Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 151 (2003) 461–473
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/dsw

Continuous Optimization

On the convergence of conditional ε-subgradient methods for convex programs and convex–concave saddle-point problems

Torbjörn Larsson [a], Michael Patriksson [b,*], Ann-Brith Strömberg [c]

[a] Department of Mathematics, Linköping University, SE-581 83 Linköping, Sweden
[b] Department of Mathematics, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden
[c] Fraunhofer–Chalmers Research Center for Industrial Mathematics, SE-412 88 Gothenburg, Sweden

Received 9 November 2001; accepted 21 August 2002

Abstract

The paper provides two contributions. First, we present new convergence results for conditional ε-subgradient algorithms for general convex programs. The results obtained here extend the classical ones by Polyak [Sov. Math. Doklady 8 (1967) 593; USSR Comput. Math. Math. Phys. 9 (1969) 14; Introduction to Optimization, Optimization Software, New York, 1987] as well as the recent ones in [Math. Program. 62 (1993) 261; Eur. J. Oper. Res. 88 (1996) 382; Math. Program. 81 (1998) 23] to a broader framework. Secondly, we establish the application of this technique to solve non-strictly convex–concave saddle point problems, such as primal-dual formulations of linear programs. Contrary to several previous solution algorithms for such problems, a saddle-point is generated by a very simple scheme in which one component is constructed by means of a conditional ε-subgradient algorithm, while the other is constructed by means of a weighted average of the (inexact) subproblem solutions generated within the subgradient method. The convergence result extends those of [Minimization Methods for Non-Differentiable Functions, Springer-Verlag, Berlin, 1985; Oper. Res. Lett. 19 (1996) 105; Math. Program. 86 (1999) 283] for Lagrangian saddle-point problems in linear and convex programming, and of [Int. J. Numer. Meth. Eng. 40 (1997) 1295] for a linear–quadratic saddle-point problem arising in topology optimization in contact mechanics.
© 2002 Elsevier B.V. All rights reserved.

Keywords: Convex programming; Non-linear programming; Game theory; Large-scale optimization

1. Introduction

Consider the convex–concave saddle-point problem to find

$$(x^*, y^*) \in X \times Y: \quad \mathscr{L}(x^*, y) \leqslant \mathscr{L}(x^*, y^*) \leqslant \mathscr{L}(x, y^*), \quad \forall (x, y) \in X \times Y. \tag{SP}$$

We assume that $X \subset \mathfrak{R}^n$ and $Y \subset \mathfrak{R}^m$ are non-empty, convex and compact sets, and $\mathscr{L}: X \times Y \mapsto \mathfrak{R}$

* Corresponding author. Tel.: +46-31-7723529; fax: +46-31-161973.
E-mail addresses: tolar@mai.liu.se (T. Larsson), mipat@math.chalmers.se (M. Patriksson), ann-brith.stromberg@fcc.chalmers.se (A.-B. Strömberg).

0377-2217/$ - see front matter © 2002 Elsevier B.V. All rights reserved.
doi:10.1016/S0377-2217(02)00629-X

is convex–concave and finite (hence continuous) on $X \times Y$, that is, convex (concave) in $x(y)$ on $X(Y)$ for every fixed value of $y \in Y$ ($x \in X$). We note that in what is to follow, the compactness assumption on $X \times Y$ can be replaced by some coercivity assumption on $\mathscr{L}$ with respect to $X \times Y$ (e.g., [12, p. 334]).

Under the above assumptions on the problem (SP), there exists a saddle-point, (x^*, y^*), of $\mathscr{L}$ on $X \times Y$, the set of which is a Cartesian product which we will denote by $X^* \times Y^*$. Further, for any choice of $(x^*, y^*) \in X^* \times Y^*$,

$$v^* = \mathscr{L}(x^*, y^*) = \min_{x \in X} \max_{y \in Y} \mathscr{L}(x, y) = \max_{y \in Y} \min_{x \in X} \mathscr{L}(x, y).$$

(These results are collected, for example, in [12, Section VII.4].)

The algorithm to be presented in Section 2 attacks the problem by means of solving the following equivalent convex problem:

$$\underset{x \in X}{\text{minimize}} f(x), \tag{P}$$

where

$$f(x) := \max_{y \in Y} \mathscr{L}(x, y), \quad x \in X. \tag{1}$$

We shall denote the set of solutions to the problem (1) by $Y(x)$. An *ε-optimal* solution, $\tilde{y}$, to the problem (1) is characterized by the relation

$$\mathscr{L}(x, \tilde{y}) \geqslant f(x) - \varepsilon, \tag{2}$$

for some $\tilde{y} \in Y$ and $\varepsilon \geqslant 0$.

Example 1 (*convex programming*). An interesting application of (SP) is to convex programming, where $\mathscr{L}(x, y) := h(x) + y^{\mathrm{T}} g(x)$, corresponding to the Lagrangian of the problem to

$$\underset{x \in X \cap G}{\text{minimize}} h(x), \tag{CP}$$

where $h : \mathfrak{R}^n \mapsto \mathfrak{R}$ is a convex function, the convex set G is described by means of convex inequalities,

$$G := \{x \in \mathfrak{R}^n | g_i(x) \leqslant 0,\ i = 1, \ldots, m\},$$

where $g_i : \mathfrak{R}^n \mapsto \mathfrak{R}$ is convex for each i, and $Y = \mathfrak{R}^m_+$. For this problem, we assume that X is bounded, and the Slater constraint qualification that the set $\{x \in X | g(x) < 0\}$ is non-empty (see [4, Theorem 6.2.4]). Under this CQ, the set Y^* is compact. Assuming we may somehow restrict the set Y to be a convex and compact set including Y^*, we thereby fulfill all the conditions on the problem (SP). The problem (P) corresponds to the (convex) Lagrangian dual problem to maximize$_{y \in \mathfrak{R}^m_+} \theta(y)$, where $\theta(y) := \min_{x \in X} \mathscr{L}(x, y)$.

For solving the problem (P), we utilize *conditional ε-subgradient optimization*, which extends traditional subgradient optimization, as analyzed, for example, in [31], to possibly inexact calculations of subgradients and to generating search directions which take the feasible set X into account. (The latter extension of traditional subgradient optimization was analyzed in depth first in [19].) Thus, we generate a point in X^*. In order to generate a point in Y^*, we propose to build the sequence of weighted averages of the (possibly inexact) solutions to (1), generated while searching for a point in X^*. We establish that this sequence converges to the set Y^*, provided that the step lengths utilized in the process of finding a point in X^* by the subgradient algorithm, and the weights used in constructing a point in Y^*, are both chosen appropriately.

Some words on notation. The notation $\|\cdot\|$ denotes the Euclidean norm; for a non-empty, closed, and convex set $S \subseteq \mathfrak{R}^n$, the *normal cone* to S is

$$N_S(x) := \begin{cases} \{z \in \mathfrak{R}^n | z^{\mathrm{T}}(y - x) \leqslant 0,\ \forall y \in S\}, & x \in S, \\ \emptyset, & x \notin S; \end{cases}$$

the *indicator function* to S is

$$\psi_S(x) := \begin{cases} 0, & x \in S, \\ +\infty, & x \notin S. \end{cases}$$

We have that the subdifferential operator of ψ_S, $\partial\psi_S$, equals N_S. Further,

$$\operatorname{proj}(x, S) := \arg \min_{y \in S} \|y - x\|$$

denotes the (Euclidean) projection of the vector x onto the set S; we further introduce

$$\operatorname{dist}(x, S) := \min_{y \in S} \|y - x\|$$

to denote the Euclidean distance from the point x to its projection $\operatorname{proj}(x, S)$ onto S. Finally, we introduce, for any $v \geqslant 0$,

$X^v := \{x \in X | f(x) \leqslant f^* + v\}$,

that is, the lower level set of f corresponding to v-optimal solutions to the problem (P). (So, $X^0 = X^*$.)

