient iterations).

5 Modified Subgradient Method under Slater

In this section, we consider a modified version of the subgradient method under the Slater assumption. The motivation is coming from the fact that under the Slater assumption, the set of dual optimal solutions is bounded (cf. Lemma 1). Therefore, it is of interest to consider a subgradient method in which dual iterates are projected onto a bounded superset of the dual optimal solution set. We consider such algorithms and generate primal solutions using averaging as described in Section 4. Also, we provide estimates for the amount of constraint violation and cost of the average primal sequence. Our goal is to compare these estimates with the error estimates obtained for the "ordinary" subgradient method in Section 4.

Formally, we consider subgradient methods of the following form:

$$\mu_{k+1} = \mathcal{P}_D\, [\mu_k + \alpha g_k], \tag{20}$$

where the set D is a compact convex set containing the set of dual optimal solutions (to be discussed shortly) and $\mathcal{P}_D$ denotes the projection on the set D. The vector $\mu_0 \in D$ is an arbitrary initial iterate and the scalar $\alpha > 0$ is a constant stepsize. The vector g_k is a subgradient of q at μ_k given by

$$g_k = g(x_k), \qquad x_k \in \operatorname*{argmin}_{x\in X}\{f(x) + \mu_k' g(x)\} \qquad \text{for all } k \geq 0$$

[see Eq. (5)].

Under the Slater condition, the dual optimal set M^* is nonempty and bounded, and a bound on the norms of the dual optimal solutions is given by

$$\sum_{j=1}^{m} \mu_j^* \leq \frac{1}{\gamma}(f(\bar{x}) - q^*) \qquad \text{for all } \mu^* \in M^*,$$

with $\gamma = \min_{1\leq j\leq m}\{-g_j(\bar{x})\}$ and $\bar{x}$ a Slater vector [cf. Lemma 1]. Thus, having the dual value $\tilde{q} = q(\tilde{\mu})$ for some $\tilde{\mu} \geq 0$, since $q^* \geq \tilde{q}$, we obtain

$$\sum_{j=1}^{m} \mu_j^* \leq \frac{1}{\gamma}(f(\bar{x}) - \tilde{q}) \qquad \text{for all } \mu^* \in M^*. \tag{21}$$

This motivates the following choice for the set D:

$$D = \left\{ \mu \geq 0 \;\middle|\; \|\mu\| \leq \frac{f(\bar{x}) - \tilde{q}}{\gamma} + r \right\}, \tag{22}$$

with a scalar $r > 0$. Clearly, the set D is compact and convex, and it contains the set of dual optimal solutions in view of relation (21) and the fact $\|y\| \leq \|y\|_1$ for any vector y; (the illustration of the set D is provided in Figure 1).

Similar to Section 4, we provide near-feasible and near-optimal primal vectors by averaging the vectors from the sequence $\{x_k\}$. In particular, we define $\hat{x}_k$ as the average of the vectors $x_0, \ldots, x_{k-1}$, i.e.,

$$\hat{x}_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i \qquad \text{for all } k \geq 1. \tag{23}$$

In the next proposition, we provide per-iterate bounds for the constraint violation and primal cost values of the average vectors $\hat{x}_k$.

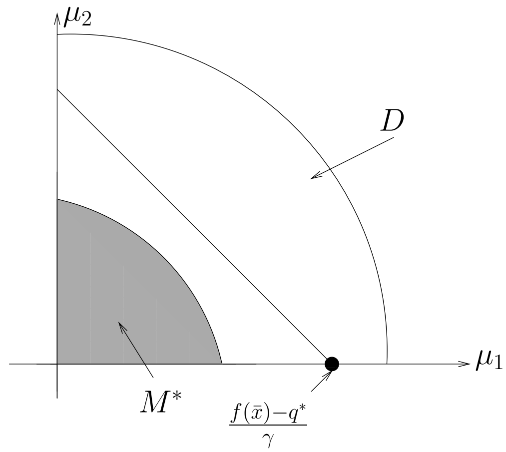

Figure 1: The dual optimal set M^* and the set D_2, which is considered in the modified subgradient method.

