ERGODIC RESULTS IN SUBGRADIENT OPTIMIZATION*

Torbjörn Larsson,[1] Michael Patriksson,[2] and Ann-Brith Strömberg[3]

[1,3] Division of Optimization
Department of Mathematics
Linköping Institute of Technology
S-581 83 Linköping, Sweden

[1] e-mail: tl@math.liu.se

[3] e-mail: anstr@math.liu.se

[2] Department of Mathematics, Box 354350
University of Washington
Seattle, WA 98195-4350
e-mail: patrikss@math.washington.edu or mipat@math.liu.se

Abstract: Subgradient methods are popular tools for nonsmooth, convex minimization, especially in the context of Lagrangean relaxation; their simplicity has been a main contribution to their success. As a consequence of the nonsmoothness, it is not straightforward to monitor the progress of a subgradient method in terms of the approximate fulfilment of optimality conditions, since the subgradients used in the method will, in general, *not* accumulate to subgradients that verify optimality of a solution obtained in the limit. Further, certain supplementary information, such as convergent estimates of Lagrange multipliers, is not directly available in subgradient schemes.

As a means for overcoming these weaknesses of subgradient optimization methods, we introduce the computation of an ergodic (averaged) sequence of subgradients. Specifically, we consider a nonsmooth, convex program solved by a conditional subgradient optimization scheme (of which the traditional subgradient optimization method is a special case) with divergent series step lengths, which generates a sequence of iterates that converges to an optimal solution. We show that the elements of the ergodic sequence of subgradients in the limit fulfil the optimality conditions at this optimal solution. Further, we use the convergence properties of the ergodic sequence of subgradients to establish convergence of an ergodic sequence of Lagrange multipliers. Finally, some potential applications of these ergodic results are briefly discussed.

Key Words: Nonsmooth minimization, Conditional subgradient optimization, Ergodic sequences, Lagrange multipliers.

1 INTRODUCTION

1.1 Background and motivation

Let the function $f : \Re^n \mapsto \Re$ be nonsmooth and convex and thus everywhere continuous and subdifferentiable. Further, let $X \subseteq \Re^n$ be a nonempty, closed and convex set, and assume that $\inf_{\mathbf{x} \in X} f(\mathbf{x})$ is attained at some $\mathbf{x} \in X$. These properties are assumed to hold throughout the paper. The program considered is

[P]

$$f^* = \min_{\mathbf{x} \in X} f(\mathbf{x}).$$

Defining the *lower level set* of the function f as

$$L_f(z) = \{ \mathbf{x} \in \Re^n \mid f(\mathbf{x}) \leq z \}, \quad z \in \Re,$$

the *solution set* of the program [P] is the nonempty, closed and convex set $X^* = X \cap L_f(f^*)$.

To enable the statement of optimality conditions for the program [P], we first give some basic definitions (e.g., Rockafellar, 1970a, and Hiriart-Urruty and Lemaréchal, 1993a). The *subdifferential* of f at $\mathbf{x} \in \Re^n$ is the nonempty, convex and compact set

$$\partial f(\mathbf{x}) = \left\{ \boldsymbol{\gamma} \in \Re^n \;\middle|\; f(\mathbf{y}) \geq f(\mathbf{x}) + \boldsymbol{\gamma}^T(\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{y} \in \Re^n \right\}, \tag{1.1}$$

the elements of which are called *subgradients*. The *normal cone* to the set X at some $\mathbf{x} \in X$ is (e.g., Lemaréchal and Hiriart-Urruty, 1993a, Definition III:5.2.3) the set

$$N_X(\mathbf{x}) = \left\{ \boldsymbol{\nu} \in \Re^n \;\middle|\; \boldsymbol{\nu}^T(\mathbf{y} - \mathbf{x}) \leq 0, \quad \forall \mathbf{y} \in X \right\}, \tag{1.2}$$

which is nonempty, closed, and convex. The *tangent cone* to the set X at $\mathbf{x} \in X$ (ibid., Definition III:5.1.1) may be expressed as the *polar cone* to $N_X(\mathbf{x})$ (ibid., Corollary III:5.2.5), that is, the set

$$T_X(\mathbf{x}) = \left\{ \boldsymbol{\tau} \in \Re^n \;\middle|\; \boldsymbol{\tau}^T \boldsymbol{\nu} \leq 0, \quad \forall \boldsymbol{\nu} \in N_X(\mathbf{x}) \right\}, \tag{1.3}$$

which is nonempty, closed and convex. Conversely, $N_X(\mathbf{x})$ is the polar cone to $T_X(\mathbf{x})$ (ibid., Proposition III:5.2.4). If $\mathbf{x} \in \mathtt{int}\, X$, then $N_X(\mathbf{x}) = \{\mathbf{0}\}$ and $T_X(\mathbf{x}) = \Re^n$.

Optimality conditions for the program [P] are given in the next proposition (ibid., Theorem VII:1.1.1 and Condition VII:(1.1.3)). Here, and in the following,

$$P_S(\mathbf{v}) = \arg\min_{\mathbf{u} \in S} \| \mathbf{u} - \mathbf{v} \|$$

denotes the *Euclidean projection* of a vector $\mathbf{v} \in \Re^n$ onto a nonempty, closed and convex set $S \subseteq \Re^n$.

Proposition 1.1 (Optimality conditions). $\mathbf{x} \in X^*$ *if and only if there exists a* $\boldsymbol{\gamma} \in \partial f(\mathbf{x})$ *such that* $P_{-T_X(\mathbf{x})}(\boldsymbol{\gamma}) = \mathbf{0}$ *or, equivalently,* $\partial f(\mathbf{x}) \cap -N_X(\mathbf{x}) \neq \emptyset$. □

The next proposition states important properties of the set $\partial f(\mathbf{x}) \cap -N_X(\mathbf{x})$, $\mathbf{x} \in X^*$, and follows from the results given above and a result of Burke and Ferris (1991, Lemma 2). It will be useful later in our analysis.

Proposition 1.2 *For any* $\mathbf{x} \in X^*$*, the nonempty set* $\partial f(\mathbf{x}) \cap -N_X(\mathbf{x})$ *is convex and compact, and* $\partial f(\mathbf{y}) \cap -N_X(\mathbf{y}) = \partial f(\mathbf{x}) \cap -N_X(\mathbf{x})$ *for all* $\mathbf{y} \in X^*$*.* $\square$

The following proposition, which is easily established, gives a geometric interpretation of the elements of the set $\partial f(\mathbf{x}) \cap -N_X(\mathbf{x})$, $\mathbf{x} \in X^*$.

Proposition 1.3 (Separation). *Let* $\mathbf{x} \in X^*$*. Then, the hyperplane*

$$H_{\boldsymbol{\gamma}}(\mathbf{x}) = \left\{ \mathbf{y} \in \Re^n \;\middle|\; \boldsymbol{\gamma}^T(\mathbf{y} - \mathbf{x}) = 0 \right\}$$

separates X *and* $L_f(f^*)$ *if and only if* $\boldsymbol{\gamma} \in \partial f(\mathbf{x}) \cap -N_X(\mathbf{x})$*.* $\square$

As compared to the case of smooth convex minimization, the task of determining whether or not a certain $\mathbf{x} \in X$ solves the program [P] is, in general, computationally more expensive; recall in particular that even if an optimal solution $\mathbf{x} \in X^*$ is at hand, it is, in general, nontrivial to find a $\boldsymbol{\gamma} \in \partial f(\mathbf{x})$ that verifies optimality (e.g., satisfies the conditions of Proposition 1.1). Both these tasks may be performed by, for example, solving the optimization problem

$$v(\mathbf{x}) = \min_{\boldsymbol{\gamma} \in \partial f(\mathbf{x})} \left\| P_{-T_X(\mathbf{x})}(\boldsymbol{\gamma}) \right\|^2.$$

The value $v(\mathbf{x})$ is zero if and only if $\mathbf{x} \in X^*$.[1] If, for example, the subdifferential set and the tangent cone are polyhedral with explicit representations, then the function v is evaluated by solving a quadratic program. If the function f is implicitly defined, for example when the program [P] is a Lagrangean dual problem, one may have to invoke a column (i.e., a subgradient) generation scheme to enable the evaluation of v.

We consider *subgradient optimization methods* for solving the program [P]. These methods originate in a work by N.Z. Shor from 1962; see Shor (1991) for a review of the early history of nonsmooth optimization. These methods are conceptually simple, in general easy to implement, and have shown to be practically useful. They have therefore been frequently applied, particularly in connection with Lagrangean duality (e.g., Fisher, 1981 and 1985, and Beasley, 1993).

Given an initial iterate $\mathbf{x}^0 \in X$, the traditional subgradient optimization method for solving the program [P] (e.g., Polyak, 1969, or Shor, 1985) is given by the iteration formula

$$\mathbf{x}^{t+\frac{1}{2}} = \mathbf{x}^t - \alpha_t \boldsymbol{\gamma}(\mathbf{x}^t), \quad \mathbf{x}^{t+1} = P_X\left(\mathbf{x}^{t+\frac{1}{2}}\right), \quad t = 0, 1, \ldots, \tag{1.4}$$

where $\boldsymbol{\gamma}(\mathbf{x}^t) \in \partial f(\mathbf{x}^t)$ and $\alpha_t > 0$ is the step length.

For the case of unconstrained optimization, Ermol'ev (1966) proved convergence of the method (1.4) using step lengths according to a divergent series, and Polyak (1967 and 1969) extended it to the case of convex constraints and gave additional convergence results.

Even though subgradient optimization methods are proven convergent, the objective value can not be guaranteed to descend in each iteration.[2] Further, in general $\{P_{-T_X(\mathbf{x}^t)}(\boldsymbol{\gamma}(\mathbf{x}^t))\} \not\to \mathbf{0}$, and even if $\mathbf{x}^t \in X^*$ it is quite likely that $P_{-T_X(\mathbf{x}^t)}(\boldsymbol{\gamma}(\mathbf{x}^t)) \neq \mathbf{0}$. Hence, the progress of subgradient optimization methods can *not* be monitored in terms of the approximate fulfilment of optimality conditions (e.g., those of Proposition 1.1).

[1] The function $v : X \mapsto \Re_+$ is an example of a *merit* function for the problem [P].

[2] However, their computational burden in each iteration is low in comparison with that of descent methods for nonsmooth minimization, and a larger number of iterations can therefore be allowed.

In contrast, termination criteria in descent methods (for smooth or nonsmooth minimization) are often based on the approximate fulfilment of optimality conditions. Further, such methods enable the calculation of supplementary and useful information. An example of this is the convergent lower bound on the optimal value in the method of Frank and Wolfe (1956) for constrained, smooth, convex minimization. Another example is the availability of convergent multiplier estimates from the solutions to the direction finding subproblems in bundle methods; their values are of particular interest in applications where the multipliers have physical interpretations (e.g., Ben-Tal and Bendsøe, 1993).

A frequently used technique for inducing properties that an original sequence is lacking is to construct an auxiliary, *ergodic*, sequence from the original one, that is, a sequence with elements that are weighted averages of those of the original sequence. Typical applications of such averaging processes are the smoothing of oscillations in a nonconvergent sequence and the filtering out of stochastic noise (see, e.g., Kornai and Lipták, 1965; Dunn, 1973; and Powell and Sheffi, 1982).