The subject of Section 2 is the convergence of conditional ε-subgradient optimization algorithms. Section 3 presents the overall algorithm and establishes its convergence to a saddle-point.

2. Convergence of conditional ε-subgradient optimization

Our first result establishes a simple relationship between ε-optimal solutions to (1) given $x \in X$ and ε-subgradients of f at x. We first note that $\gamma_\varepsilon(x)$ is an *ε-subgradient* of f at x (that is, $\gamma_\varepsilon(x)$ is an element of the *ε-subdifferential*, $\partial_\varepsilon f(x)$, of f at x) for some $\varepsilon \geqslant 0$ if and only if

$$f(z) \geqslant f(x) + \gamma_\varepsilon(x)^{\mathrm{T}}(z - x) - \varepsilon, \quad z \in \Re^n; \tag{3}$$

the definition of a subgradient (that is, an element of the subdifferential) follows by setting $\varepsilon = 0$.

Proposition 2 (ε-optimal solutions provide ε-subgradients). *Suppose that, given $x \in X$, $\tilde{y}$ is an ε-optimal solution to* (1). *Then, any subgradient $\tilde{\gamma}(x)$ of $\mathscr{L}(\cdot, \tilde{y})$ at x is an ε-subgradient to f at x.*

Proof. Fix any $x \in X$ and $\varepsilon \geqslant 0$. For an arbitrary $z \in X$ then follows that

$$\begin{aligned} f(z) &\geqslant \mathscr{L}(z, \tilde{y}) \geqslant f(x) + [\mathscr{L}(z, \tilde{y}) - \mathscr{L}(x, \tilde{y}) - \varepsilon] \\ &\geqslant f(x) + \tilde{\gamma}(x)^{\mathrm{T}}(z - x) - \varepsilon, \end{aligned}$$

which yields that $\tilde{\gamma}(x) \in \partial_\varepsilon f(x)$, the first inequality coming from the definition of the value $f(z)$, the second inequality following from the ε-optimality of $\tilde{y}$ in (1), and the right-most inequality following from the convexity of $\mathscr{L}(\cdot, \tilde{y})$. □

Example 1 (*continued*). For the special case of the problem (CP), the above result states that an ε-optimal solution to the Lagrangian subproblem provides an ε-subgradient of the Lagrangian dual function θ. This property was described independently by Larsson et al. [21] and Bertsekas [5, p. 615], but is most probably folklore, and a much older result.

Given the iteration point x^t at iteration t, let $\gamma_{\varepsilon_t}(x^t)$ be an ε_t-subgradient of f at $x^t \in X$. Let $\gamma^X_{\varepsilon_t}(x^t)$ be a *conditional* ε_t-subgradient of f at $x^t \in X$, that is, $\gamma^X_{\varepsilon_t}(x^t) = \gamma_{\varepsilon_t}(x^t) + v(x^t)$ for some $v(x^t) \in N_X(x^t)$. (This is equivalent to replacing $z \in \Re^n$ with $z \in X$ in (3) or, in other words, $\gamma^X_{\varepsilon_t}(x^t)$ is an element of the ε_t-subdifferential of the function $f + \psi_X$ at x^t; see [9,19].) We will in the following analyze the convergence of conditional ε-subgradient algorithms for the solution of (P) using the divergent series step length rule,

$$\alpha_t > 0, \ \forall t, \quad \lim_{t\to\infty} \alpha_t = 0, \quad \text{and} \quad \sum_{t=0}^{\infty} \alpha_t = \infty, \tag{4}$$

in cases also under the additional requirement that

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty, \tag{5}$$

and under different scalings of the search directions.

In the analysis that follows, it is assumed that the sequences are infinite. In the case that $\gamma^X_{\varepsilon_t}(x^t) = 0$ for some t, x^t is ε_t-optimal in (P), and the procedure may be terminated (or the iteration considered void and the value of ε_t decreased).

For the sake of reaching a maximal generality, the analysis in this section for the problem (P) will ignore that its origin is the saddle-point problem discussed in the previous section, and hence assume temporarily that $f : \Re^n \mapsto \Re$ is a general convex function, and drop the assumption that the non-empty, closed and convex set X is necessarily bounded. Although we will study the problem only under the assumption that there exist optimal solutions to (P), we note that the algorithms described below are optimizing in the sense that $\liminf_{t\to\infty} f(x^t) = f^* := \inf_{x\in X} f(x)$ holds even if X^* is empty.

2.1. Divergent series step lengths, unscaled direction

We begin by considering unscaled directions.

The conditional ε-subgradient optimization method is given by

$$\begin{aligned} &x^{t+1/2} := x^t - \alpha_t \gamma^X_{\varepsilon_t}(x^t), \quad x^{t+1} := \mathrm{proj}(x^{t+1/2}, X), \\ &\quad t = 0, 1, \ldots, \end{aligned} \tag{6}$$

We note that the requirements of the algorithm are (a) that we have at hand a convergent algorithm for solving the problem (1), (b) a procedure for generating subgradients of $\mathscr{L}(\cdot, y^t_{\varepsilon_t})$, where $y^t_{\varepsilon_t}$ is an ε_t-optimal solution to the problem (1) given x^t and (c) that projections onto X are easily performed. In the case of the problem (CP) solved via its Lagrangian dual, the latter two requirements are of course trivial to fulfill.

Our main convergence result for the method (6), (4) establishes convergence to the optimal set X^*.

Theorem 3 (convergence to the optimal set using divergent series step lengths). *Let* $\{x^t\}$ *be generated by the method* (6), (4) *applied to* (P). *If* X^* *is bounded*, $\mathfrak{R}_+ \ni \{\varepsilon_t\} \to 0$, *and the sequence* $\{\gamma^X_{\varepsilon_t}(x^t)\}$ *is bounded, then* $\{f(x^t)\} \to f^*$ *and* $\{\operatorname{dist}(x^t, X^*)\} \to 0$.

Proof. Let $\delta > 0$ and $B^\delta = \{x \in \mathfrak{R}^n \mid \|x\| \leqslant \delta\}$. Since f is convex, X is non-empty, closed and convex, and X^* is bounded, it follows (from [28, Theorem 27.2], applied to the lower semicontinuous, proper and convex function $f + \psi_X$) that there exist $\epsilon = \epsilon(\delta) > 0$ and $\sigma = \sigma(\delta) > 0$ such that the lower level set $X^{\epsilon(1+\sigma)} \subseteq X^* + B^{\delta/2}$. Moreover, since $\{\gamma^X_{\varepsilon_t}(x^t)\}$ is bounded and $\{\alpha_t\} \to 0$, there exists an $N(\delta)$ such that $\alpha_t \|\gamma^X_{\varepsilon_t}(x^t)\|^2 \leqslant \epsilon$, $\varepsilon_t \leqslant \sigma\epsilon$ and $\alpha_t \|\gamma^X_{\varepsilon_t}(x^t)\| \leqslant \delta/2$ for all $t \geqslant N(\delta)$.

The sequel of the proof is based on induction and is organized as follows. In the first part, we show that there exists a finite $t(\delta) \geqslant N(\delta)$ such that $x^{t(\delta)} \in X^* + B^\delta$. In the second part, we establish that if x^t belongs to $X^* + B^\delta$ for some $t \geqslant N(\delta)$ then so does x^{t+1}; this is done by showing that $\operatorname{dist}(x^{t+1}, X^*) < \operatorname{dist}(x^t, X^*)$, or $x^t \in X^\epsilon$ so that $x^{t+1} \in X^* + B^\delta$ since the step taken is not longer than $\delta/2$.