Proposition 3 Let the Slater condition and the bounded subgradient assumption hold [cf. Assumptions 1 and 2]. Let the dual sequence $\{\mu_k\}$ be generated by the modified subgradient method of Eq. (20). Let $\{\hat{x}_k\}$ be the average sequence defined in Eq. (23). Then, for all $k \geq 1$, we have:

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

$$\|g(\hat{x}_k)^+\| \leq \frac{2}{k\alpha r}\left(\frac{f(\bar{x})-\tilde{q}}{\gamma}+r\right)^2+\frac{\alpha L^2}{2r}.$$

(b) An upper bound on the primal cost of the vector $\hat{x}_k$ is given by

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

(c) A lower bound on the primal cost of the vector $\hat{x}_k$ is given by

$$f(\hat{x}_k) \geq f^* - \left(\frac{f(\bar{x})-\tilde{q}}{\gamma}\right)\|g(\hat{x}_k)^+\|.$$

Here, the scalars $r > 0$ and $\tilde{q}$ with $\tilde{q} \leq q^*$ are those from the definition of the set D in Eq. (22), $\gamma = \min_{1\leq j\leq m}\{-g_j(\bar{x})\}$, $\bar{x}$ is the Slater vector of Assumption 1, and L is the subgradient norm bound of Assumption 2.

Proof.

(a) Using the definition of the iterate μ_{k+1} in Eq. (20) and the nonexpansive property of projection on a closed convex set, we obtain for all $\mu \in D$ and all $i \geq 0$,

$$\begin{aligned}
\|\mu_{i+1}-\mu\|^2 &= \|\mathcal{P}_D\,[\mu_i+\alpha g_i]-\mu\|^2 \\
&\leq \|\mu_i+\alpha g_i-\mu\|^2 \\
&\leq \|\mu_i-\mu\|^2+2\alpha g_i'(\mu_i-\mu)+\alpha^2\|g_i\|^2 \\
&\leq \|\mu_i-\mu\|^2+2\alpha g_i'(\mu_i-\mu)+\alpha^2 L^2,
\end{aligned}$$

where the last inequality follows from the bounded subgradient assumption (cf. Assumption 2). Therefore, for any $\mu \in D$,

$$g_i'(\mu - \mu_i) \le \frac{\|\mu_i - \mu\|^2 - \|\mu_{i+1} - \mu\|^2}{2\alpha} + \frac{\alpha L^2}{2} \qquad \text{for all } i \ge 0. \tag{24}$$

Since g_i is a subgradient of the dual function q at μ_i, using the subgradient inequality [cf. Eq. (1)], we obtain for any dual optimal solution μ^*,

$$g_i'(\mu_i - \mu^*) \le q(\mu_i) - q(\mu^*) \le 0 \qquad \text{for all } i \ge 0,$$

where the last inequality follows from the optimality of μ^* and the feasibility of each $\mu_i \in D$ [i.e., $\mu_i \ge 0$]. We then have for all $\mu \in D$ and all $i \ge 0$,

$$g_i'(\mu - \mu^*) = g_i'(\mu - \mu^* - \mu_i + \mu_i) = g_i'(\mu - \mu_i) + g_i'(\mu_i - \mu^*) \le g_i'(\mu - \mu_i).$$

From the preceding relation and Eq. (24), we obtain for any $\mu \in D$,

$$g_i'(\mu - \mu^*) \le \frac{\|\mu_i - \mu\|^2 - \|\mu_{i+1} - \mu\|^2}{2\alpha} + \frac{\alpha L^2}{2} \qquad \text{for all } i \ge 0.$$

Summing over $i = 0, \dots, k-1$ for $k \ge 1$, we obtain for any $\mu \in D$ and $k \ge 1$,

$$\sum_{i=0}^{k-1} g_i'(\mu - \mu^*) \le \frac{\|\mu_0 - \mu\|^2 - \|\mu_k - \mu\|^2}{2\alpha} + \frac{\alpha k L^2}{2} \le \frac{\|\mu_0 - \mu\|^2}{2\alpha} + \frac{\alpha k L^2}{2}.$$

Therefore, for any $k \ge 1$,

$$\max_{\mu \in D} \left\{ \sum_{i=0}^{k-1} g_i'(\mu - \mu^*) \right\} \le \frac{1}{2\alpha} \max_{\mu \in D} \|\mu_0 - \mu\|^2 + \frac{\alpha k L^2}{2}. \tag{25}$$