We will show that the elements of an ergodic sequence of subgradients will, in the limit, fulfil the optimality conditions of Proposition 1.1; hence, the construction of such a sequence introduces the possibility to employ a termination criterion which is related to the approximate fulfilment of these conditions and, further, it enables the calculation of convergent Lagrange multiplier estimates. Subgradient optimization methods thus have an inherent capability of utilizing information assembled during the iteration process. This observation reveals a new characteristic of subgradient optimization methods and deepens the understanding of their behaviour. The practical usefulness of the ergodic results in this work will be demonstrated in a subsequent report.

1.2 Basic concepts

The ergodic results to be presented are derived for the conditional subgradient optimization method (Larsson et al., 1992), which generalizes the method (1.4) in the sense that the feasible set X may be taken into account when determining the step direction. It was shown (ibid.) that the practical performance of the resulting method when applied to three types of structured optimization models is better than that of the method (1.4).

The basis for the development of the generalized method is a relaxation of the subdifferential definition (1.1). The *conditional subdifferential* of f with respect to the feasible set X at $\mathbf{x} \in X$ (see Dem'janov and Šomesova, 1978 and 1980) is the nonempty, closed and convex set

$$\partial^X f(\mathbf{x}) = \left\{ \boldsymbol{\gamma} \in \Re^n \;\middle|\; f(\mathbf{y}) \geq f(\mathbf{x}) + \boldsymbol{\gamma}^T(\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{y} \in X \right\},$$

the elements of which will be referred to as *conditional subgradients*. Clearly, for all $\mathbf{x} \in X$, $\partial^X f(\mathbf{x}) \supseteq \partial f(\mathbf{x})$. Dem'janov and Šomesova (1978 and 1980) show the following result.

Proposition 1.4 (Conditional subdifferential). $\partial^X f(\mathbf{x}) = \partial f(\mathbf{x}) + N_X(\mathbf{x})$, $\mathbf{x} \in X$. □

The optimality conditions in Proposition 1.1 can be stated in terms of the conditional subdifferential.

Proposition 1.5 (Optimality condition). $\mathbf{x} \in X^*$ *if and only if* $\mathbf{0} \in \partial^X f(\mathbf{x})$. □

Given an initial iterate $\mathbf{x}^0 \in X$, the *conditional subgradient optimization* method (Larsson et al., 1992) is defined by the iteration formula

$$\mathbf{x}^{t+\frac{1}{2}} = \mathbf{x}^t - \alpha_t \boldsymbol{\gamma}^X(\mathbf{x}^t), \quad \mathbf{x}^{t+1} = P_X\left(\mathbf{x}^{t+\frac{1}{2}}\right), \quad t = 0, 1, \ldots, \tag{1.5}$$

where $\alpha_t > 0$ and $\boldsymbol{\gamma}^X(\mathbf{x}^t)$ is a conditional subgradient of f at $\mathbf{x}^t$, that is,

$$\boldsymbol{\gamma}^X(\mathbf{x}^t) = \gamma(\mathbf{x}^t) + \boldsymbol{\nu}(\mathbf{x}^t), \quad \text{with} \quad \gamma(\mathbf{x}^t) \in \partial f(\mathbf{x}^t) \quad \text{and} \quad \boldsymbol{\nu}(\mathbf{x}^t) \in N_X(\mathbf{x}^t).$$

Note that the vector $\mathbf{x}^{t+\frac{1}{2}} - \mathbf{x}^{t+1} \in N_X(\mathbf{x}^{t+1})$. Clearly, if $\{\boldsymbol{\nu}(\mathbf{x}^t)\} = \{\mathbf{0}\}$, then the method (1.5) reduces to the method (1.4). In the case where the set X is stated in terms of explicit constraints (and fulfils some constraint qualification) the normal vector $\boldsymbol{\nu}(\mathbf{x}^t)$ is a nonnegative linear combination of the outward normals of the constraints that are active at $\mathbf{x}^t$.

Choosing $\boldsymbol{\nu}(\mathbf{x}^t) = P_{N_X(\mathbf{x}^t)}(-\gamma(\mathbf{x}^t))$ results in $\boldsymbol{\gamma}^X(\mathbf{x}^t) = P_{-T_X(\mathbf{x}^t)}(\gamma(\mathbf{x}^t))$ (see Larsson et al., 1992, Lemma 3.2), defining the special case of the method (1.5) that is called the *subgradient projection* method; for the case of a polyhedral feasible set, the direction $-\boldsymbol{\gamma}^X(\mathbf{x}^t)$ from $\mathbf{x}^t$ is then feasible.

We consider step lengths satisfying the divergent series conditions

$$\alpha_t > 0, \ \forall t, \quad \{\alpha_t\} \to 0, \quad \text{and} \quad \sum_{t=0}^{\infty} \alpha_t = \infty, \tag{1.6}$$

and the additional condition

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty. \tag{1.7}$$

The following convergence result, where the conditions (1.6) yield $\{f(\mathbf{x}^t)\} \to f^*$ and the addition of the condition (1.7) leads to the convergence also of the sequence $\{\mathbf{x}^t\}$, is given by Larsson et al. (1992, Theorem 2.7).

Proposition 1.6 (Convergence of conditional subgradient optimization). *Let the method (1.5)–(1.7) be applied to the program* [P]. *If the sequences $\{\gamma(\mathbf{x}^t)\}$ and $\{\boldsymbol{\nu}(\mathbf{x}^t)\}$ are bounded, then $\{\mathbf{x}^t\} \to \mathbf{x}^\infty \in X^*$.* □

The boundedness of the sequence $\{\gamma(\mathbf{x}^t)\}$ is ensured if the program [P] has some suitable additional property, for example that X is bounded or that f is piecewise linear with a finite number of pieces. In contrast, boundedness of the sequence $\{\boldsymbol{\nu}(\mathbf{x}^t)\}$ may always be imposed.[3] If subgradient projection is used, then $\|\boldsymbol{\nu}(\mathbf{x}^t)\| \leq \|\gamma(\mathbf{x}^t)\|$ so that the sequence $\{\boldsymbol{\nu}(\mathbf{x}^t)\}$ is bounded whenever this is the case for the sequence $\{\gamma(\mathbf{x}^t)\}$.

To establish convergence of the method (1.5) for *adaptive* step length selection rules, based on line searches or formulas involving estimates of the optimal value, the *almost complete relaxation* strategy of Dem'yanov and Vasil'ev (1985, Section 3.4) can be used (Corollary 2.8 in Larsson et al., 1992). It works as follows. Define two sequences, $\{\underline{\alpha}_t\}$ and $\{\overline{\alpha}_t\}$, with $\underline{\alpha}_t \leq \overline{\alpha}_t$ for all t, and both satisfying the conditions (1.6) and (1.7).[4] Suppose that, at some iteration t, an adaptive rule has generated a (tentative) step length β_t. The step length α_t used is defined as the projection of β_t onto the interval

[3] For example, by choosing the elements to satisfy $\|\boldsymbol{\nu}(\mathbf{x}^t)\| \leq c$ for some $c \geq 0$.

[4] It may, for example, be appropriate to let the sequences be given by $\underline{\alpha}_t = \underline{c}/(t+1)$ and $\overline{\alpha}_t = \overline{c}/(t+1)$, $t = 0, 1, \ldots$, where $\underline{c} > 0$ and $\overline{c} > 0$ are very small and very large constants, respectively.

$[\underline{\alpha}_t, \overline{\alpha}_t]$. The sequence $\{\alpha_t\}$, resulting from this safeguarding strategy, satisfies the conditions (1.6) and (1.7).[5]

Now, consider the application of the method (1.5)–(1.7) to the program [P] and suppose that during the first t iterations the subgradients $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}_{s=0}^{t-1}$ have been used to define the step directions. We then define the sequence $\{\mathbf{g}^t\}$ of weighted averages of subgradients used up to iteration t as

$$\mathbf{g}^t = \sum_{s=0}^{t-1} \frac{\alpha_s}{\sum_{r=0}^{t-1} \alpha_r} \boldsymbol{\gamma}(\mathbf{x}^s), \quad t = 1, 2, \dots.$$

It will be shown (in Theorem 3.4) that, provided that the sequences $\{\boldsymbol{\gamma}(\mathbf{x}^t)\}$ and $\{\boldsymbol{\nu}(\mathbf{x}^t)\}$ are bounded, the elements of the ergodic sequence $\{\mathbf{g}^t\}$ in the limit fulfil the optimality conditions given in Proposition 1.1.

Note that the sequence $\{\mathbf{g}^t\}$ is calculated beside the sequence $\{\mathbf{x}^t\}$ and that it therefore does *not* affect the course of the method (1.5)–(1.7). The use of the sequence $\{\mathbf{g}^t\}$ for accelerating subgradient optimization algorithms is a subject for further research.

1.3 Outline

In Section 2 we define ergodic sequences of subgradients, normal cone elements, and objective values evaluated in the method (1.5), and of certain auxiliary functions. We also state convergence properties of the sequences of auxiliary functions. Then, in Section 3, we derive the main result that the elements of the ergodic sequence of subgradients tend toward the set $\partial f(\mathbf{x}) \cap -N_X(\mathbf{x})$, for some optimal solution $\mathbf{x}$. We also give ergodic results for sequences of conditional subgradients. Section 4 contains convergence results for an ergodic sequence of Lagrange multipliers. It is shown that, under a strong Slater constraint qualification, the elements of this ergodic sequence tend toward the set of Lagrange multipliers at an optimal solution; a corresponding result holds for multipliers associated with objective pieces. Finally, in Section 5, we discuss some opportunities for further research.

2 TECHNICAL PRELIMINARIES

We first define the sequences to be studied; these are ergodic sequences of the subgradients and the normal cone elements, respectively, generated in the method (1.5). The weights employed in the construction of the ergodic sequences are based on the step lengths used in this method. Then, we establish properties of sequences of certain affine functions in order to derive convergence results for the ergodic sequence of subgradients.

We denote by A_t the cumulative step length of the method (1.5) up to iteration t, that is,

$$A_t = \sum_{s=0}^{t-1} \alpha_s, \quad t = 1, 2, \dots. \tag{2.1}$$

Definition 2.1 (Ergodic sequence of subgradients). *Let the method (1.5) be applied to the program* [P]. *The ergodic sequence* $\{\mathbf{g}^t\}$ *of subgradients is defined by*

$$\mathbf{g}^t = A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \boldsymbol{\gamma}(\mathbf{x}^s), \quad t = 1, 2, \dots. \qquad \square$$

[5]Traditionally, the divergent series conditions (1.6) and (1.7) are fulfilled through the choice of step lengths according to some appropriate and a priori determined formula, e.g., $\alpha_t = 1/(t+1)$, $t = 0, 1, \dots$.