Let $x^* \in X^*$ be arbitrary. In every iteration t we then have

$$\begin{aligned}\left\|x^* - x^{t+1}\right\|^2 &= \left\|x^* - \operatorname{proj}\left(x^t - \alpha_t \gamma^X_{\varepsilon_t}(x^t), X\right)\right\|^2 \\ &\leqslant \left\|x^* - x^t + \alpha_t \gamma^X_{\varepsilon_t}(x^t)\right\|^2 \\ &= \|x^* - x^t\|^2 + \alpha_t\Big(2\gamma^X_{\varepsilon_t}(x^t)^{\mathrm{T}}(x^* - x^t) \\ &\quad + \alpha_t \|\gamma^X_{\varepsilon_t}(x^t)\|^2\Big), \end{aligned} \tag{7}$$

where the inequality follows from the projection property. Now, suppose that

$$2\gamma^X_{\varepsilon_s}(x^s)^{\mathrm{T}}(x^* - x^s) + \alpha_s \|\gamma^X_{\varepsilon_s}(x^s)\|^2 < -\epsilon \tag{8}$$

for all $s \geqslant N(\delta)$. Then, using (7) repeatedly, we obtain that for any $t \geqslant N(\delta)$,

$$\|x^* - x^{t+1}\|^2 < \|x^* - x^{N(\delta)}\|^2 - \epsilon \sum_{s=N(\delta)}^{t} \alpha_s,$$

and from (4) it follows that the right-hand side of this inequality tends to minus infinity as $t \to \infty$, which clearly is impossible. Therefore,

$$2\gamma^X_{\varepsilon_t}(x^t)^{\mathrm{T}}(x^* - x^t) + \alpha_t \|\gamma^X_{\varepsilon_t}(x^t)\|^2 \geqslant -\epsilon \tag{9}$$

holds for at least one $t \geqslant N(\delta)$, say $t = t(\delta)$. From the definition of $N(\delta)$, it follows that $\gamma^X_{\varepsilon_{t(\delta)}}(x^{t(\delta)})^{\mathrm{T}}(x^* - x^{t(\delta)}) \geqslant -\epsilon$. By convexity we have that $f(x^*) - f(x^{t(\delta)}) \geqslant \gamma^X_{\varepsilon_{t(\delta)}}(x^{t(\delta)})^{\mathrm{T}}(x^* - x^{t(\delta)}) - \varepsilon_{t(\delta)}$, since $x^*, x^{t(\delta)} \in X$. Hence, $f(x^{t(\delta)}) \leqslant f^* + \epsilon + \varepsilon_{t(\delta)}$, that is, $x^{t(\delta)} \in X^{\epsilon+\varepsilon_{t(\delta)}} \subseteq X^{\epsilon(1+\sigma)} \subseteq X^* + B^{\delta/2} \subset X^* + B^\delta$.

Now, suppose that $x^t \in X^* + B^\delta$ for some $t \geqslant N(\delta)$. If (8) holds, then, using (7), we have that $\|x^* - x^{t+1}\| < \|x^* - x^t\|$ for any $x^* \in X^*$. Hence,

$$\begin{aligned}\operatorname{dist}(x^{t+1}, X^*) &\leqslant \|\operatorname{proj}(x^t, X^*) - x^{t+1}\| \\ &< \|\operatorname{proj}(x^t, X^*) - x^t\| \\ &= \operatorname{dist}(x^t, X^*) \leqslant \delta.\end{aligned}$$

Thus, $x^{t+1} \in X^* + B^\delta$. Otherwise, (9) must hold and, using the same arguments as above, we obtain that $f(x^t) \leqslant f^* + \epsilon + \varepsilon_t \leqslant f^* + \epsilon(1+\sigma)$, i.e., $x^t \in X^{\epsilon(1+\sigma)} \subseteq X^* + B^{\delta/2}$. As

$$\begin{aligned}\|x^{t+1} - x^t\| &= \|\operatorname{proj}(x^t - \alpha_t \gamma^X_{\varepsilon_t}(x^t), X) - x^t\| \\ &\leqslant \|x^t - \alpha_t \gamma^X_{\varepsilon_t}(x^t) - x^t\| = \alpha_t \|\gamma^X_{\varepsilon_t}(x^t)\| \leqslant \frac{\delta}{2}\end{aligned}$$

whenever $t \geqslant N(\delta)$, it follows that $x^{t+1} \in X^* + B^{\delta/2} + B^{\delta/2} = X^* + B^\delta$.

By induction with respect to $t \geqslant t(\delta)$, it follows that $x^t \in X^* + B^\delta$ for all $t \geqslant t(\delta)$. Since this holds for arbitrarily small values of $\delta > 0$ and f is continuous, the theorem follows. □

Remark 4 (*on the convergence conditions*). From the proof, the requirement that $\{\gamma^X_{\varepsilon_t}(x^t)\}$ is bounded can be replaced by the weaker requirement that $\{\alpha_t \|\gamma^X_{\varepsilon_t}(x^t)\|^2\} \to 0$ holds. Further, if X is

bounded, and not just the set X^*, then the sequence $\{\gamma_{\varepsilon_t}(x^t)\}$ is bounded automatically, while the sequence $\{v(x^t)\}$ may always be constructed so that it is bounded. For more details on the possible choices of this sequence, we refer to [19].

Remark 5 (*relations*). With $\varepsilon_t = 0$, Theorem 3 reduces to a result by Larsson et al. [19]. Further letting $v^t = 0^n$ reduces the algorithm to traditional subgradient optimization, and the result to one by Ermol'ev [10, Section 9].

2.2. Divergent series step lengths, scaled direction

The scaled conditional ε-subgradient optimization method is given by

$$x^{t+1/2} := x^t - \alpha_t \frac{\gamma^X_{\varepsilon_t}(x^t)}{\|\gamma^X_{\varepsilon_t}(x^t)\|},$$
$$x^{t+1} = \mathrm{proj}(x^{t+1/2}, X), \quad t = 0, 1, \ldots, \tag{10}$$

given some rule for choosing $\{\alpha_t\}$.

This scaling of the search direction allows us to remove the condition that the sequence $\{\gamma^X_{\varepsilon_t}(x^t)\}$ is bounded.

Theorem 6 (convergence to the optimal set using divergent series step lengths). *Let* $\{x^t\}$ *be generated by the method* (10), (4) *applied to* (P). *If* X^* *is bounded and* $\Re_+ \ni \{\varepsilon_t\} \to 0$, *then* $\{f(x^t)\} \to f^*$ *and* $\{\mathrm{dist}(x^t, X^*)\} \to 0$.

Proof. The proof technique is similar to that of Theorem 3. We define $\mu_t := \|\gamma^X_{\varepsilon_t}(x^t)\|$. The definition of $N(\delta)$ is here altered to mean that for all $t \geqslant N(\delta)$, $\alpha_t \leqslant \epsilon$, $\varepsilon_t \leqslant \sigma\epsilon$ and $\alpha_t \leqslant \delta/2$.

The inequality (7) is here replaced by

$$\|x^* - x^{t+1}\|^2 \leqslant \|x^* - x^t\|^2 + \alpha_t\left(\frac{2}{\mu_t}\gamma^X_{\varepsilon_t}(x^t)^{\mathrm{T}}(x^* - x^t) + \alpha_t\right),$$

and, consequently, (8) by

$$\frac{2}{\mu_t}\gamma^X_{\varepsilon_s}(x^s)^{\mathrm{T}}(x^* - x^s) + \alpha_s < -\epsilon.$$

We conclude as in the previous proof that

$$\frac{2}{\mu_t}\gamma^X_{\varepsilon_t}(x^t)^{\mathrm{T}}(x^* - x^t) + \alpha_t \geqslant -\epsilon$$

holds for at least one $t \geqslant N(\delta)$, say $t = t(\delta)$, which implies that $\gamma^X_{\varepsilon_{t(\delta)}}(x^{t(\delta)})^{\mathrm{T}}(x^* - x^{t(\delta)}) \geqslant -\epsilon\mu_{t(\delta)}$, and, by convexity, that $f(x^{t(\delta)}) \leqslant f^* + \epsilon\mu_{t(\delta)} + \varepsilon_{t(\delta)}$, that is, $x^{t(\delta)} \in X^{\epsilon\mu_{t(\delta)}+\varepsilon_{t(\delta)}} \subseteq X^{\epsilon(\mu_{t(\delta)}+\sigma)} \subseteq X^* + B^{\delta/2} \subset X^* + B^{\delta}$.