We now provide a lower estimate on the left-hand side of the preceding relation. Let $k \ge 1$ be arbitrary and, for simplicity, we suppress the explicit dependence on k by letting

$$s = \sum_{i=0}^{k-1} g_i. \tag{26}$$

Let s^+ be the component-wise maximum of s and the zero vector, i.e., the j-th entry of the vector s^+ is given by $s_j^+ = \max\{s_j, 0\}$. If $s^+ = 0$, then the bound in part (a) of this proposition trivially holds. Thus, assume that $s^+ \ne 0$ and define a vector $\bar{\mu}$ as follows:

$$\bar{\mu} = \mu^* + r \frac{s^+}{\|s^+\|}.$$

Note that $\bar{\mu} \ge 0$ since $\mu^* \ge 0$, $s^+ \ge 0$ and $r > 0$. By Lemma 1, the dual optimal solution set is bounded and, in particular, $\|\mu^*\| \le \frac{f(\bar{x}) - q^*}{\gamma}$. Furthermore, since

$\tilde{q} \leq q$, it follows that $\|\mu^*\| \leq \frac{f(\bar{x})-\tilde{q}}{\gamma}$ for any dual solution μ^*. Therefore, by the definition of the vector $\bar{\mu}$, we have

$$\|\bar{\mu}\| \leq \|\mu^*\| + r \leq \frac{f(\bar{x}) - \tilde{q}}{\gamma} + r, \tag{27}$$

implying that $\bar{\mu} \in D$. Using the definition of the vector s in Eq. (26) and relation (25), we obtain

$$s'(\bar{\mu}-\mu^*) = \sum_{i=0}^{k-1} g_i'(\bar{\mu}-\mu^*) \leq \max_{\mu \in D} \left\{ \sum_{i=0}^{k-1} g_i'(\mu - \mu^*) \right\} \leq \frac{1}{2\alpha} \max_{\mu \in D} \|\mu_0 - \mu\|^2 + \frac{\alpha k L^2}{2}.$$

Since $\bar{\mu} - \mu^* = r\frac{s^+}{\|s^+\|}$, we have $s'(\bar{\mu} - \mu^*) = r\|s^+\|$. Thus, by the definition of s in Eq. (26) and the fact $g_i = g(x_i)$, we have

$$s'(\bar{\mu} - \mu^*) = r\left\| \left[\sum_{i=0}^{k-1} g(x_i) \right]^+ \right\|.$$

Combining the preceding two relations, it follows that

$$\left\| \left[\sum_{i=0}^{k-1} g(x_i) \right]^+ \right\| \leq \frac{1}{2\alpha r} \max_{\mu \in D} \|\mu_0 - \mu\|^2 + \frac{\alpha k L^2}{2r}.$$

Dividing both sides of this relation by k, and using the convexity of the functions g_j in $g = (g_1, \ldots, g_m)$ and the definition of the average primal vector $\hat{x}_k$, we obtain

$$\|g(\hat{x}_k)^+\| \leq \frac{1}{k}\left\| \left[\sum_{i=0}^{k-1} g(x_i) \right]^+ \right\| \leq \frac{1}{2k\alpha r} \max_{\mu \in D} \|\mu_0 - \mu\|^2 + \frac{\alpha L^2}{2r}. \tag{28}$$

Since $\mu_0 \in D$, we have

$$\max_{\mu \in D} \|\mu_0 - \mu\|^2 \leq \max_{\mu \in D}(\|\mu_0\| + \|\mu\|)^2 \leq 4 \max_{\mu \in D} \|\mu\|^2.$$

By using the definition of the set D of Eq. (22), we have

$$\max_{\mu \in D} \|\mu\| \leq \frac{f(\bar{x}) - \tilde{q}}{\gamma} + r.$$

By substituting the preceding two estimates in relation (28), we obtain

$$\|g(\hat{x}_k)^+\| \leq \frac{2}{k\alpha r} \left(\frac{f(\bar{x}) - \tilde{q}}{\gamma} + r \right)^2 + \frac{\alpha L^2}{2r}.$$

(b) The proof follows from an identical argument to that used in the proof of Proposition 1(c), and therefore is omitted.

(c) The result follows from an identical argument to that used in the proof of Proposition 1(c) and the bound on the dual optimal solution set that follows in view of the Slater condition (cf. Assumption 1 and Lemma 1 with $\bar{\mu} = \mu^*$).