This ergodic sequence of subgradients is of primary interest in this work. Clearly, each vector $\mathbf{g}^t$ is a convex combination of the subgradients $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}_{s=0}^{t-1}$. A recursive computation of $\mathbf{g}^t$ can be implemented using the formula[6]

$$\mathbf{g}^1 = \boldsymbol{\gamma}(\mathbf{x}^0), \quad \mathbf{g}^t = \frac{A_{t-1}}{A_t}\mathbf{g}^{t-1} + \frac{\alpha_{t-1}}{A_t}\boldsymbol{\gamma}(\mathbf{x}^{t-1}), \quad t = 2, 3, \ldots.$$

Definition 2.2 (Ergodic sequence of normal vectors). *Let the method (1.5) be applied to the program* [P]. *The ergodic sequence of normal terms of the conditional subgradients is defined by*

$$\mathbf{n}_c^t = A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \boldsymbol{\nu}(\mathbf{x}^s), \quad t = 1, 2, \ldots,$$

the ergodic sequence of the projection steps is defined by

$$\mathbf{n}_p^t = A_t^{-1} \sum_{s=0}^{t-1} (\mathbf{x}^{s+\frac{1}{2}} - \mathbf{x}^{s+1}), \quad t = 1, 2, \ldots,$$

and the ergodic sequence of normal vectors is then defined by

$$\mathbf{n}^t = \mathbf{n}_c^t + \mathbf{n}_p^t, \quad t = 1, 2, \ldots. \qquad \square$$

Clearly, $\mathbf{n}_c^t$ is a convex combination of the normal terms of the conditional subgradients used up to iteration t of the method (1.5) while $\mathbf{n}_p^t$ is a positive linear combination of its projection steps. If the traditional subgradient optimization method (1.4) is employed then, for all t, $\mathbf{n}_c^t = \mathbf{0}$ and $\mathbf{n}^t = \mathbf{n}_p^t$.

The following result is the first step towards a convergence result for the ergodic sequence $\{\mathbf{g}^t\}$.

Lemma 2.3 (Boundedness of $\{\mathbf{g}^t\}$ and $\{\mathbf{n}^t\}$). *Let the sequences $\{\mathbf{g}^t\}$ and $\{\mathbf{n}^t\}$ be given by Definitions 2.1 and 2.2, respectively. If the sequences $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}$ and $\{\boldsymbol{\nu}(\mathbf{x}^s)\}$ are bounded, then this is also true for the sequences $\{\mathbf{g}^t\}$ and $\{\mathbf{n}^t\}$.*

Proof. Since the elements $\mathbf{g}^t$ are convex combinations of $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}_{s=0}^{t-1}$, the sequence $\{\mathbf{g}^t\}$ is also bounded. Analogously, since each $\mathbf{n}_c^t$ is a convex combination of $\{\boldsymbol{\nu}(\mathbf{x}^s)\}_{s=0}^{t-1}$, the sequence $\{\mathbf{n}_c^t\}$ is also bounded. By the iteration formula (1.5), $\|\mathbf{x}^{s+\frac{1}{2}} - \mathbf{x}^{s+1}\| \leq \alpha_s \|\boldsymbol{\gamma}^X(\mathbf{x}^s)\| \leq \alpha_s(\|\boldsymbol{\gamma}(\mathbf{x}^s)\| + \|\boldsymbol{\nu}(\mathbf{x}^s)\|)$ for all s. Using the same arguments as above, it follows that the sequence $\{\mathbf{n}_p^t\}$ is bounded. Hence, the sequence $\{\mathbf{n}^t\}$ is bounded. $\square$

The ergodic sequence of objective values obtained in the method (1.5) is defined as

$$\varphi_t = A_t^{-1} \sum_{s=0}^{t-1} \alpha_s f(\mathbf{x}^s), \quad t = 1, 2, \ldots. \tag{2.2}$$

It follows that $\varphi_t \geq f^*$ for all t, since, for all s, $\mathbf{x}^s \in X$.

Utilizing that, for all s, $\boldsymbol{\gamma}(\mathbf{x}^s) \in \partial f(\mathbf{x}^s)$, and the definition (1.1), we may derive the *surrogate inequalities*

$$\sum_{s=0}^{t-1} \alpha_s f(\mathbf{x}) \geq \sum_{s=0}^{t-1} \alpha_s \left(f(\mathbf{x}^s) + \boldsymbol{\gamma}(\mathbf{x}^s)^T (\mathbf{x} - \mathbf{x}^s) \right), \quad \mathbf{x} \in \Re^n, \quad t = 1, 2, \ldots,$$

[6]This remark applies to all ergodic sequences to be defined in this report.

and define the *affine* functions $f_t : \Re^n \mapsto \Re$, $t = 1, 2, \ldots$, as

$$f_t(\mathbf{x}) = A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \left(f(\mathbf{x}^s) + \boldsymbol{\gamma}(\mathbf{x}^s)^T(\mathbf{x} - \mathbf{x}^s) \right), \quad \mathbf{x} \in \Re^n. \tag{2.3}$$

Then, $f_t(\mathbf{x}) \leq f(\mathbf{x})$ holds for all $\mathbf{x} \in \Re^n$ and all t, that is, each function f_t is an affine *underestimate* of the objective f. Clearly, the hyperplane defined by the function f_t is a weighted average of all supporting hyperplanes generated in the previous iterations of the conditional subgradient optimization method (1.5). From Definition 2.1 follows that $\nabla f_t(\mathbf{x}) = \mathbf{g}^t$ and $f_t(\mathbf{x}) = f_t(\mathbf{y}) + (\mathbf{g}^t)^T(\mathbf{x} - \mathbf{y})$ for any $\mathbf{x}, \mathbf{y} \in \Re^n$ and all t.

We let the convex functions $\delta_t : \Re^n \mapsto \Re$, $t = 1, 2, \ldots$, describe the pointwise vertical distance between the functions f and f_t, that is,

$$\delta_t(\mathbf{x}) = f(\mathbf{x}) - f_t(\mathbf{x}), \quad \mathbf{x} \in \Re^n. \tag{2.4}$$

It follows that $\delta_t(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \Re^n$ and all t.

Further, we define the affine functions $\varrho_t : \Re^n \mapsto \Re$, $t = 1, 2, \ldots$, as

$$\varrho_t(\mathbf{x}) = A_t^{-1} \sum_{s=0}^{t-1} \left(\alpha_s \boldsymbol{\nu}(\mathbf{x}^s)^T (\mathbf{x}^s - \mathbf{x}) + \left(\mathbf{x}^{s+\frac{1}{2}} - \mathbf{x}^{s+1}\right)^T \left(\mathbf{x}^{s+1} - \mathbf{x}\right) \right), \quad \mathbf{x} \in \Re^n. \tag{2.5}$$

Using that, for all s, $\boldsymbol{\nu}(\mathbf{x}^s) \in N_X(\mathbf{x}^s)$ and $\mathbf{x}^{s+\frac{1}{2}} - \mathbf{x}^{s+1} \in N_X(\mathbf{x}^{s+1})$, and the definition (1.2), it follows that $\varrho_t(\mathbf{x}) \geq 0$ for $\mathbf{x} \in X$ and all t, that is, each of these inequalities is a *valid inequality* for the set X. From Definition 2.2 follows that $\nabla \varrho_t(\mathbf{x}) = -\mathbf{n}^t$ and $\varrho_t(\mathbf{x}) = \varrho_t(\mathbf{y}) - (\mathbf{n}^t)^T(\mathbf{x} - \mathbf{y})$ for any $\mathbf{x}, \mathbf{y} \in \Re^n$ and all t.

The following lemma is crucial for our continued analysis.

Lemma 2.4 (Convergence of $\{\delta_t(\mathbf{x}^\infty)\}$ and $\{\varrho_t(\mathbf{x}^\infty)\}$ to 0 and of $\{\varphi_t\}$ to f^*). *Let the method (1.5)–(1.7) be applied to the program* [P] *and let the sequences* $\{\delta_t(\mathbf{x})\}$, $\{\varrho_t(\mathbf{x})\}$, *and* $\{\varphi_t\}$ *be given by the definitions (2.4), (2.5), and (2.2), respectively. Suppose that the sequences* $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}$ *and* $\{\boldsymbol{\nu}(\mathbf{x}^s)\}$ *are bounded. Then,* $\{\delta_t(\mathbf{x}^\infty)\} \to 0$, $\{\varrho_t(\mathbf{x}^\infty)\} \to 0$, *and* $\{\varphi_t\} \to f^*$.

Proof. By Proposition 1.6, $\{\mathbf{x}^t\} \to \mathbf{x}^\infty$. From the iteration formula (1.5),

$$\begin{aligned}
0 &\leq \left\|\mathbf{x}^{t+1} - \mathbf{x}^\infty\right\|^2 \leq \left\|\mathbf{x}^{t+1} - \mathbf{x}^\infty\right\|^2 + \left\|\mathbf{x}^{t+\frac{1}{2}} - \mathbf{x}^{t+1}\right\|^2 \\
&= \left\|\mathbf{x}^{t+\frac{1}{2}} - \mathbf{x}^\infty\right\|^2 - 2\left(\mathbf{x}^{t+\frac{1}{2}} - \mathbf{x}^{t+1}\right)^T\left(\mathbf{x}^{t+1} - \mathbf{x}^\infty\right) \\
&= \left\|\mathbf{x}^t - \mathbf{x}^\infty\right\|^2 + \alpha_t^2\left\|\boldsymbol{\gamma}^X(\mathbf{x}^t)\right\|^2 - 2\left(\alpha_t \boldsymbol{\gamma}^X(\mathbf{x}^t)^T\left(\mathbf{x}^t - \mathbf{x}^\infty\right) + \left(\mathbf{x}^{t+\frac{1}{2}} - \mathbf{x}^{t+1}\right)^T\left(\mathbf{x}^{t+1} - \mathbf{x}^\infty\right)\right).
\end{aligned}$$

Repeated application of the second inequality and utilization of the definitions (2.4), (2.5), and (2.2), result in the inequality

$$0 \leq \left\|\mathbf{x}^0 - \mathbf{x}^\infty\right\|^2 + \sum_{s=0}^{t-1} \alpha_s^2 \left\|\boldsymbol{\gamma}^X(\mathbf{x}^s)\right\|^2 - 2A_t \left(\delta_t(\mathbf{x}^\infty) + \varrho_t(\mathbf{x}^\infty) + \varphi_t - f^*\right).$$

Since $\delta_t(\mathbf{x}^\infty) \geq 0$, $\varrho_t(\mathbf{x}^\infty) \geq 0$, $\varphi_t \geq f^*$, and $A_t > 0$, it is immediate that

$$0 \leq \delta_t(\mathbf{x}^\infty) + \varrho_t(\mathbf{x}^\infty) + \varphi_t - f^* \leq \frac{1}{2A_t} \left(\left\|\mathbf{x}^0 - \mathbf{x}^\infty\right\|^2 + \sum_{s=0}^{t-1} \alpha_s^2 \left\|\boldsymbol{\gamma}(\mathbf{x}^s) + \boldsymbol{\nu}(\mathbf{x}^s)\right\|^2 \right).$$

Let $t \to \infty$ and invoke the conditions (1.6) and (1.7). $\square$

Corollary 2.5 *Let the assumptions of Lemma 2.4 hold. Then,* $\{\delta_t(\mathbf{x}^t)\} \to 0$ *and* $\{\varrho_t(\mathbf{x}^t)\} \to 0$.