The rest of the proof follows as in the proof of Theorem 3, noting that

$$\begin{aligned}\|x^{t+1} - x^t\| &= \left\|\mathrm{proj}\left(x^t - \alpha_t\frac{\gamma^X_{\varepsilon_t}(x^t)}{\mu_t}, X\right) - x^t\right\| \\ &\leqslant \left\|x^t - \alpha_t\frac{\gamma^X_{\varepsilon_t}(x^t)}{\mu_t} - x^t\right\| \\ &= \alpha_t \leqslant \frac{\delta}{2}, \quad t \geqslant t(\delta).\end{aligned}$$

The result follows. □

Remark 7 (*relations*). With $v^t = 0^n$, the result reduces essentially to one by Polyak [25;27, pp. 144–145] (the first one also assumes that $\varepsilon_t = 0$, while the second one also assumes that $X = \Re^n$). Convergence is there established only for the sequence $\{f(x^t)\}$. In [1,32], convergence results are established for a subgradient algorithm (still assuming that $v^t = 0^n$ holds), where the search direction is given by $-(\gamma(x^t) + r^t)$, where $\{r^t\} \subset \Re^n$ is a sequence of error vectors which tends to zero [1] or stays bounded [32].

2.3. Quadratically convergent step lengths, non-scaled direction

We now introduce the additional requirement that (5) holds. As can be seen from the proof of the below theorem, this step length condition implies the boundedness of the sequence of iterates, whence that boundedness condition, present in Theorem 3, here can be removed.

Theorem 8 (convergence to an optimal solution using divergent series step lengths). *Let* $\{x^t\}$ *be generated by the method* (6), (4), (5) *applied to* (P). *If* $\Re_+ \ni \{\varepsilon_t\} \to 0$, *the sequence* $\{\gamma^X_{\varepsilon_t}(x^t)\}$ *is bounded, and if* $\sum_{s=0}^{\infty}\alpha_s\varepsilon_s < \infty$, *then* $\{x^t\}$ *converges to an element of* X^*.

Proof. Let $x^* \in X^*$. Define $\mu_t := \|\gamma^X_{\varepsilon_t}(x^t)\|$. In every iteration t we have that

$$\begin{aligned}\|x^* - x^{t+1}\|^2 &= \left\|x^* - \operatorname{proj}\left(x^t - \alpha_t \gamma^X_{\varepsilon_t}(x^t), X^*\right)\right\|^2 \\ &\leqslant \|x^* - x^t + \alpha_t \gamma^X_{\varepsilon_t}(x^t)\|^2 = \|x^* - x^t\|^2 \\ &\quad + \alpha_t\left(2\gamma^X_{\varepsilon_t}(x^t)^{\mathrm{T}}(x^* - x^t) + \alpha_t \mu_t^2\right),\end{aligned} \tag{11}$$

where the inequality follows from the projection property. Repeated application of (11) yields that

$$\|x^* - x^t\|^2 \leqslant \|x^* - x^0\|^2 + 2\sum_{s=0}^{t-1} \alpha_s \gamma^X_{\varepsilon_s}(x^s)^{\mathrm{T}}(x^* - x^s) + \sum_{s=0}^{t-1} \alpha_s^2 \mu_s^2. \tag{12}$$

Since $x^* \in X^*$ and $\gamma^X_{\varepsilon_s}(x^s) \in \partial^X_{\varepsilon_s} f(x^s)$ for all $s \geqslant 0$ we obtain that

$$f(x^s) \geqslant f^* \geqslant f(x^s) + \gamma^X_{\varepsilon_s}(x^s)^{\mathrm{T}}(x^* - x^s) - \varepsilon_s, \quad s \geqslant 0, \tag{13}$$

and hence that $\gamma^X_{\varepsilon_s}(x^s)^{\mathrm{T}}(x^* - x^s) \leqslant \varepsilon_s$ for all $s \geqslant 0$. We define $c := \sup_t\{\mu_t\}$, $p := \sum_{t=0}^{\infty} \alpha_t^2$ and $d := \sum_{s=0}^{\infty} \alpha_s \varepsilon_s$. From (12) we then conclude that $\|x^* - x^t\|^2 < \|x^* - x^0\|^2 + pc^2 + 2d$ for any $t \geqslant 1$, and thus that the sequence $\{x^t\}$ is bounded.

Assume now that there is no subsequence $\{x^{t_i}\}$ of $\{x^t\}$ with $\{\gamma^X_{\varepsilon_{t_i}}(x^{t_i})^{\mathrm{T}}(x^* - x^{t_i})\} \to 0$. Then there must exist an $\epsilon > 0$ with $\gamma^X_{\varepsilon_s}(x^s)^{\mathrm{T}}(x^* - x^s) \leqslant -\epsilon$ for all sufficiently large values of s. From (12) and the conditions on the step lengths it follows that $\{\|x^* - x^t\|\} \to -\infty$, which clearly is impossible. The sequence $\{x^t\}$ must therefore contain a subsequence $\{x^{t_i}\}$ such that $\{\gamma^X_{\varepsilon_{t_i}}(x^{t_i})^{\mathrm{T}}(x^* - x^{t_i})\} \to 0$. From (13) and the assumption that $\{\varepsilon_t\} \to 0$ it follows that $\{f(x^{t_i})\} \to f^*$. The boundedness of $\{x^t\}$ implies the existence of an accumulation point of $\{x^{t_i}\}$, say x^∞. From the continuity of f follows that $x^\infty \in X^*$.

To show that x^∞ is the only accumulation point of $\{x^t\}$, let $\delta > 0$ and let $M(\delta)$ be such that $\|x^\infty - x^{M(\delta)}\|^2 \leqslant \delta/3$, $\sum_{s=M(\delta)}^{\infty} \alpha_s^2 \leqslant \delta/(3c^2)$ and $\sum_{s=M(\delta)}^{\infty} \times \alpha_s \varepsilon_s \leqslant \delta/6$. Consider any $t > M(\delta)$. Analogously to the derivation of (12), and using (13), we then obtain that

$$\begin{aligned}\|x^\infty - x^t\|^2 &\leqslant \|x^\infty - x^{M(\delta)}\|^2 + \sum_{s=M(\delta)}^{t-1} \alpha_s^2 \mu_s^2 + 2\sum_{s=M(\delta)}^{t-1} \alpha_s \varepsilon_s \\ &< \frac{\delta}{3} + \frac{\delta}{3c^2}c^2 + \frac{2\delta}{6} = \delta.\end{aligned}$$

Since this holds for arbitrarily small values of $\delta > 0$, the theorem follows. □

Remark 9 (*relations*). With $v^t = 0^n$ and $X = \Re^n$, the result reduces to one in [6]. With $\varepsilon_t = 0$ the result reduces to one obtained in [19]. Under the assumption that $X = \Re^n$, Polyak [27, Section 5.5] establishes the convergence of inexact subgradient algorithms where the search direction is $-(\gamma(x^t) + r^t)$, and where the error sequence $\{r^t\}$ tends to zero, stays bounded or is some random sequence of vectors with bounded variance.