■

We note here that the subgradient method of Eq. (20) with the set D given in Eq. (22) couples the computation of multipliers. In some applications, it might be desirable to accommodate distributed computation models whereby the multiplier components μ_j^* are processed in a distributed manner among a set of processors or agents. To accommodate such computations, one may modify the subgradient method of Eq. (20) by replacing the set D of Eq. (22) with the following set

$$D_\infty = \left\{ \mu \geq 0 \;\middle|\; \|\mu\|_\infty \leq \frac{f(\bar{x}) - \tilde{q}}{\gamma} + r \right\}.$$

It can be seen that the results of Proposition 3 also hold for this choice of the projection set. In particular, this can be seen by following the same line of argument as in the proof of Proposition 3 and by using the following relation

$$\|\bar{\mu}\|_\infty \leq \|\mu^*\| + r \leq \frac{f(\bar{x}) - \tilde{q}}{\gamma} + r$$

[cf. Eq. (27) and the fact $\|y\|_\infty \leq \|y\|$ for any vector y].

We next consider selecting the parameter r, which is used in the definition of the set D, such that the right-hand side of the bound in part (a) of Proposition 3 is minimized at each iteration k. Given some $k \geq 1$, we choose r as the optimal solution of the problem

$$\min_{r>0} \left\{ \frac{2}{k\alpha r} \left(\frac{f(\bar{x}) - \tilde{q}}{\gamma} + r \right)^2 + \frac{\alpha L^2}{2r} \right\}.$$

It can be seen that the optimal solution of the preceding problem, denoted by $r^*(k)$, is given by

$$r^*(k) = \sqrt{\left(\frac{f(\bar{x}) - \tilde{q}}{\gamma} \right)^2 + \frac{\alpha^2 L^2 k}{4}} \qquad \text{for } k \geq 1. \tag{29}$$

Consider now an algorithm where the dual iterates are obtained by

$$\mu_{i+1} = \mathcal{P}_{D_k} \left[\mu_i + \alpha g_k \right] \qquad \text{for each } i \geq 0,$$

with $\mu_0 \in D_0$ and the set D_k given by

$$D_k = \left\{ \mu \geq 0 \mid \|\mu\| \leq \frac{f(\bar{x}) - \tilde{q}}{\gamma} + r^*(k) \right\},$$

where $r^*(k)$ is given by Eq. (29). Hence, at each iteration i, the algorithm projects onto the set D_k, which contains the set of dual optimal solutions M^*.

Substituting $r^*(k)$ in the bound of Proposition 3(a), we can see that

$$\begin{aligned}
\|g(\hat{x}_k)^+\| &\leq \frac{4}{k\alpha}\left(\frac{f(\bar{x})-\tilde{q}}{\gamma}+\sqrt{\left(\frac{f(\bar{x})-\tilde{q}}{\gamma}\right)^2+\frac{\alpha^2L^2k}{4}}\right)\\
&\leq \frac{4}{k\alpha}\left(\frac{2(f(\bar{x})-\tilde{q})}{\gamma}+\frac{\alpha L\sqrt{k}}{2}\right)\\
&= \frac{8}{k\alpha}\left(\frac{f(\bar{x})-\tilde{q}}{\gamma}\right)+\frac{2L}{\sqrt{k}}.
\end{aligned}$$

The preceding discussion combined with Proposition 3(a) immediately yields the following result:

Proposition 4 Let the Slater condition and the bounded subgradient assumption hold [cf. Assumptions 1 and 2]. Given some $k \geq 1$, define the set D_k as

$$D_k = \left\{\mu \geq 0 \;\middle|\; \|\mu\|_2 \leq \frac{f(\bar{x})-\tilde{q}}{\gamma}+\sqrt{\left(\frac{f(\bar{x})-\tilde{q}}{\gamma}\right)^2+\frac{\alpha^2L^2k}{4}}\right\}, \tag{30}$$

where $\bar{x}$ is the Slater vector of Assumption 1, $\gamma = \min_{1\leq j\leq m}\{-g_j(\bar{x})\}$, and L is the bound on the subgradient norm of Assumption 2. Let the dual sequence $\{\mu_i\}$ be generated by the following modified subgradient method: Let $\mu_0 \in D_k$ and for each $i \geq 0$, the dual iterate μ_i is obtained by

$$\mu_{i+1} = \mathcal{P}_{D_k}\,[\mu_i + \alpha g_i].$$

Then, an upper bound on the amount of feasibility violation of the vector $\hat{x}_k$ is given by

$$\|g(\hat{x}_k)^+\| \leq \frac{8}{k\alpha}\left(\frac{f(\bar{x})-\tilde{q}}{\gamma}\right)+\frac{2L}{\sqrt{k}}. \tag{31}$$

This result shows that at a given k, the error estimate provided in Eq. (31) can be achieved if we use a modified subgradient method where each dual iterate is projected on the set D_k defined in Eq. (30). Given a pre-specified accuracy for the amount of feasibility violation, this bound can be used to select the stepsize value and the set D_k for the algorithm.