Proof. Definition 2.1, and the definitions (2.3) and (2.4), yield

$$0 \leq \delta_t(\mathbf{x}^t) = \delta_t(\mathbf{x}^\infty) + f(\mathbf{x}^t) - f(\mathbf{x}^\infty) - \left(\mathbf{g}^t\right)^T \left(\mathbf{x}^t - \mathbf{x}^\infty\right), \quad t = 1, 2, \ldots.$$

From Definition 2.2, and the definition (2.5), follow that

$$0 \leq \varrho_t(\mathbf{x}^t) = \varrho_t(\mathbf{x}^\infty) - \left(\mathbf{n}^t\right)^T \left(\mathbf{x}^t - \mathbf{x}^\infty\right), \quad t = 1, 2, \ldots.$$

Proposition 1.6, Lemma 2.3, the continuity of f, and Lemma 2.4 yield that the right hand sides of both the above equations tend to zero as t approaches infinity. The result follows. □

A small value of $\delta_t(\mathbf{x}^t) \geq 0$ implies that the affine underestimate f_t of the function f defines a hyperplane that is close to a support to $\texttt{epi}\, f$ at $(\mathbf{x}^t, f(\mathbf{x}^t))$, and a small value of $\varrho_t(\mathbf{x}^t)$ yields that the hyperplane $\{\mathbf{x} \in \Re^n \mid \varrho_t(\mathbf{x}) = 0\}$ is close to a support to X at $\mathbf{x}^t$ (while $X \subseteq \{\mathbf{x} \in \Re^n \mid \varrho_t(\mathbf{x}) \geq 0\}$ always holds).

Lemma 2.6 ($\{\mathbf{g}^t + \mathbf{n}^t\}$ tends to $\mathbf{0}$). *Let the method (1.5)–(1.7) be applied to the program* [P] *and let the sequences* $\{\mathbf{g}^t\}$ *and* $\{\mathbf{n}^t\}$ *be given by Definitions 2.1 and 2.2, respectively. Suppose that the sequences* $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}$ *and* $\{\boldsymbol{\nu}(\mathbf{x}^s)\}$ *are bounded. Then,* $\{\mathbf{g}^t + \mathbf{n}^t\} \to \mathbf{0}$.

Proof. By Definitions 2.1 and 2.2, and the iteration formula (1.5),

$$\begin{aligned}
\mathbf{g}^t + \mathbf{n}^t &= A_t^{-1} \sum_{s=0}^{t-1} \left(\alpha_s \boldsymbol{\gamma}^X(\mathbf{x}^s) + \mathbf{x}^{s+\frac{1}{2}} - \mathbf{x}^{s+1}\right) \\
&= A_t^{-1} \sum_{s=0}^{t-1} \left(\mathbf{x}^s - \mathbf{x}^{s+\frac{1}{2}} + \mathbf{x}^{s+\frac{1}{2}} - \mathbf{x}^{s+1}\right) \;=\; A_t^{-1}\left(\mathbf{x}^0 - \mathbf{x}^t\right).
\end{aligned}$$

Proposition 1.6 yields that $\{\mathbf{x}^t\} \to \mathbf{x}^\infty$. The result then follows from the definition (2.1) and the condition (1.6). □

In our continued analysis we will by $d_Y(\mathbf{x})$ denote the Euclidean distance between a point $\mathbf{x} \in \Re^n$ and a convex set $Y \subseteq \Re^n$, that is,

$$d_Y(\mathbf{x}) = \min_{\mathbf{y} \in Y} \|\mathbf{y} - \mathbf{x}\|.$$

The function $d_Y : \Re^n \mapsto \Re_+$ is convex and continuous.

3 ERGODIC RESULTS FOR SUBGRADIENTS

We first show that the distance between $\mathbf{g}^t$ and the set $\partial f(\mathbf{x}^\infty)$ tends to zero. Then, we establish that the projection of $\mathbf{g}^t$ onto the set $-T_X(\mathbf{x}^\infty)$ tends to zero. The combination of these results yields that the elements of the sequence $\{\mathbf{g}^t\}$ in the limit provide fulfilment of the optimality conditions given in Proposition 1.1. Finally, we give corresponding results for an ergodic sequence of *conditional* subgradients.

As concluded above, the affine functions f_t, $t = 1, 2, \ldots$, underestimate the convex function f everywhere, and from Lemma 2.4 follows that $\{f_t(\mathbf{x}^\infty)\} \to f(\mathbf{x}^\infty)$. Hence, in the limit, the sequence $\{f_t\}$ of affine functions provides supporting hyperplanes to $\texttt{epi}\, f$ at $(\mathbf{x}^\infty, f(\mathbf{x}^\infty))$; this observation is further developed in the following theorem.

Theorem 3.1 (Convergence of $\{\mathbf{g}^t\}$ to $\partial f(\mathbf{x}^\infty)$). *Let the method (1.5)–(1.7) be applied to the program* [P] *and let the sequence* $\{\mathbf{g}^t\}$ *be given by Definition 2.1. Suppose that the sequences* $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}$ *and* $\{\boldsymbol{\nu}(\mathbf{x}^s)\}$ *are bounded. Then,*

$$\left\{d_{\partial f(\mathbf{x}^\infty)}(\mathbf{g}^t)\right\} \to 0.$$

Proof. By the properties of $\partial f(\mathbf{x}^\infty)$, the sequence $\{d_{\partial f(\mathbf{x}^\infty)}(\mathbf{g}^t)\}$ is well defined. Let $\mathcal{T}$ be a subsequence such that

$$\lim_{\substack{t\in\mathcal{T}\\ t\to\infty}} d_{\partial f(\mathbf{x}^\infty)}\left(\mathbf{g}^t\right) = \limsup_{t\to\infty} d_{\partial f(\mathbf{x}^\infty)}\left(\mathbf{g}^t\right).$$

By Lemma 2.3, the sequence $\{\mathbf{g}^t\}_{t\in\mathcal{T}}$ is bounded and thus has an accumulation point, say $\tilde{\mathbf{g}}$. From the definitions (2.3) and (2.4) follow that, for all t and any $\mathbf{y} \in \Re^n$,

$$f(\mathbf{y}) \geq f_t(\mathbf{y}) = f_t(\mathbf{x}^\infty) + \left(\mathbf{g}^t\right)^T(\mathbf{y}-\mathbf{x}^\infty) = f(\mathbf{x}^\infty) + \left(\mathbf{g}^t\right)^T(\mathbf{y}-\mathbf{x}^\infty) - \delta_t(\mathbf{x}^\infty).$$

From Lemma 2.4 then follows that, for any $\mathbf{y} \in \Re^n$,

$$\tilde{\mathbf{g}}^T(\mathbf{y}-\mathbf{x}^\infty) \leq \limsup_{\substack{t\in\mathcal{T}\\ t\to\infty}} \left(\mathbf{g}^t\right)^T(\mathbf{y}-\mathbf{x}^\infty) \leq f(\mathbf{y}) - f(\mathbf{x}^\infty) + \lim_{\substack{t\in\mathcal{T}\\ t\to\infty}} \delta_t(\mathbf{x}^\infty) = f(\mathbf{y}) - f(\mathbf{x}^\infty).$$

Hence, $\tilde{\mathbf{g}} \in \partial f(\mathbf{x}^\infty)$, implying that $d_{\partial f(\mathbf{x}^\infty)}(\tilde{\mathbf{g}}) = 0$. By the continuity of the function $d_{\partial f(\mathbf{x}^\infty)}$, $d_{\partial f(\mathbf{x}^\infty)}(\tilde{\mathbf{g}}) = \lim_{t\in\mathcal{T},\, t\to\infty} d_{\partial f(\mathbf{x}^\infty)}(\mathbf{g}^t)$, and the result follows. □

For any $\varepsilon \geq 0$, the *ε-subdifferential* of f at $\mathbf{x} \in \Re^n$ is defined as (e.g., Hiriart-Urruty and Lemaréchal, 1993b, Definition XI:1.1.1)

$$\partial_\varepsilon f(\mathbf{x}) = \left\{ \boldsymbol{\gamma} \in \Re^n \;\middle|\; f(\mathbf{y}) \geq f(\mathbf{x}) + \boldsymbol{\gamma}^T(\mathbf{y}-\mathbf{x}) - \varepsilon, \quad \forall \mathbf{y} \in \Re^n \right\}.$$

For any $\varepsilon_2 > \varepsilon_1 \geq 0$ and any $\mathbf{x} \in \Re^n$, $\partial_{\varepsilon_1} f(\mathbf{x}) \subseteq \partial_{\varepsilon_2} f(\mathbf{x})$, and $\partial_0 f(\mathbf{x}) = \partial f(\mathbf{x})$. Using Definition 2.1 and the definitions (2.3) and (2.4) we derive that $\mathbf{g}^t$ is a $\delta_t(\mathbf{x})$-subgradient of f at $\mathbf{x}$ for any $\mathbf{x} \in \Re^n$; note in particular that $\mathbf{g}^t$ is a $\delta_t(\mathbf{x}^t)$-subgradient of f at $\mathbf{x}^t$ and that $\{\delta_t(\mathbf{x}^t)\} \to 0$ by Corollary 2.5.

From the definition (2.5) follows that, for every t, $\varrho_t(\mathbf{x}) \geq 0$ is a valid inequality for the convex set X, and from Lemma 2.4 that $\{\varrho_t(\mathbf{x}^\infty)\} \to 0$. Hence, in the limit, the sequence $\{\varrho_t\}$ of affine functions provides supporting hyperplanes to the set X at $\mathbf{x}^\infty$; the next theorem further develops this remark.

Theorem 3.2 (Convergence of $\{\mathbf{g}^t\}$ to $-N_X(\mathbf{x}^\infty)$). *Let the method (1.5)–(1.7) be applied to the program* [P] *and let the sequence* $\{\mathbf{g}^t\}$ *be given by Definition 2.1. Suppose that the sequences* $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}$ *and* $\{\boldsymbol{\nu}(\mathbf{x}^s)\}$ *are bounded. Then,*

$$\left\{d_{-N_X(\mathbf{x}^\infty)}(\mathbf{g}^t)\right\} \to 0.$$

Proof. By the properties of $N_X(\mathbf{x}^\infty)$, the sequences $\{d_{-N_X(\mathbf{x}^\infty)}(\mathbf{g}^t)\}$ and $\{d_{N_X(\mathbf{x}^\infty)}(\mathbf{n}^t)\}$, where $\mathbf{n}^t$ is given by Definition 2.2, are well defined. Let $\mathcal{T}$ be a subsequence such that

$$\lim_{\substack{t\in\mathcal{T}\\ t\to\infty}} d_{N_X(\mathbf{x}^\infty)}\left(\mathbf{n}^t\right) = \limsup_{t\to\infty} d_{N_X(\mathbf{x}^\infty)}\left(\mathbf{n}^t\right).$$

By Lemma 2.3, the sequence $\{\mathbf{n}^t\}_{t\in\mathcal{T}}$ is bounded and thus has an accumulation point, say $\tilde{\mathbf{n}}$. From the definition (2.5) follows that, for all t and any $\mathbf{y} \in X$,

$$\left(\mathbf{n}^t\right)^T (\mathbf{y} - \mathbf{x}^\infty) = -\varrho_t(\mathbf{y}) + \varrho_t(\mathbf{x}^\infty) \leq \varrho_t(\mathbf{x}^\infty).$$

From Lemma 2.4 follow that, for any $\mathbf{y} \in X$,

$$\tilde{\mathbf{n}}^T (\mathbf{y} - \mathbf{x}^\infty) \leq \limsup_{\substack{t\in\mathcal{T}\\ t\to\infty}} \left(\mathbf{n}^t\right)^T (\mathbf{y} - \mathbf{x}^\infty) \leq \lim_{\substack{t\in\mathcal{T}\\ t\to\infty}} \varrho_t(\mathbf{x}^\infty) = 0.$$