2.4. Quadratically convergent step lengths, scaled direction

When introducing a scaling of the step direction in the case of ε-subgradients and the use of the quadratically convergent step length rule (5), not only do we need to introduce the condition that $\sum_{s=0}^{\infty} \alpha_s \varepsilon_s < \infty$ holds (as in Section 2.3), but we also need to ensure that the length of the step direction does not tend to zero too quickly. Hence, the norm of the direction vector is projected onto the half-line $\{\ell | \ell \geqslant 1\}$. We further note that the scaling again allows us to remove the condition that $\{\gamma^X_{\varepsilon_t}(x^t)\}$ is bounded.

The scaled conditional ε-subgradient optimization method is given by

$$\begin{aligned}x^{t+1/2} &:= x^t - \alpha_t \frac{\gamma^X_{\varepsilon_t}(x^t)}{\max\{1, \|\gamma^X_{\varepsilon_t}(x^t)\|\}}, \\ x^{t+1} &= \operatorname{proj}(x^{t+1/2}, X), \quad t = 0, 1, \ldots\end{aligned} \tag{14}$$

Theorem 10 (convergence to an optimal solution using divergent series step lengths). *Let* $\{x^t\}$ *be generated by the method* (14), (4), (5) *applied to* (P). *If* $\Re_+ \ni \{\varepsilon_t\} \to 0$ *and* $\sum_{s=0}^{\infty} \alpha_s \varepsilon_s < \infty$, *then* $\{x^t\}$ *converges to an element of* X^*.

Proof. The proof follows the same line of arguments as that of Theorem 8. Let $\mu_t := \|\gamma^X_{\varepsilon_t}(x^t)\|$ and

$\eta_t := \max\{1, \mu_t\}$. Let $x^* \in X^*$. Analogously to the proof of Theorem 8, we obtain for every t that

$$\|x^* - x^{t+1}\|^2 \leqslant \|x^* - x^t\|^2 + \frac{\alpha_t}{\eta_t}\left(2\gamma^X_{\varepsilon_t}(x^t)^{\mathrm{T}}(x^* - x^t) + \frac{\alpha_t \mu_t^2}{\eta_t}\right).$$

Used repeatedly, we obtain

$$\|x^* - x^t\|^2 \leqslant \|x^* - x^0\|^2 + 2\sum_{s=0}^{t-1} \frac{\alpha_s}{\eta_s}\gamma^X_{\varepsilon_s}(x^s)^{\mathrm{T}}(x^* - x^s) + \sum_{s=0}^{t-1} \frac{\alpha_s^2 \mu_s^2}{\eta_s^2}. \tag{15}$$

From (13) and (15) we then obtain that

$$\|x^* - x^t\|^2 \leqslant \|x^* - x^0\|^2 + 2\sum_{s=0}^{t-1} \frac{\alpha_s \varepsilon_s}{\eta_s} + \sum_{s=0}^{t-1} \frac{\alpha_s^2 \mu_s^2}{\eta_s^2} \leqslant \|x^* - x^0\|^2 + 2\sum_{s=0}^{t-1} \alpha_s \varepsilon_s + \sum_{s=0}^{t-1} \alpha_s^2,$$

so the sequence $\{x^t\}$ is bounded, from the assumptions on the step lengths.

That every accumulation point of $\{x^t\}$ is optimal then follows as in the proof of Theorem 8, using (15) in place of (12), and noting that $\mu_t/\eta_t \leqslant 1$ for all t.

To show that x^∞ is the only limit point of $\{x^t\}$, let $\delta > 0$ and let $M(\delta)$ be such that $\|x^\infty - x^{M(\delta)}\|^2 \leqslant \delta/3$, $\sum_{s=M(\delta)}^\infty \alpha_s^2 \leqslant \delta/3$ and $\sum_{s=M(\delta)}^\infty \alpha_s \varepsilon_s \leqslant \delta/6$. Consider any $t > M(\delta)$. Analogously to the derivation of (15), and using (13) and again noting that $\mu_t/\eta_t \leqslant 1$ for all t, we then obtain that

$$\|x^\infty - x^t\|^2 \leqslant \|x^\infty - x^{M(\delta)}\|^2 + \sum_{s=M(\delta)}^{t-1} \frac{\alpha_s^2 \mu_s^2}{\eta_s^2} + 2\sum_{s=M(\delta)}^{t-1} \frac{\alpha_s \varepsilon_s}{\eta_s} < \frac{\delta}{3} + \frac{\delta}{3} + \frac{2\delta}{6} = \delta.$$

Since this holds for arbitrarily small values of $\delta > 0$, this completes the proof. □

Remark 11 (*relations*). With $\varepsilon_t = 0$ the result reduces to one in [19]. With $v^t = 0^n$ the result reduces to ones previously reached by Polyak [25] [1] and Alber et al. [2] (where the condition that $\sum_{s=0}^\infty \alpha_s \varepsilon_s < \infty$ holds is replaced by the condition that $\varepsilon_t \leqslant \eta\alpha_t$, $\eta > 0$ for all t).

Having presented some generally useful results on the convergence of conditional ε-subgradient algorithms in constrained, convex optimization, we now turn to the solution of the saddle-point problem (SP) which will use a subset of the results obtained in this section.

3. Ergodic convergence to the set of saddle-points

In order to solve the saddle-point problem (SP) it is in general not enough to find any solution to the maximization problem over y with x fixed to a value $x^* \in X^*$, as discussed, for example, in [12, Remark VII.4.2.6], unless $\mathscr{L}$ is strictly concave in y. (The corresponding result in the case of the convex program (CP) is known as the non-coordinability phenomenon; see, e.g., [22].) We will establish that an ergodic (that is, averaged) sequence of the inexact solutions $y^t_{\varepsilon_t}$ to the problem (1), generated from the sequence x^t in the above subgradient algorithm, tends to Y^*. Before moving on to state and establish this result, we will however discuss some related algorithms.

Under an assumption that $\mathscr{L}$ is strictly convex–concave, it is known that f, defined in (1), is continuously differentiable, with $\nabla f(x) = \nabla_x \times \mathscr{L}(x, y)$ from Danskin's Theorem [7], y being the unique solution to (1) given x. In order to produce a saddle-point in this situation, it is enough to find the minimizer of f over X, and apply (1) to get the corresponding component y of the saddle-point. Algorithms of this type include the descent algorithm of Demyanov and Malozemov [8, p. 230] and the gradient projection algorithm of Zhu and Rockafellar [36]. The former reference also suggests adding a strictly convex–concave quadratic term in the absence of strict convex–concavity, in effect thus producing a proximal point-like algorithm. Kallio and Ruszczyński [14] (see also [11,13,23]) propose a gradient projection algorithm defined by the partial gradients of the function $\mathscr{L}$ evaluated at perturbed points. (We, however, do not assume $\mathscr{L}$ to be differentiable.)

[1] Polyak [26] established that this algorithm cannot be linearly convergent, and proceeded to introduce the step length rule that now bears his name.

Kiwiel [15] solves the saddle-point problem through the use of a bundle method for the problem (P), and establishes that a convergent component in y is generated automatically in the search-direction finding quadratic programming problems. As for the special case of the convex program (CP), one can also envisage applying column generation and/or cutting plane approaches, where coordinability is induced through the solution of restricted master problems; see [21] for more detailed discussions.

The inspiration for the algorithm proposed here is however much more simple approaches to the solution of the problem (CP) through the use of Lagrangian dualization, subgradient optimization, and the construction of an optimal primal solution as simple or weighted averages of the Lagrangian subproblem solutions. Thus, a saddle-point is generated without solving any auxiliary problems. The origin is Shor's [31, pp. 116–118] work on linear programming, followed by Larsson and Liu [16], Sherali and Choi [30], Barahona and Anbil [3] (still in the context of linear programming, and the latter reference lacking a convergence proof), Larsson et al. [17] for a special large-scale convex programming problem arising in transportation planning, and Larsson et al. [21] for the general convex program (CP). The first convergence result of this type for saddle-point problems *not* arising from a Lagrangian was given by Petersson and Patriksson [24], who studied a special such large-scale problem arising in contact mechanics.