Using the estimate (31) in Proposition 3(c), we can obtain a lower bound on the cost $f(\hat{x}_k)$. As indicated from the preceding proposition, the modified algorithm with "optimal choice" of the parameter r of the projection set results in a primal near-optimal solution with the error of the order $1/\sqrt{k}$.

We now compare the feasibility violation bound of Proposition 2 (a) for the ordinary subgradient method with the result of Proposition 4 for the modified subgradient method. In particular, by Proposition 4(a) and by using $q^* \geq q(\tilde{\mu})$ for $\tilde{\mu} \geq 0$ in the definition of the bound B^* in Eq. (13), we obtain for all $k \geq 1$:

$$\|g(\hat{x}_k)^+\| \leq \frac{2}{k\alpha\gamma}\,(f(\bar{x})-\tilde{q}) + \frac{1}{k\alpha}\,\max\left\{\|\mu_0\|, \frac{1}{\gamma}\,(f(\bar{x})-\tilde{q}) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\}. \tag{32}$$

This bound, as compared to that of Proposition 4, is better when k is very large, since the feasibility violation decreases in the order of $1/k$, while the feasibility violation in Proposition 4 decreases at order $1/\sqrt{k}$. However, initially, the feasibility violation in Eq. (32) for the ordinary subgradient method can be worse because it depends on the initial iterate μ_0. Suppose that the initial iterate is $\mu_0 = 0$. Then, the bound in Eq. (32) reduces to

$$\|g(\hat{x}_k)^+\| \leq \frac{3}{k\alpha\gamma}\,[\,f(\bar{x}) - \tilde{q}\,] + \frac{1}{k}\left(\frac{L^2}{2\gamma} + L\right).$$

Even in this case, initially, this bound can be worse than that of Proposition 4 because the bound depends inversely on γ which can be very small [recall $\gamma = \min_{1\leq j\leq m}\{-g_j(\bar{x})\}$ with $\bar{x}$ being a Slater vector]. To complement our theoretical analysis of the algorithms, we need to perform some numerical experiments to further study and compare these algorithms.

6 Conclusions

In this paper, we have studied the application of dual subgradient algorithms for generating primal near-feasible and near-optimal optimal solutions. We have proposed and analyzed two such algorithms under Slater condition. Both of the proposed algorithms use projections to generate a dual sequence and an averaging scheme to produce approximate primal vectors. The algorithms employ the same averaging scheme in the primal space. However, they operate on different sets when projecting in the dual space. One algorithm uses the projections on the nonnegative orthant, while the other algorithm uses the projections on nested compact convex sets that change with each iteration but always contain the dual optimal solutions.

We have provided bounds on the amount of feasibility violation at approximate primal solutions for both of the subgradient schemes, as well as bounds on the primal function values. As indicated by our results, both algorithms produce primal vectors whose infeasibility diminishes to zero and whose function value converges to the primal optimal within an error. However, the rate at which the feasibility diminishes and the primal values converge are different. As revealed by our analysis, the convergence rate is $1/k$ in the number of iterations k for the ordinary subgradient method with the simple projections on the positive orthant, while the convergence is $1/\sqrt{k}$ for the subgradient method projections on nested sets.

We attribute the *theoretically* better performance of the ordinary subgradient method to the "quality" of information contained in the primal vectors that define the subgradients. Intuitively, we believe that the subgradients produced with simple projection carry more information about the primal problem than those generated by projections on more restricted sets. We postulate that by projecting on more restricted sets, even when these sets closely "locate" the dual solutions, some information from subgradients about the primal problem is lost, thus resulting in a more computations in order to compensate for the information loss.

Our comparison of the two methods is purely based on our *theoretical analysis*, which need not reflect the real behavior of these algorithms for practical implementations. Our future goal is to numerically test and evaluate these algorithms in order to gain deeper insights into their behavior.