Hence, $\tilde{\mathbf{n}} \in N_X(\mathbf{x}^\infty)$, implying that $d_{N_X(\mathbf{x}^\infty)}(\tilde{\mathbf{n}}) = 0$. By the continuity of the function $d_{N_X(\mathbf{x}^\infty)}$, $d_{N_X(\mathbf{x}^\infty)}(\tilde{\mathbf{n}}) = \lim_{t\in\mathcal{T},\ t\to\infty} d_{N_X(\mathbf{x}^\infty)}(\mathbf{n}^t)$. Therefore, for any $\varepsilon > 0$ and all t that are sufficiently large, $d_{N_X(\mathbf{x}^\infty)}(\mathbf{n}^t) < \varepsilon/2$, and, by Lemma 2.6, $\|\mathbf{g}^t + \mathbf{n}^t\| < \varepsilon/2$, so that $d_{-N_X(\mathbf{x}^\infty)}(\mathbf{g}^t) \leq d_{-N_X(\mathbf{x}^\infty)}(-\mathbf{n}^t) + \|\mathbf{g}^t + \mathbf{n}^t\| < \varepsilon$. The theorem follows. $\square$

From Definition 2.2 and the definition (2.5) follow that, for any $\mathbf{x} \in X$ and all t,

$$\mathbf{n}^t \in \left\{ \boldsymbol{\nu} \in \Re^n \;\middle|\; \boldsymbol{\nu}^T(\mathbf{y} - \mathbf{x}) \leq \varrho_t(\mathbf{x}), \quad \forall \mathbf{y} \in X \right\}.$$

This set is nonempty, closed, and convex, and it clearly contains $N_X(\mathbf{x})$ (cf. the relationship between the ε-subdifferential and the subdifferential).

Theorem 3.3 (Projection of $\mathbf{g}^t$ onto $-T_X(\mathbf{x}^\infty)$ tends to $\mathbf{0}$). *Let the assumptions of Theorem 3.2 hold and let $T_X(\mathbf{x})$ be defined by (1.3). Then,*

$$\left\{P_{-T_X(\mathbf{x}^\infty)}(\mathbf{g}^t)\right\} \to \mathbf{0}.$$

Proof. The sequence $\{P_{-T_X(\mathbf{x}^\infty)}(\mathbf{g}^t)\}$ is well defined. Applying the Moreau decomposition (Zarantonello, 1971, Lemma 2.2) to $\mathbf{g}^t$, with respect to the polar cones $N_X(\mathbf{x}^\infty)$ and $T_X(\mathbf{x}^\infty)$, yields

$$\mathbf{g}^t = P_{-N_X(\mathbf{x}^\infty)}\left(\mathbf{g}^t\right) + P_{-T_X(\mathbf{x}^\infty)}\left(\mathbf{g}^t\right), \quad \forall t,$$

which implies that $\|P_{-T_X(\mathbf{x}^\infty)}(\mathbf{g}^t)\| = \|\mathbf{g}^t - P_{-N_X(\mathbf{x}^\infty)}(\mathbf{g}^t)\| = d_{-N_X(\mathbf{x}^\infty)}(\mathbf{g}^t)$. The result then follows from Theorem 3.2. $\square$

Theorem 3.4 (Optimality fulfilment). *Let the method (1.5)–(1.7) be applied to the program* [P] *and let the sequence $\{\mathbf{g}^t\}$ be given by Definition 2.1. Suppose that the sequences $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}$ and $\{\boldsymbol{\nu}(\mathbf{x}^s)\}$ are bounded. Then,*

$$\left\{d_{\partial f(\mathbf{x}^\infty)\cap -N_X(\mathbf{x}^\infty)}(\mathbf{g}^t)\right\} \to 0$$

or, equivalently,

$$\left\{d_{\partial f(\mathbf{x}^\infty)}(\mathbf{g}^t)\right\} \to 0 \quad \textit{and} \quad \left\{P_{-T_X(\mathbf{x}^\infty)}(\mathbf{g}^t)\right\} \to \mathbf{0}.$$

Proof. The first statement follows from Theorems 3.1 and 3.2, and the second statement is then implied by Theorem 3.3. $\square$

The corresponding ergodic result for the sequence $\{\mathbf{n}^t\}$ is $\{d_{N_X(\mathbf{x}^\infty)\cap -\partial f(\mathbf{x}^\infty)}(\mathbf{n}^t)\} \to 0$ or, equivalently, $\{d_{-\partial f(\mathbf{x}^\infty)}(\mathbf{n}^t)\} \to 0$ and $\{P_{T_X(\mathbf{x}^\infty)}(\mathbf{n}^t)\} \to \mathbf{0}$.

In the interesting special case of *unconstrained minimization* the optimality conditions given in Proposition 1.1 reduce to $\mathbf{0} \in \partial f(\mathbf{x})$ and Theorem 3.4 reduces to the following.

Corollary 3.5 (Convergence of $\{\mathbf{g}^t\}$ to $\mathbf{0}$). *Suppose that $X = \Re^n$ and that the assumptions of Theorem 3.4 hold. Then,*

$$\left\{\mathbf{g}^t\right\} \to \mathbf{0}. \qquad \square$$

Finally in this section, we define the ergodic sequence of *conditional* subgradients by

$$\mathbf{g}^t + \mathbf{n}_c^t = A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \gamma^X(\mathbf{x}^s), \quad t = 1, 2, \ldots, \tag{3.1}$$

where $\mathbf{g}^t$ and $\mathbf{n}_c^t$ are given by Definitions 2.1 and 2.2, respectively.

Theorem 3.6 (Optimality fulfilment). *Let the assumptions of Theorem 3.4 hold and let the sequence $\{\mathbf{g}^t + \mathbf{n}_c^t\}$ be defined by (3.1). Then,*

$$\left\{d_{\partial^X f(\mathbf{x}^\infty)\cap -N_X(\mathbf{x}^\infty)}(\mathbf{g}^t + \mathbf{n}_c^t)\right\} \to 0$$

or, equivalently,

$$\left\{d_{\partial^X f(\mathbf{x}^\infty)}(\mathbf{g}^t + \mathbf{n}_c^t)\right\} \to 0 \quad \textit{and} \quad \left\{P_{-T_X(\mathbf{x}^\infty)}(\mathbf{g}^t + \mathbf{n}_c^t)\right\} \to \mathbf{0}.$$

Proof. Analogously to the proof of Theorem 3.2 we can show that $\{d_{N_X(\mathbf{x}^\infty)}(\mathbf{n}_c^t)\} \to 0$ and $\{d_{N_X(\mathbf{x}^\infty)}(\mathbf{n}_p^t)\} \to 0$. From Lemma 2.6 and Definition 2.2, $\{\mathbf{g}^t + \mathbf{n}_c^t + \mathbf{n}_p^t\} \to \mathbf{0}$, which then implies that $\{d_{-N_X(\mathbf{x}^\infty)}(\mathbf{g}^t + \mathbf{n}_c^t)\} \to 0$. From Theorem 3.1 then follows that $\{d_{\partial f(\mathbf{x}^\infty)+N_X(\mathbf{x}^\infty)}(\mathbf{g}^t + \mathbf{n}_c^t)\} \to 0$. Utilizing the polarity of the normal and tangent cones we can, analogously to the proof of Theorem 3.3 show that the projection of $\mathbf{g}^t + \mathbf{n}_c^t$ onto $-T_X(\mathbf{x}^\infty)$ vanishes in the limit. $\square$

The corresponding ergodic result for the sequence $\{\mathbf{n}_p^t\}$ is $\{d_{N_X(\mathbf{x}^\infty)\cap -\partial^X f(\mathbf{x}^\infty)}(\mathbf{n}_p^t)\} \to 0$ or, equivalently, $\{d_{-\partial^X f(\mathbf{x}^\infty)}(\mathbf{n}_p^t)\} \to 0$ and $\{P_{T_X(\mathbf{x}^\infty)}(\mathbf{n}_p^t)\} \to \mathbf{0}$.

4 ERGODIC RESULTS FOR LAGRANGE MULTIPLIERS

Here, we consider the cases when the feasible set and/or the objective function of the program [P] have explicit representations, and utilize the previous ergodic results to construct ergodic sequences of the Lagrange multipliers associated with these representations of the program [P]. In Subsection 4.1 we assume that the feasible set is defined by a finite number of smooth, convex inequality constraints and/or linear equality constraints, and satisfies a strong Slater constraint qualification. The ergodic sequences of multipliers associated with these constraints are shown to converge to Lagrange multipliers. In Subsection 4.2 we assume that the objective function is represented as the pointwise maximum of a finite number of smooth, convex functions. We show that the ergodic sequences of multipliers associated with the pieces of the objective function converge to Lagrange multipliers.

4.1 Multipliers of constraints

Throughout this subsection we assume that the feasible set X of the program [P] is explicitly described as

$$X = \left\{ \mathbf{x} \in \Re^n \;\middle|\; a_i(\mathbf{x}) \leq 0,\ i \in \mathcal{M};\quad \mathbf{b}_j^T \mathbf{x} = e_j,\ j \in \mathcal{P} \right\}, \tag{4.1}$$

where the functions $a_i : \Re^n \mapsto \Re$, $i \in \mathcal{M} = \{1, \ldots, m\}$, are convex and differentiable, and $\mathbf{b}_j \in \Re^n$ and $e_j \in \Re$, $j \in \mathcal{P} = \{m+1, \ldots, m+p\}$. The convergence analysis of this subsection relies on the assumption that the set X fulfils the following constraint qualification (e.g., Hiriart-Urruty and Lemaréchal, 1993a, Definition VII:2.3.1).

Assumption 4.1 (Strong Slater assumption). *The vectors $\mathbf{b}_j$, $j \in \mathcal{P}$, are linearly independent, and the set X has a nonempty relative interior, that is,*

$$\left\{ \mathbf{x} \in \Re^n \;\middle|\; a_i(\mathbf{x}) < 0,\ i \in \mathcal{M};\quad \mathbf{b}_j^T \mathbf{x} = e_j,\ j \in \mathcal{P} \right\} \neq \emptyset. \qquad \square$$

We let, for $\mathbf{x} \in X$, $\mathcal{I}(\mathbf{x}) = \{ i \in \mathcal{M} \mid a_i(\mathbf{x}) = 0 \}$ denote the set of inequality constraints that are active at $\mathbf{x}$, and $\Pi_X(\mathbf{x})$ the set of multiplier values that satisfy the complementary slackness conditions for the program [P] at $\mathbf{x}$, that is,

$$\Pi_X(\mathbf{x}) = \left\{ \boldsymbol{\pi} \in \Re^m_+ \times \Re^p \;\middle|\; \pi_i = 0,\ i \in \mathcal{M} \setminus \mathcal{I}(\mathbf{x}) \right\} \ni \mathbf{0}.$$

Letting $\boldsymbol{\omega} \in \Re^{m+p}$, the normal cone of X at $\mathbf{x} \in X$ can then be expressed as

$$N_X(\mathbf{x}) = \left\{ \boldsymbol{\nu} \in \Re^n \;\middle|\; \boldsymbol{\nu} = \sum_{i \in \mathcal{M}} \omega_i \nabla a_i(\mathbf{x}) + \sum_{j \in \mathcal{P}} \omega_j \mathbf{b}_j;\quad \boldsymbol{\omega} \in \Pi_X(\mathbf{x}) \right\}.$$

For any $\mathbf{x} \in X$, let

$$M_X(\mathbf{x}) = \left\{ \boldsymbol{\mu} \in \Pi_X(\mathbf{x}) \;\middle|\; \mathbf{0} \in \partial f(\mathbf{x}) + \sum_{i \in \mathcal{M}} \mu_i \nabla a_i(\mathbf{x}) + \sum_{j \in \mathcal{P}} \mu_j \mathbf{b}_j \right\} \tag{4.2}$$

be the (possibly empty) set of Lagrange multipliers associated with the constraints defining the set X.