In the above algorithms, it is assumed that the computations of the subproblems are performed exactly. We extend the scope of these algorithm not only to the more general convex–concave saddle-point problems, but also to possibly inexact solutions of the subproblems. In the context of the problem (CP) and subgradient methods, it has previously been shown in [5,34,35] that convergence to the solution to (P) can be achieved through the solution of inexact subproblem solutions, but primal convergence results that can be extracted from this development have not been studied previously.

The interest in inexact computations is perhaps the most pronounced in the solution of combinatorial optimization problems through Lagrangian dualization. (This area is also one where Lagrangian dualization is quite popular.) Although combinatorial optimization problems are not convex problems in general, and there may not exist a saddle point to such a Lagrangian, there does however exist a saddle point for the convexified problem associated with the combinatorial problem and its Lagrangian formulation, and the maximum value of the Lagrangian dual function is the optimal value of the convexified problem. (Some of this theory is outlined by Wolsey [33, Section 10.2].) If the relaxation does not satisfy the integrality property (so that the Lagrangian subproblem cannot be reduced to a linear program), then the Lagrangian subproblem will be a (potentially) computationally difficult combinatorial problem, and the use of inexact methods such as heuristics to solve them will therefore be of computational advantage; then, also, the resulting solution will provide an ε-subgradient of the Lagrangian dual function.

The scope is also extended to include the possible use of conditional subgradients. Although such an algorithm can, in principle, be incorporated into a standard subgradient algorithm for the extended objective function $f + \psi_X$, this function is not finite everywhere, and moreover, it has been found in the numerical investigations performed in [19] that adding a normal cone element to the search direction can substantially enhance convergence in practice for some difficult problems.

3.1. Preliminaries

The optimality conditions for x^* in (P) is given by

$$-\partial f(x^*) \cap N_X(x^*) \neq \emptyset \tag{16}$$

(e.g., [29, Proposition 5A and Equation (5.5)]). The non-coordinability phenomenon discussed above, which is inherent in every non-strictly convex–concave saddle-point problem, can be equivalently described as the failure of the entire set $-\partial f(x^*)$ to be included in $N_X(x^*)$, something which obviously does hold whenever f is differentiable at x^*. We will establish the convergence of an averaged (or, ergodic) sequence of subproblem

solutions by means of establishing that an averaged sequence of the ε_t-subgradients $\gamma_{\varepsilon_t}(x^t)$ accumulates at subgradients which verify optimality in accordance with (16). The representative algorithm which we have chosen among those in the previous section is that validated in Theorem 8.

The properties of ergodic sequences of elements generated in a subgradient scheme have previously been analyzed in [20]. We extend some of their analysis to the use of ε-subgradients.

Let

$$A_t := \sum_{s=0}^{t-1} \alpha_s, \tag{17}$$

$$\hat{y}^t := A_t^{-1} \sum_{s=0}^{t-1} \alpha_s y^s_{\varepsilon_s}, \tag{18}$$

$$g^t := A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \gamma_{\varepsilon_s}(x^s), \tag{19}$$

$$n^t_c := A_t^{-1} \sum_{s=0}^{t-1} \alpha_s v^s, \tag{20}$$

$$n^t_p := A_t^{-1} \sum_{s=0}^{t-1} (x^{s+1/2} - x^{s+1}), \tag{21}$$

$$n^t := n^t_c + n^t_p, \tag{22}$$

$$\varphi^t := A_t^{-1} \sum_{s=0}^{t-1} \alpha_s f(x^s), \tag{23}$$

$$f_t(x) := A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \left[f(x^s) + (\gamma_{\varepsilon_s}(x^s))^{\mathrm{T}}(x - x^s) - \varepsilon_s \right], \quad x \in X, \tag{24}$$

$$\delta_t(x) := f(x) - f_t(x), \quad x \in X, \tag{25}$$

$$\varrho_t(x) := A_t^{-1} \sum_{s=0}^{t-1} \left[\alpha_s (v^s)^{\mathrm{T}}(x^s - x) + \left(x^{s+1/2} - x^{s+1}\right)^{\mathrm{T}}(x^{s+1} - x) \right], \quad x \in X. \tag{26}$$

Here, A_t is the accumulated step length up to iteration t, $\hat{y}^t$ the weighted average of the inexact solutions to (1), g^t the weighted average of the ε_s-subgradients of f, and n^t_c and n^t_p ergodic normal elements and projection steps, respectively. We note that $\{\hat{y}^t\} \subset Y$, and this sequence is therefore bounded, by the boundedness assumption on Y. Continuing, $\varphi^t \geqslant v^*$ clearly holds. The affine function f_t is derived as a surrogate of the convexity inequality, and therefore $\delta_t(x) \geqslant 0$ on X. Further, since $v^s \in N_X(x^s)$ and $x^{s+1/2} - x^{s+1} \in N_X(x^{s+1})$, $\varrho_t(x) \geqslant 0$ on X, and thus defines a valid inequality for X.

We will establish that any accumulation point, $\hat{y}^\infty$, of the sequence $\{\hat{y}^t\}$ together with the solution x^∞ obtained from the subgradient scheme forms a saddle-point of $\mathscr{L}$.

Lemma 12 ($Y(\cdot)$ is a closed map). *Let the sequence* $\{x^t\} \subset X$, $\{\varepsilon_t\} \subset \mathfrak{R}_+$, *the map* $Y_\varepsilon(\cdot) : X \mapsto 2^Y$ *be given by the definition*

$$Y_\varepsilon(x) := \{y \in Y | \mathscr{L}(x, y) \geqslant f(x) - \varepsilon\}, \quad x \in X,$$

and the sequence $\{y^t_{\varepsilon_t}\}$ *be given by the inclusion* $y^t_{\varepsilon_t} \in Y_{\varepsilon_t}(x^t)$. *If the sequences* $\{x^t\} \to x$ *and* $\{\varepsilon_t\} \to 0$, *then* $\{\mathrm{dist}(y^t_{\varepsilon_t}, Y(x))\} \to 0$. *If, in addition,* $Y(x) = \{y\}$, *then* $\{y^t_{\varepsilon_t}\} \to y$.

Proof. By the definition of $Y_{\varepsilon_t}(x^t)$, $\mathscr{L}(x^t, y^t_{\varepsilon_t}) \geqslant f(x^t) - \varepsilon_t$ holds for all t. It follows from the continuity of the functions $\mathscr{L}$ and f, the compactness of X and the construction of the sequence $\{\varepsilon_t\}$ that $y \in Y(x)$ holds for any accumulation point y of the sequence $\{y^t_{\varepsilon_t}\}$. The result $\{\mathrm{dist}(y^t_{\varepsilon_t}, Y(x))\} \to 0$ then follows from the boundedness of the sequence $\{y^t_{\varepsilon_t}\}$. The second result is then immediate. □

3.2. Main results

We next utilize this result to establish that the sequence $\{\hat{y}^t\}$ accumulates in the set $Y(x^\infty)$, establishing the left-most inequality in (SP).

Theorem 13 ($\mathscr{L}(x^\infty, \hat{y}^\infty) \geqslant \mathscr{L}(x^\infty, y)$ for all $y \in Y$). $\{\mathrm{dist}(\hat{y}^t, Y(x^\infty))\} \to 0$.