References

[1] F. Barahona and R. Anbil, *The volume algorithm: producing primal solutions with a subgradient method*, Mathematical Programming **87** (2000), 385–399.

[2] A. Ben-Tal, T. Margalit, and A. Nemirovski, *The ordered subsets mirror descent optimization method and its use for the positron emission tomography reconstruction*, Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications (D. Butnariu, Y. Censor, and S. Reich, eds.), Studies Comput. Math., Elsevier, 2001.

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

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

[5] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Mathematical Programming **62** (1993), 261–271.

[6] V.F. Demyanov and L.V. Vasilyev, *Nondifferentiable optimization*, Optimization Software, Inc., New York, 1985.

[7] Y.M. Ermoliev, *Methods for solving nonlinear extremal problems*, Kibernetika (1969), no. 4, 1–17.

[8] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms*, Springer-Verlag, Berlin, 1996.

[9] K.C. Kiwiel, T. Larsson, and P.O. Lindberg, *Lagrangian relaxation via ballstep subgradient methods*, Mathematics of Operations Research **32** (2007), no. 3, 669–686.

[10] K.C. Kiwiel and P.O. Lindberg, *Parallel subgradient methods for convex optimization*, Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications (D. Butnariu, Y. Censor, and S. Reich, eds.), Studies Comput. Math., Elsevier, 2001.

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

[12] T. Larsson, M. Patriksson, and A. Strömberg, *Ergodic results and bounds on the optimal value in subgradient optimization*, Operations Research Proceedings (P. Kelinschmidt et al., ed.), Springer, 1995, pp. 30–35.

[13] T. Larsson, M. Patriksson, and A. Strömberg, *Ergodic convergence in subgradient optimization*, Optimization Methods and Software **9** (1998), 93–120.

[14] ______, *Ergodic primal convergence in dual subgradient schemes for convex programming*, Mathematical Programming **86** (1999), 283–312.

[15] S. Low and D.E. Lapsley, *Optimization flow control, I: Basic algorithm and convergence*, IEEE/ACM Transactions on Networking **7** (1999), no. 6, 861–874.

[16] A. Nedić and D.P. Bertsekas, *Convergence rate of incremental subgradient algorithms*, Stochastic Optimization: Algorithms and Applications (S. Uryasev and P. Pardalos, eds.), Kluwer Academic Publishers, 2000.

[17] ______, *Incremental subgradient methods for nondifferentiable optimization*, SIAM Journal on Optimization **56** (2001), no. 1, 109–138.

[18] A. Nedić, D.P. Bertsekas, and V.S. Borkar, *Distributed asynchronous incremental subgradient methods*, Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications (D. Butnariu, Y. Censor, and S. Reich, eds.), Studies Comput. Math., Elsevier, 2001.

[19] A.S. Nemirovskii and D.B. Judin, *Cezare convergence of gradient method approximation of saddle points for convex-concave functions*, Doklady Akademii Nauk SSSR **239** (1978), 1056–1059.

[20] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Tech. Report 67, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 2005.

[21] B.T. Polyak, *A general method for solving extremum problems*, Soviet Math. Doklady **8** (1967), no. 3, 593–597.

[22] ______, *Introduction to optimisation*, Optimization Software, Inc., New York, 1987.

[23] A. Ruszczyński, *A merit function approach to the subgradient method with averaging*, Optimization Methods and Software (2007), to appear.

[24] S. Sen and H.D. Sherali, *A class of convergent primal-dual subgradient algorithms for decomposable convex programs*, Mathematical Programming **35** (1986), 279–297.

[25] 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.

[26] N.Z. Shor, *Minimization methods for nondifferentiable functions*, Translated from Russian by K.C. Kiwiel and A. Ruszczynski, Springer, Berlin, 1985.

[27] R. Srikant, *Mathematics of Internet congestion control*, Birkhauser, 2004.

[28] H. Uzawa, *Iterative methods in concave programming*, Studies in Linear and Nonlinear Programming (K. Arrow, L. Hurwicz, and H. Uzawa, eds.), Stanford University Press, 1958, pp. 154–165.

[29] X. Zhao, P.B. Luh, and J. Wang, *Surrogate gradient algorithm for lagrangian relaxation*, Optimization Methods and Software **9** (1998), 93–120.