The optimality conditions for the program [P] can be stated in terms of the set $M_X(\mathbf{x})$ (e.g., Hiriart-Urruty and Lemaréchal, 1993a, Theorems VII:2.1.4 and VII:2.3.2).[7]

Proposition 4.2 (Optimality conditions). *Suppose that the set X, defined by (4.1), fulfils Assumption 4.1, and let $\mathbf{x} \in X$. Then, $\mathbf{x} \in X^*$ if and only if $M_X(\mathbf{x}) \neq \emptyset$.* $\square$

Proposition 4.3 (Properties of $M_X(\mathbf{x})$, $\mathbf{x} \in X^*$). *Suppose that the set X, defined by (4.1), fulfils Assumption 4.1. For any $\mathbf{x} \in X^*$ the nonempty set $M_X(\mathbf{x})$ is convex and compact, and $M_X(\mathbf{y}) = M_X(\mathbf{x})$ for all $\mathbf{y} \in X^*$.*

Proof. The closedness and convexity of $M_X(\mathbf{x})$ follow from its definition (4.2). The invariance over the set X^* and the boundedness follow from Proposition VII:3.1.1 and Theorem VII:2.3.2, respectively, in Hiriart-Urruty and Lemaréchal (1993a). $\square$

Our derivation of convergence results for the ergodic sequence of Lagrange multipliers starts with the following lemma.

[7] Parts of Propositions 4.2 and 4.3 are true also under weaker constraint qualifications.

Lemma 4.4 *Suppose that the set X, defined by (4.1), fulfils Assumption 4.1, and let $\mathbf{x} \in X$ and $\boldsymbol{\pi} \in \Re^{m+p}$. Then, the system*

$$\sum_{i\in\mathcal{I}(\mathbf{x})} \pi_i \nabla a_i(\mathbf{x}) + \sum_{j\in\mathcal{P}} \pi_j \mathbf{b}_j = \mathbf{0} \tag{4.3a}$$

$$\boldsymbol{\pi} \in \Pi_X(\mathbf{x}) \tag{4.3b}$$

is satisfied if and only if $\boldsymbol{\pi} = \mathbf{0}$.

Proof. Clearly, $\boldsymbol{\pi} = \mathbf{0}$ satisfies the system (4.3).

Conversely, suppose that $\boldsymbol{\pi}$ satisfies the system (4.3). From Assumption 4.1 follows that there is an $\widehat{\mathbf{x}} \neq \mathbf{x}$ such that $\widehat{\mathbf{x}} \in \texttt{rint}\, X$. The convexity and differentiability of the functions a_i, together with the fact that $a_i(\mathbf{x}) = 0$ for all $i \in \mathcal{I}(\mathbf{x})$, then yield that $\nabla a_i(\mathbf{x})^T(\widehat{\mathbf{x}} - \mathbf{x}) \leq a_i(\widehat{\mathbf{x}})$ for all $i \in \mathcal{I}(\mathbf{x})$, and, since $\pi_i \geq 0$ and $a_i(\widehat{\mathbf{x}}) < 0$ for all $i \in \mathcal{I}(\mathbf{x})$, that

$$\sum_{i\in\mathcal{I}(\mathbf{x})} \pi_i \nabla a_i(\mathbf{x})^T (\widehat{\mathbf{x}} - \mathbf{x}) \leq \sum_{i\in\mathcal{I}(\mathbf{x})} \pi_i a_i(\widehat{\mathbf{x}}) \leq 0. \tag{4.4}$$

However, since, for all $j \in \mathcal{P}$, $\mathbf{b}_j^T(\widehat{\mathbf{x}} - \mathbf{x}) = 0$, it follows from the equations (4.3a) that $\sum_{i\in\mathcal{I}(\mathbf{x})} \pi_i \nabla a_i(\mathbf{x})^T(\widehat{\mathbf{x}} - \mathbf{x}) = 0$, and from (4.4) we then obtain that $\sum_{i\in\mathcal{I}(\mathbf{x})} \pi_i a_i(\widehat{\mathbf{x}}) = 0$. From the relations $\pi_i \geq 0$ and $a_i(\widehat{\mathbf{x}}) < 0$ for all $i \in \mathcal{I}(\mathbf{x})$ it follows that $\pi_i = 0$ for all $i \in \mathcal{I}(\mathbf{x})$. The system (4.3) then reduces to the equations $\sum_{j\in\mathcal{P}} \pi_j \mathbf{b}_j = \mathbf{0}$ and $\pi_i = 0$, $i \in \mathcal{M}$. By Assumption 4.1, the vectors $\mathbf{b}_j$, $j \in \mathcal{P}$, are linearly independent, thus implying that $\pi_j = 0$ for all $j \in \mathcal{P}$. Hence, $\boldsymbol{\pi} = \mathbf{0}$. □

Definition 4.5 (Ergodic sequence of Lagrange multipliers). *Let the set X be defined by (4.1), and let the method (1.5) be applied to the program* [P]. *The sequences $\{\boldsymbol{\omega}(\mathbf{x}^s)\}$ and $\{\boldsymbol{\lambda}(\mathbf{x}^s)\}$ are defined by*

$$\boldsymbol{\omega}(\mathbf{x}^s) \in \left\{ \boldsymbol{\omega} \in \Pi_X(\mathbf{x}^s) \;\middle|\; \boldsymbol{\nu}(\mathbf{x}^s) = \sum_{i\in\mathcal{M}} \omega_i \nabla a_i(\mathbf{x}^s) + \sum_{j\in\mathcal{P}} \omega_j \mathbf{b}_j \right\}, \quad s = 0, 1, \ldots,$$

and

$$\boldsymbol{\lambda}(\mathbf{x}^s) \in \left\{ \boldsymbol{\lambda} \in \Pi_X(\mathbf{x}^{s+1}) \;\middle|\; \mathbf{x}^{s+\frac{1}{2}} - \mathbf{x}^{s+1} = \sum_{i\in\mathcal{M}} \lambda_i \nabla a_i(\mathbf{x}^{s+1}) + \sum_{j\in\mathcal{P}} \lambda_j \mathbf{b}_j \right\}, \quad s = 0, 1, \ldots,$$

where $\boldsymbol{\nu}(\mathbf{x}^s) \in N_X(\mathbf{x}^s)$ and $\mathbf{x}^{s+\frac{1}{2}} - \mathbf{x}^{s+1} \in N_X(\mathbf{x}^{s+1})$ for all s. The ergodic sequence $\{\boldsymbol{\mu}^t\}$ of Lagrange multipliers is defined by

$$\boldsymbol{\mu}^t = A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \left(\boldsymbol{\omega}(\mathbf{x}^s) + \frac{\boldsymbol{\lambda}(\mathbf{x}^s)}{\alpha_s} \right), \quad t = 1, 2, \ldots . \qquad \square$$

Note that $\mu_i^t \geq 0$ for all $i \in \mathcal{M}$ and all t.

Clearly, the value of the vector $\boldsymbol{\omega}(\mathbf{x}^s)$ depends of the choice of the normal term $\boldsymbol{\nu}(\mathbf{x}^s)$ in the conditional subgradients used in the method (1.5). In the special case of the traditional subgradient optimization method (1.4), $\boldsymbol{\nu}(\mathbf{x}^s) = \mathbf{0}$ for all s, and then, from Lemma 4.4, $\boldsymbol{\omega}(\mathbf{x}^s) = \mathbf{0}$ for all s. In the special case of the method (1.5) where $\boldsymbol{\gamma}^X(\mathbf{x}^s)$ is a *projected subgradient*,

$$\boldsymbol{\nu}(\mathbf{x}^s) = \arg\min_{\boldsymbol{\nu}} \left\{ \|\boldsymbol{\nu}\|^2 \;\middle|\; \nabla a_i(\mathbf{x}^s)^T (\boldsymbol{\nu} + \boldsymbol{\gamma}(\mathbf{x}^s)) \geq 0, i \in \mathcal{I}(\mathbf{x}^s);\; \mathbf{b}_j^T (\boldsymbol{\nu} + \boldsymbol{\gamma}(\mathbf{x}^s)) = 0, j \in \mathcal{P} \right\},$$

and the elements $\omega_i(\mathbf{x}^s)$, $i \in \mathcal{I}(\mathbf{x}^s)$, and $\omega_j(\mathbf{x}^s)$, $j \in \mathcal{P}$, are Lagrange multipliers associated with the constraints of this projection problem.

The vector $\boldsymbol{\lambda}(\mathbf{x}^s)$ always contains Lagrange multipliers associated with the constraints of the projection problem

$$\mathbf{x}^{s+1} = \arg\min_{\mathbf{x}} \left\{ \left\| \mathbf{x}^{s+\frac{1}{2}} - \mathbf{x} \right\|^2 \,\middle|\, a_i(\mathbf{x}) \leq 0, \; i \in \mathcal{M}; \quad \mathbf{b}_j^T \mathbf{x} = e_j, \; j \in \mathcal{P} \right\}.$$

Lemma 4.6 (Boundedness of $\{\boldsymbol{\omega}(\mathbf{x}^s)\}$ and $\{\boldsymbol{\lambda}(\mathbf{x}^s)/\alpha_s\}$). *Suppose that the set X is defined by (4.1) and fulfils Assumption 4.1, and let the method (1.5)–(1.7) be applied to the program* [P]. *If the sequences $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}$ and $\{\boldsymbol{\nu}(\mathbf{x}^s)\}$ are bounded, then this is also true for the sequences $\{\boldsymbol{\omega}(\mathbf{x}^s)\}$ and $\{\boldsymbol{\lambda}(\mathbf{x}^s)/\alpha_s\}$, given by Definition 4.5.*

Proof. We first prove, by contradiction, that the sequence $\{\boldsymbol{\omega}(\mathbf{x}^s)\}$ is bounded.

Suppose that this is not the case. Then there exists a subsequence $\mathcal{S}$ such that $\|\boldsymbol{\omega}(\mathbf{x}^s)\| > 0$ for all $s \in \mathcal{S}$ and $\{\|\boldsymbol{\omega}(\mathbf{x}^s)\|\}_{s\in\mathcal{S}} \to \infty$. Let $\boldsymbol{\pi}^s = \boldsymbol{\omega}(\mathbf{x}^s)/\|\boldsymbol{\omega}(\mathbf{x}^s)\|$, implying that $\boldsymbol{\pi}^s \geq \mathbf{0}$ and $\|\boldsymbol{\pi}^s\| = 1$ for all $s \in \mathcal{S}$. According to Proposition 1.6, $\{\mathbf{x}^s\} \to \mathbf{x}^\infty$. Suppose that, for some $i \in \mathcal{M}$, $a_i(\mathbf{x}^\infty) < 0$. By the continuity of the function a_i, it must then hold for all s which are sufficiently large that $a_i(\mathbf{x}^s) < 0$, $\omega_i(\mathbf{x}^s) = 0$, and $\pi_i^s = 0$; hence, for any $i \in \mathcal{M} \setminus \mathcal{I}(\mathbf{x}^\infty)$, $\pi_i^s = 0$ for all $s \in \mathcal{S}$ that are large enough. The sequence $\{\boldsymbol{\pi}^s\}_{s\in\mathcal{S}}$ is clearly bounded and thus has an accumulation point, say $\tilde{\boldsymbol{\pi}}$. Then, from the preceding, $\tilde{\boldsymbol{\pi}} \in \Pi_X(\mathbf{x}^\infty)$ and $\|\tilde{\boldsymbol{\pi}}\| = 1$.