Proof. Fix any $\epsilon > 0$. By Theorem 8 and Lemma 12, for any large enough τ,

$$\mathrm{dist}(y^s_{\varepsilon_s}, Y(x^\infty)) \leqslant \epsilon/2, \quad s \geqslant \tau.$$

By the convexity of the function $\mathrm{dist}(\cdot, Y(x^\infty))$ (e.g., [28, Theorem 4.3]), we have that

$$\mathrm{dist}(\hat{y}^t, Y(x^\infty)) \leqslant A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \mathrm{dist}(y^s_{\varepsilon_s}, Y(x^\infty)), \quad s \geqslant \tau.$$

Since $\{A_t\} \to \infty$ and τ is fixed, $A_t^{-1} \sum_{s=\tau}^{t-1} \alpha_s \mathrm{dist}(y^s_{\varepsilon_s}, Y(x^\infty)) \leqslant (1 - A_t^{-1} A_\tau)\epsilon/2$ holds for every $t > \tau$. Hence, $\mathrm{dist}(\hat{y}^t, Y(x^\infty)) \leqslant \epsilon$ holds for all $t > \tau$ that are large enough, and the desired result follows. □

The next result, which is a direct consequence of the definition (19) and the inequality $\|x^{s+1/2} - x^{s+1}\| \leqslant \alpha_s \|\gamma^X_{\varepsilon_s}(x^s)\|$, is the first step towards a convergence result for the ergodic sequence $\{g^t\}$.

Lemma 14. *The sequences $\{g^t\}$ and $\{n^t\}$ are bounded.*

The following lemma concerns the convergence properties of some of our ergodic sequences.

Lemma 15. $\{\delta_t(x^\infty)\} \to 0$, $\{\varrho_t(x^\infty)\} \to 0$, *and* $\{\varphi_t\} \to v^*$. *Further*, $\{\delta_t(x^t)\} \to 0$ *and* $\{\varrho_t(x^t)\} \to 0$.

Proof. By Theorem 8, $\{x^t\} \to x^\infty$. From the iteration formula (6) it follows that

$$\begin{aligned}\|x^{t+1} - x^\infty\|^2 \leqslant \|x^t - x^\infty\|^2 + \alpha_t^2 \|\gamma^X_{\varepsilon_t}(x^t)\|^2 \\ - 2\Big(\alpha_t \big(\gamma^X_{\varepsilon_t}(x^t)\big)^{\mathrm{T}} (x^t - x^\infty) \\ + \big(x^{t+1/2} - x^{t+1}\big)^{\mathrm{T}} \big(x^{t+1} - x^\infty\big)\Big).\end{aligned}$$

Repeated application of this inequality and utilizing the definitions (23), (25) and (26) result in

$$\begin{aligned}\|x^0 - x^\infty\|^2 + \sum_{s=0}^{t-1} \alpha_s^2 \|\gamma^X_{\varepsilon_s}(x^s)\|^2 \\ - 2A_t \left(\delta_t(x^\infty) + \varrho_t(x^\infty) + \varphi_t - v^* - A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \varepsilon_s \right) \geqslant 0.\end{aligned}$$

Since $\delta_t(x^\infty) \geqslant 0$, $\varrho_t(x^\infty) \geqslant 0$, $\varphi_t \geqslant f^*$ and $A_t > 0$, the immediate result is that

$$\begin{aligned}0 \leqslant \delta_t(x^\infty) + \varrho_t(x^\infty) + \varphi_t - f^* \\ \leqslant \frac{1}{2A_t} \left(\left\|x^0 - x^\infty\right\|^2 + \sum_{s=0}^{t-1} \left[\alpha_s^2 \left\|\gamma^X_{\varepsilon_s}(x^s)\right\|^2 + \alpha_s \varepsilon_s \right] \right).\end{aligned}$$

Let $t \to \infty$ and invoke the conditions of Theorem 8.

For the latter result, we note that the definitions (19), (24) and (25) yield

$$0 \leqslant \delta_t(x^t) = \delta_t(x^\infty) + f(x^t) - f(x^\infty) - (g^t)^{\mathrm{T}}(x^t - x^\infty), \quad t = 1, 2, \ldots$$

From the definitions (19) and (26) follow that

$$0 \leqslant \varrho_t(x^t) = \varrho_t(x^\infty) - (n^t)^{\mathrm{T}}(x^t - x^\infty), \quad t = 1, 2, \ldots$$

Theorem 8, Lemma 14, the continuity of f, and the first part of this Lemma yield that the right-hand sides of both the above equations tend to zero as t approaches infinity. The result follows. □

We will also utilize the following lemma in our continued analysis, when proving optimality fulfillment of the sequence $\{g^t\}$ in the limit.

Lemma 16. $\{g^t + n^t\} \to 0$.

Proof. By the definition (19) and the iteration formula (6),

$$\begin{aligned}g^t + n^t &= A_t^{-1} \sum_{s=0}^{t-1} (x^s - x^{s+1/2} + x^{s+1/2} - x^{s+1}) \\ &= A_t^{-1}(x^0 - x^t).\end{aligned}$$

Theorem 8 yields that $\{x^t\} \to x^\infty$. The result then follows from the definition (17) and the condition (4). □

We are now ready to establish that the ergodic sequence $\{g^t\}$ of subgradients accumulates at subgradients which verify optimality, according to (16). We divide the result into two parts, extending, respectively, Theorem 3.7 and Theorem 3.8 in [20] to the use of ε-subgradients.

Proposition 17 (convergence of $\{g^t\}$ to $\partial f(x^\infty)$). $\{\mathrm{dist}(g^t, \partial f(x^\infty))\} \to 0$.

Proof. The definitions (19), (24) and (25) imply that g^t is a $\delta_t(x)$-subgradient of f at any $x \in X$; applying this result to $x = x^t$ yields

$$f(y) \geqslant f(x^t) + (g^t)^{\mathrm{T}}(y - x^t) - \delta_t(x^t), \quad y \in X.$$

By Lemma 14, the sequence $\{g^t\}$ is bounded. Let $\tilde{g}$ be an accumulation point of $\{g^t\}$, corresponding to a convergent subsequence $\mathscr{T}$. Then, by Lemma 15, in the limit of $\mathscr{T}$, from the above inequality, we obtain that $\tilde{g} \in \partial f(x^\infty)$. The boundedness of $\partial f(x^\infty)$ then yields the desired result. □

Proposition 18 (convergence of $\{g^t\}$ to $-N_X(x^\infty)$). $\{\mathrm{dist}(g^t, -N_X(x^\infty))\} \to 0$.

Proof. From the definition (26) it follows that, for all t and any $z \in X$, $(n^t)^{\mathrm{T}}(z - x^\infty) = -\varrho_t(z) + \varrho_t(x^\infty) \leqslant \varrho_t(x^\infty)$. By Lemma 14, the sequence $\{n^t\}$ is bounded. Let $\tilde{n}$ be an accumulation point of the sequence $\{n^t\}_{t\in\mathscr{T}}$ for some convergent subsequence $\mathscr{T}$. From Lemma 15 it then follows that $\tilde{n}^{\mathrm{T}}(z - x^\infty) \leqslant 0$ for any $z \in X$. By the continuity of the function $\mathrm{dist}(\cdot, N_X(x^\infty))$, for any $\epsilon > 0$ and all t that are sufficiently large, $\mathrm{dist}(n^t, N_X(x^\infty)) < \epsilon/2$; further, by Lemma 16, $\|g^t + n^t\| < \epsilon/2$. This yields that $\mathrm{dist}(g^t, -N_X(x^\infty)) \leqslant \mathrm{dist}(-n^t, -N_X(x^\infty)) + \|g^t + n^t\| < \epsilon$, and the result follows. □

Theorem 13 established one of the inequalities in the definition of (SP); the second inequality now follows. (Note that the accumulation point $\hat{y}^\infty$ still is arbitrary.)

Theorem 19 ($\mathscr{L}(x, \hat{y}^\infty) \geqslant \mathscr{L}(x^\infty, \hat{y}^\infty)$ for all $x \in X$). $\{\mathrm{dist}(x^t, X(\hat{y}^\infty))\} \to 0$.