From Definition 4.5 we have that

$$\sum_{i\in\mathcal{M}} \pi_i^s \nabla a_i(\mathbf{x}^s) + \sum_{j\in\mathcal{P}} \pi_j^s \mathbf{b}_j = \frac{\boldsymbol{\nu}(\mathbf{x}^s)}{\|\boldsymbol{\omega}(\mathbf{x}^s)\|}, \quad \forall s \in \mathcal{S}.$$

Taking the limit of the subsequence of $\mathcal{S}$ corresponding to $\tilde{\boldsymbol{\pi}}$ we obtain

$$\sum_{i\in\mathcal{I}(\mathbf{x}^\infty)} \tilde{\pi}_i \nabla a_i(\mathbf{x}^\infty) + \sum_{j\in\mathcal{P}} \tilde{\pi}_j \mathbf{b}_j = \mathbf{0},$$

since $\{\mathbf{x}^s\} \to \mathbf{x}^\infty$ and the functions ∇a_i, $i \in \mathcal{M}$, are continuous, and since the sequence $\{\boldsymbol{\nu}(\mathbf{x}^t)\}$ is bounded and the sequence $\{\|\boldsymbol{\omega}(\mathbf{x}^s)\|\}_{s\in\mathcal{S}}$ is presumed to tend to infinity. Then, by Lemma 4.4, $\tilde{\boldsymbol{\pi}} = \mathbf{0}$ must hold, and we have arrived at a contradiction. Hence, the sequence $\{\boldsymbol{\omega}(\mathbf{x}^s)\}$ is bounded.

To establish that the sequence $\{\boldsymbol{\lambda}(\mathbf{x}^s)/\alpha_s\}$ is bounded, we first observe that, from (1.5), $\|\mathbf{x}^{s+\frac{1}{2}} - \mathbf{x}^{s+1}\|/\alpha_s \leq \|\boldsymbol{\gamma}(\mathbf{x}^s) + \boldsymbol{\nu}(\mathbf{x}^s)\|$ for all s, which implies that the sequence $\{(\mathbf{x}^{s+\frac{1}{2}} - \mathbf{x}^{s+1})/\alpha_s\}$ is bounded. Assume then that the sequence $\{\boldsymbol{\lambda}(\mathbf{x}^s)/\alpha_s\}$ is not bounded and define the index set $\mathcal{S}$ analogously to the above. By, for all $s \in \mathcal{S}$, letting $\boldsymbol{\pi}^s = \boldsymbol{\lambda}(\mathbf{x}^s)/\|\boldsymbol{\lambda}(\mathbf{x}^s)\|$, Definition 4.5 yields that

$$\sum_{i\in\mathcal{M}} \pi_i^s \nabla a_i(\mathbf{x}^{s+1}) + \sum_{j\in\mathcal{P}} \pi_j^s \mathbf{b}_j = \frac{(\mathbf{x}^{s+\frac{1}{2}} - \mathbf{x}^{s+1})/\alpha_s}{\|\boldsymbol{\lambda}(\mathbf{x}^s)\|/\alpha_s}, \quad \forall s \in \mathcal{S}.$$

Clearly, the right hand side of this equality tends to the zero vector and we may analogously to the first part of the proof derive a contradiction. $\square$

From Definition 4.5 and Lemma 4.6 follow that, under the conditions (1.6) and (1.7), the sequence $\{\boldsymbol{\mu}^t\}$ is bounded.

Theorem 4.7 (Convergence of $\{\boldsymbol{\mu}^t\}$ to $M_X(\mathbf{x}^\infty)$). *Suppose that the set X is defined by (4.1) and fulfils Assumption 4.1. Let the method (1.5)–(1.7) be applied to the program*

[P] *and let the sequence* $\{\boldsymbol{\mu}^t\}$ *be given by Definition 4.5. Suppose that the sequences* $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}$ *and* $\{\boldsymbol{\nu}(\mathbf{x}^s)\}$ *are bounded. Then,*

$$\left\{d_{M_X(\mathbf{x}^\infty)}(\boldsymbol{\mu}^t)\right\} \to 0.$$

Proof. From the continuity of the functions a_i, $i \in \mathcal{M}$, follows that $\mathcal{I}(\mathbf{x}^s) \subseteq \mathcal{I}(\mathbf{x}^\infty)$ for all s that are sufficiently large. From Definition 4.5 then follows that $\omega_i(\mathbf{x}^s) = \lambda_i(\mathbf{x}^s) = 0$ for all $i \in \mathcal{M} \setminus \mathcal{I}(\mathbf{x}^\infty)$ and all s that are sufficiently large, so that $\{\mu_i^t\} \to 0$, $i \in \mathcal{M} \setminus \mathcal{I}(\mathbf{x}^\infty)$. Hence, $\{d_{\Pi_X(\mathbf{x}^\infty)}(\boldsymbol{\mu}^t)\} \to 0$.

To simplify the notation, we define $\boldsymbol{\nu}^t = \sum_{i\in\mathcal{M}} \mu_i^t \nabla a_i(\mathbf{x}^\infty) + \sum_{j\in\mathcal{P}} \mu_j^t \mathbf{b}_j$. From Definitions 2.2 and 4.5 then follow that

$$\begin{aligned}
\boldsymbol{\nu}^t \;&=\; \mathbf{n}^t - \sum_{i\in\mathcal{M}} A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \omega_i(\mathbf{x}^s)\left(\nabla a_i(\mathbf{x}^s) - \nabla a_i(\mathbf{x}^\infty)\right) \\
&\quad - \sum_{i\in\mathcal{M}} A_t^{-1} \sum_{s=0}^{t-1} \lambda_i(\mathbf{x}^s)\left(\nabla a_i(\mathbf{x}^{s+1}) - \nabla a_i(\mathbf{x}^\infty)\right).
\end{aligned} \tag{4.5}$$

Let $\varepsilon > 0$. Lemma 2.6 and Theorem 3.1 then yield that $d_{-\partial f(\mathbf{x}^\infty)}(\mathbf{n}^t) \le \varepsilon/3$ for all t that are sufficiently large. We next show that the second term of the right hand side of the equation (4.5) tends to the zero vector. Lemma 4.6 implies that the sequence $\{\omega_i(\mathbf{x}^s)\}$ is bounded for each $i \in \mathcal{M}$. Hence, since the function ∇a_i is continuous, there is an iterate $t(\varepsilon)$ such that

$$\omega_i(\mathbf{x}^s)\left\|\nabla a_i(\mathbf{x}^s) - \nabla a_i(\mathbf{x}^\infty)\right\| \le \frac{\varepsilon}{6m}, \quad \forall s \ge t(\varepsilon).$$

Therefore, for all $t > t(\varepsilon)$,

$$A_t^{-1} \sum_{s=t(\varepsilon)}^{t-1} \alpha_s \omega_i(\mathbf{x}^s)\left\|\nabla a_i(\mathbf{x}^s) - \nabla a_i(\mathbf{x}^\infty)\right\| \le A_t^{-1} \sum_{s=t(\varepsilon)}^{t-1} \alpha_s \frac{\varepsilon}{6m} \le \frac{\varepsilon}{6m}.$$

Since $\{A_t\} \to \infty$ we have, for all t that are large enough,

$$A_t^{-1} \sum_{s=0}^{t(\varepsilon)-1} \alpha_s \omega_i(\mathbf{x}^s)\left\|\nabla a_i(\mathbf{x}^s) - \nabla a_i(\mathbf{x}^\infty)\right\| \le \frac{\varepsilon}{6m}.$$

From the derivations above follow that, for all $t > t(\varepsilon)$ that are large enough,

$$\begin{aligned}
0 \;&\le\; \left\| \sum_{i\in\mathcal{M}} A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \omega_i(\mathbf{x}^s)\left(\nabla a_i(\mathbf{x}^s) - \nabla a_i(\mathbf{x}^\infty)\right) \right\| \\
&\le\; \sum_{i\in\mathcal{M}} A_t^{-1} \sum_{s=0}^{t-1} \alpha_s\, \omega_i(\mathbf{x}^s)\, \left\|\, \nabla a_i(\mathbf{x}^s) - \nabla a_i(\mathbf{x}^\infty) \,\right\| \;\le\; \frac{\varepsilon}{3}.
\end{aligned}$$

It can, analogously to the preceding, be shown that the same conclusion holds for the last term of the right hand side of equation (4.5). It follows that $d_{-\partial f(\mathbf{x}^\infty)}(\boldsymbol{\nu}^t) \le \varepsilon$ for all t that are sufficiently large, and hence that $\{d_{-\partial f(\mathbf{x}^\infty)}(\boldsymbol{\nu}^t)\} \to 0$. Now, let $\mathcal{T}$ be a subsequence such that

$$\lim_{\substack{t\in\mathcal{T}\\ t\to\infty}} d_{M_X(\mathbf{x}^\infty)}(\boldsymbol{\mu}^t) = \limsup_{t\to\infty} d_{M_X(\mathbf{x}^\infty)}(\boldsymbol{\mu}^t),$$

and let $\tilde{\boldsymbol{\mu}}$ be an accumulation point of the sequence $\{\boldsymbol{\mu}^t\}_{t\in\mathcal{T}}$. From the preceding and the closedness of the sets $\Pi_X(\mathbf{x}^\infty)$ and $\partial f(\mathbf{x}^\infty)$ then follow that $\tilde{\boldsymbol{\mu}} \in \Pi_X(\mathbf{x}^\infty)$ and $\sum_{i\in\mathcal{M}} \tilde{\mu}_i \nabla a_i(\mathbf{x}^\infty) + \sum_{i\in\mathcal{P}} \tilde{\mu}_j \mathbf{b}_j \in -\partial f(\mathbf{x}^\infty)$, thus implying that $\tilde{\boldsymbol{\mu}} \in M_X(\mathbf{x}^\infty)$. The theorem follows. $\square$

4.2 Multipliers of objective pieces

Throughout this subsection we assume that the nonsmooth objective function f of the program [P] has the form

$$f(\mathbf{x}) = \max_{k \in \mathcal{L}} c_k(\mathbf{x}), \quad \mathbf{x} \in \Re^n, \tag{4.6}$$

where the functions $c_k : \Re^n \mapsto \Re$, $k \in \mathcal{L} = \{1, \dots, \ell\}$, are convex and differentiable. We let $\mathcal{K}(\mathbf{x}) = \{ k \in \mathcal{L} \mid c_k(\mathbf{x}) = f(\mathbf{x}) \}$ denote the set of objective pieces that are active at $\mathbf{x} \in \Re^n$ and

$$\Pi_f(\mathbf{x}) = \left\{ \boldsymbol{\pi} \in \Re^\ell_+ \;\middle|\; \pi_k = 0,\ k \in \mathcal{L} \setminus \mathcal{K}(\mathbf{x});\quad \sum_{k \in \mathcal{L}} \pi_k = 1 \right\}, \quad \mathbf{x} \in \Re^n.$$