Proof. For every $x \in X$,

$$\begin{aligned}\mathscr{L}(x, \hat{y}^\infty) - \mathscr{L}(x^\infty, \hat{y}^\infty) &\geqslant f(x) - \mathscr{L}(x^\infty, \hat{y}^\infty)\\ &= f(x) - f(x^\infty)\\ &\geqslant \tilde{g}^{\mathrm{T}}(x - x^\infty) \geqslant 0,\end{aligned}$$

where the first inequality follows from the definition of $f(x)$ and the fact that $\hat{y}^\infty \in Y$, the equality from Theorem 13, and the two final inequalities from the convexity of f and, respectively, Propositions 17 and 18. □

We summarize the results of the Theorems 13 and 19 as follows:

Theorem 20 ($(x^\infty, \hat{y}^\infty)$ solves (SP)). $\{\mathrm{dist}((x^t, \hat{y}^t), \{x^\infty\} \times Y^*)\} \to 0$.

4. Further research

As outlined at the beginning of Section 3, among the possible application areas of this type of methods perhaps the most interesting one is to use them in order to generate an approximate solution to the convexification of an integer program. The ergodic sequence would then be terminated finitely, for example when the duality gap fails to be reduced significantly. The averaged solution, or the result of a primal feasibility heuristic applied from it, is there after used as a starting point for (or is embedded within) an algorithm devised to close the duality gap, such as a cutting plane or branch and bound algorithm. The use of this technique is well-known in circumstances when the relaxation has the integrality property, because it is then equivalent to solve the linear relaxation, but it has been tested only to a limited extent for more difficult Lagrangian subproblems. In theory, these subproblems must of course be solved exactly in the limit according to our convergence conditions, and this may not be practically feasible. Two of the authors of this article are currently investigating the theory and practice of using Lagrangian near-optimal solutions and ergodic sequences in computations in combinatorial optimization, combined with approximate solution methods using core problems and column generation [18].

The proofs of the results of the previous section relies on the essential element that the sequence $\{x^t\}$ converges. Moreover, the analysis at present utilizes rather heavily the condition (5) on the sequence of step lengths. It would be of interest for practical purposes to be able to avoid this condition, as the possibility to enable the use of the algorithm of Theorem 3 in place of that of Theorem 8 would also imply that the condition that $\sum_{s=0}^{\infty} \alpha_s \varepsilon_s < \infty$ holds can be removed, which in turn would allow for the subproblem solutions to be even less exact.

The possible use of the algorithms of Theorems 6 and 10, where the search directions are scaled, are left as a topic for further research, as are the possibilities to use other convexity weights in the construction of the ergodic sequence $\{\hat{y}^t\}$ as well as other step length rules in the subgradient algorithm.

Acknowledgements

The authors thank an anonymous referee for pertinent remarks which improved the presentation.

References

[1] Ya.I. Alber, Recurrence relations and variational inequalities, Soviet Mathematics Doklady 27 (1983) 511–517.
[2] Ya.I. Alber, A.N. Iusem, M.V. Solodov, On the projected subgradient method for nonsmooth convex optimization in a Hilbert space, Mathematical Programming 81 (1998) 23–35.
[3] F. Barahona, R. Anbil, The volume algorithm: Producing primal solutions with a subgradient method, Mathematical Programming 87 (2000) 385–399.
[4] M.S. Bazaraa, H.D. Sherali, C.M. Shetty, Nonlinear Programming: Theory and Algorithms, second ed., John Wiley & Sons, New York, NY, 1993.
[5] D.P. Bertsekas, Nonlinear Programming, second ed., Athena Scientific, Belmont, MA, 1999.
[6] R. Correa, C. Lemaréchal, Convergence of some algorithms for convex minimization, Mathematical Programming 62 (1993) 261–275.
[7] J.M. Danskin, The Theory of Max–Min, Springer-Verlag, Berlin, 1967.
[8] V.F. Demyanov, V.N. Malozemov, Introduction to Minimax, John Wiley & Sons, New York, NY, 1974.
[9] V.F. Dem'janov, V.K. Šomesova, Conditional subdifferentials of convex functions, Soviet Mathematics Doklady 19 (1980) 1181–1185.
[10] Yu.M. Ermol'ev, Methods for solving nonlinear extremal problems, Cybernetics 2 (1966) 1–17.
[11] D.W. Hearn, S. Lawphongpanich, Operations Research Letters 8 (1989) 189–196.
[12] J.-B. Hiriart-Urruty, C. Lemaréchal, Convex Analysis and Minimization Algorithms, I: Fundamentals, Springer-Verlag, Berlin, 1993.
[13] M. Kallio, C.H. Rosa, Large-scale convex optimization via saddle point computation, Operations Research 47 (1999) 93–101.
[14] M. Kallio, A. Ruszczyński, Perturbation methods for saddle point computation, Working paper WP-94-38, IIASA, Laxenburg, Austria.
[15] K.C. Kiwiel, Proximal level bundle methods for convex nondifferentiable optimization, saddle-point problems and variational inequalities, Mathematical Programming 69 (1995) 89–109.
[16] T. Larsson, Z. Liu, A Lagrangian relaxation scheme for structured linear programs with application to multicommodity network flows, Optimization 40 (1997) 247–284.
[17] T. Larsson, Z. Liu, M. Patriksson, A dual scheme for traffic assignment problems, Optimization 42 (1997) 323–358.
[18] T. Larsson, M. Patriksson, Global optimality conditions and Lagrangian heuristics for nonconvex optimization, report, Department of Mathematics, Chalmers University of Technology, Gothenburg, in preparation.
[19] T. Larsson, M. Patriksson, A.-B. Strömberg, Conditional subgradient optimization—theory and applications, European Journal of Operational Research 88 (1996) 382–403.
[20] T. Larsson, M. Patriksson, A.-B. Strömberg, Ergodic convergence in subgradient optimization, Optimization Methods & Software 9 (1998) 93–120.
[21] T. Larsson, M. Patriksson, A.-B. Strömberg, Ergodic, primal convergence in dual subgradient schemes for convex programming, Mathematical Programming 86 (1999) 283–312.
[22] L.S. Lasdon, Optimization Theory for Large Systems, Macmillan, New York, 1970.
[23] A. Ouorou, A primal–dual algorithm for monotropic programming and its application to network optimization, Computational Optimization and Applications 15 (2000) 125–143.
[24] J. Petersson, M. Patriksson, Topology optimization of sheets in contact by a subgradient method, International Journal of Numerical Methods in Engineering 40 (1997) 1295–1321.
[25] B.T. Polyak, A general method of solving extremum problems, Soviet Mathematics Doklady 8 (1967) 593–597.
[26] B.T. Polyak, Minimization of unsmooth functionals, USSR Computational Mathematics and Mathematical Physics 9 (1969) 14–29.
[27] B.T. Polyak, Introduction to Optimization, Optimization Software New York, 1987.
[28] R.T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, NJ, 1970.
[29] R.T. Rockafellar, The Theory of Subgradients and its Applications to Problems of Optimization: Convex and Nonconvex Functions, Heldermann Verlag, Berlin, 1981.
[30] H.D. Sherali, 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.
[31] N.Z. Shor, Minimization Methods for Non-Differentiable Functions, Springer-Verlag, Berlin, 1985.
[32] M.V. Solodov, S.K. Zavriev, Error stability properties of generalized gradient-type algorithms, Journal of Optimization Theory and Applications 98 (1998) 663–680.
[33] L.A. Wolsey, Integer Programming, John Wiley & Sons, New York, 1998.
[34] X. Zhao, P.B. Luh, J. Wang, New bundle methods for solving Lagrangian relaxation dual problems, Journal of Optimization Theory and Applications 113 (2002) 373–397.

[35] X. Zhao, P.B. Luh, J. Wang, Surrogate gradient algorithm for Lagrangian relaxation, Journal of Optimization Theory and Applications 100 (1999) 699–712.

[36] C. Zhu, R.T. Rockafellar, Primal–dual projected gradient algorithms for extended linear–quadratic programming, SIAM Journal on Optimization 3 (1993) 751–783.