Letting $\boldsymbol{\xi} \in \Re^\ell$, the subdifferential of f at $\mathbf{x}$ can then be expressed as

$$\partial f(\mathbf{x}) = \left\{ \boldsymbol{\gamma} \in \Re^n \;\middle|\; \boldsymbol{\gamma} = \sum_{k \in \mathcal{L}} \xi_k \nabla c_k(\mathbf{x});\quad \boldsymbol{\xi} \in \Pi_f(\mathbf{x}) \right\}.$$

Let the program [P] be reformulated as

$$\begin{aligned} \min \quad & z, && (4.7\text{a}) \\ \text{s.t.} \quad & c_k(\mathbf{x}) - z \leq 0, \quad k \in \mathcal{L}, && (4.7\text{b}) \\ & \mathbf{x} \in X, && (4.7\text{c}) \end{aligned}$$

and introduce nonnegative multipliers η_k, $k \in \mathcal{L}$, associated with the constraints (4.7b). Optimality of an $\mathbf{x} \in X$ can then be characterized in terms of the nonemptiness of the set

$$M_f(\mathbf{x}) = \left\{ \boldsymbol{\eta} \in \Pi_f(\mathbf{x}) \;\middle|\; \mathbf{0} \in \sum_{k \in \mathcal{L}} \eta_k \nabla c_k(\mathbf{x}) + N_X(\mathbf{x}) \right\} \tag{4.8}$$

of Lagrange multipliers associated with the constraints (4.7b) (e.g., Hiriart-Urruty and Lemaréchal, 1993a, Theorem VII:2.1.4). Note that, in contrast to Propositions 4.2 and 4.3, a constraint qualification is here always fulfilled (since $\texttt{int}(\texttt{epi}\, f) \neq \emptyset$).

Proposition 4.8 (Optimality conditions). *Let $\mathbf{x} \in X$. Then, $\mathbf{x} \in X^*$ if and only if $M_f(\mathbf{x}) \neq \emptyset$.* □

Proposition 4.9 (Properties of $M_f(\mathbf{x})$, $\mathbf{x} \in X^*$). *For any $\mathbf{x} \in X^*$ the nonempty set $M_f(\mathbf{x})$ is convex and compact, and $M_f(\mathbf{y}) = M_f(\mathbf{x})$ for all $\mathbf{y} \in X^*$.*

Proof. The convexity and compactness of $M_f(\mathbf{x})$ are immediate from its definition (4.8), and its invariance over X^* follows from Proposition VII:3.1.1 in Hiriart-Urruty and Lemaréchal (1993a) applied to the reformulation (4.7) of the program [P]. □

Definition 4.10 (Ergodic sequence of multipliers). *Let the objective function f be defined by (4.6) and let the method (1.5) be applied to the program* [P]. *The sequence $\{\boldsymbol{\xi}(\mathbf{x}^s)\}$ is defined by*

$$\boldsymbol{\xi}(\mathbf{x}^s) \in \left\{ \boldsymbol{\xi} \in \Pi_f(\mathbf{x}^s) \;\middle|\; \boldsymbol{\gamma}(\mathbf{x}^s) = \sum_{k \in \mathcal{L}} \xi_k \nabla c_k(\mathbf{x}^s) \right\}, \quad s = 0, 1, \dots,$$

where $\boldsymbol{\gamma}(\mathbf{x}^s) \in \partial f(\mathbf{x}^s)$ for all s. The ergodic sequence $\{\boldsymbol{\eta}^t\}$ of multipliers is then defined by

$$\boldsymbol{\eta}^t = A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \boldsymbol{\xi}(\mathbf{x}^s), \quad t = 1, 2, \dots . \qquad \square$$

Note that $\boldsymbol{\eta}^t \geq \mathbf{0}$ and $\sum_{k \in \mathcal{L}} \eta_k^t = 1$ for all t; obviously, the sequence $\{\boldsymbol{\eta}^t\}$ is bounded.

Theorem 4.11 (Convergence of $\{\boldsymbol{\eta}^t\}$ to $M_f(\mathbf{x}^\infty)$). *Let the objective function f be defined by (4.6), let the method (1.5)–(1.7) be applied to the program* [P], *and let the sequence $\{\boldsymbol{\eta}^t\}$ be given by Definition 4.10. Suppose that the sequences $\{\boldsymbol{\gamma}(\mathbf{x}^t)\}$ and $\{\boldsymbol{\nu}(\mathbf{x}^t)\}$ are bounded. Then,*

$$\left\{ d_{M_f(\mathbf{x}^\infty)}(\boldsymbol{\eta}^t) \right\} \to 0.$$

Proof. Similarly to the proof of Theorem 4.7 we can show that $\{d_{\Pi_f(\mathbf{x}^\infty)}(\boldsymbol{\eta}^t)\} \to 0$. Defining $\boldsymbol{\gamma}^t = \sum_{k \in \mathcal{L}} \eta_k^t \nabla c_k(\mathbf{x}^\infty)$, it follows from Definitions 2.1 and 4.10 that

$$\boldsymbol{\gamma}^t = \mathbf{g}^t - \sum_{k \in \mathcal{L}} A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \xi_k(\mathbf{x}^s) \left(\nabla c_k(\mathbf{x}^s) - \nabla c_k(\mathbf{x}^\infty) \right).$$

Let $\varepsilon > 0$. Theorem 3.1 then yields that $d_{-N_X(\mathbf{x}^\infty)}(\mathbf{g}^t) \leq \varepsilon/2$ for all t that are sufficiently large. Using the same technique as in the proof of Theorem 4.7 and utilizing the boundedness of the sequence $\{\boldsymbol{\xi}(\mathbf{x}^s)\}$, we can show that

$$0 \leq \left\| \sum_{k \in \mathcal{L}} A_t^{-1} \sum_{s=0}^{t-1} \alpha_s \xi_k(\mathbf{x}^s) \left(\nabla c_k(\mathbf{x}^s) - \nabla c_k(\mathbf{x}^\infty) \right) \right\| \leq \frac{\varepsilon}{2}.$$

It follows that $d_{-N_X(\mathbf{x}^\infty)}(\boldsymbol{\gamma}^t) \leq \varepsilon$ for all t that are sufficiently large, and hence that $\{d_{-N_X(\mathbf{x}^\infty)}(\boldsymbol{\gamma}^t)\} \to 0$. Using the same technique as in the proof of Theorem 4.7, we can also show that all accumulation points, $\tilde{\boldsymbol{\eta}}$, of the sequence $\{\boldsymbol{\eta}^t\}$ fulfil $\tilde{\boldsymbol{\eta}} \in \Pi_f(\mathbf{x}^\infty)$ and $\sum_{k \in \mathcal{L}} \tilde{\eta}_k \nabla c_k(\mathbf{x}^\infty) \in -N_X(\mathbf{x}^\infty)$, thus implying $\tilde{\boldsymbol{\eta}} \in M_f(\mathbf{x}^\infty)$. The theorem follows. $\square$

5 FURTHER RESEARCH

The ergodic results of Section 4 can immediately be generalized to the case when both the objective function and the feasible set have explicit representations. A straightforward topic for continued studies is the derivation of ergodic results for multipliers when the constraint functions a_i, $i \in \mathcal{M}$, and/or the objective pieces c_k, $k \in \mathcal{L}$, are allowed to be nonsmooth.

We have recently constructed a convergent lower bounding procedure for the program [P] (cf. the lower bounding principle of the Frank–Wolfe method) based on the ergodic results presented in this paper, and studied its application in a subgradient scheme for a nonsmooth formulation of the linear minimum cost multicommodity network flow problem (Larsson et al., 1995). The application of the lower bounding procedure to other classes of structured, nonsmooth programs, especially Lagrangean dual problems, will be studied further.

A further application of the sequence $\{\mathbf{g}^t\}$ is the construction of an auxiliary, optimizing sequence $\{\overline{\mathbf{x}}^t\}$ of feasible solutions to the program [P], by

$$\overline{\mathbf{x}}^t = P_X \left(\mathbf{x}^t - \sigma_t \mathbf{g}^t \right), \quad t = 1, 2, \dots,$$

where $0 < \underline{\sigma} \leq \sigma_t \leq \overline{\sigma} < +\infty$ for all t. This sequence, which is generated in a manner akin to gradient projection methods for smooth optimization, has stronger properties than the original sequence $\{\mathbf{x}^t\}$; for example, under very mild assumptions it identifies *finitely* the set of active constraints at the optimal solution $\mathbf{x}^\infty$. This is an extension of results from the differentiable case, and it illustrates well the smoothing effect of the process defining the sequence $\{\mathbf{g}^t\}$.

Consider the *general* problem of finding an $\mathbf{x} \in \Re^n$ such that $\mathbf{0} \in F(\mathbf{x})$, where $F : \Re^n \mapsto 2^{(\Re \cup \{\pm\infty\})^n}$ is a maximal monotone point–to–set map. A natural method for solving this problem can be described as follows. Let

$$\mathbf{x}^0 \in \mathtt{dom}\, F, \quad \gamma^t \in F(\mathbf{x}^t), \quad \alpha_t > 0, \quad \text{and} \quad \mathbf{x}^{t+1} = P_{\mathtt{dom}\, F}(\mathbf{x}^t - \alpha_t \gamma^t), \quad t = 0, 1, \dots . \tag{5.1}$$

For any nonempty, closed and convex set $X \subseteq \Re^n$ and any convex function $f : \Re^n \mapsto \Re$, the mapping $F \equiv \partial^X f$ is maximal monotone, since it equals the sum of the two maximal monotone operators ∂f and N_X (Rockafellar, 1970b, Theorem B, and 1970c, Theorem 1); by Proposition 1.5, the problem [P] is hence a special case of the general problem and, further, the conditional subgradient method (1.5) is a special case of the method (5.1). Proposition 1.6 then establishes that $\{d_{F^{-1}(\mathbf{0})}(\mathbf{x}^t)\} \to 0$, that is, $\mathbf{0} \in F(\mathbf{x}^\infty)$. A subject for future research is to derive ergodic results for the method (5.1) applied to the general problem.

Acknowledgement The research was supported financially by grants (No. 94-292 and 282-93-1195) from the Swedish Research Council for Engineering Sciences (TFR).

References

[1] J.E. Beasley, 1993, "Lagrangean relaxation", *in*: "Modern Heuristic Techniques for Combinatorial Problems," C.R. Reeves, ed., Blackwell Scientific Publications, Oxford, pp. 243–303.

[2] A. Ben-Tal and M.P. Bendsøe, 1993, "A new method for optimal truss topology design", *SIAM Journal on Optimization*, 3:322–358.

[3] J.V. Burke and M.C. Ferris, 1991, "Characterization of solution sets of convex programs", *Operations Research Letters*, 10:57–60.

[4] V.F. Dem'janov and V.K. Šomesova, 1978, "Conditional subdifferentials of convex functions", *Soviet Mathematics Doklady*, 19:1181–1185.

[5] V.F. Dem'yanov and V.K. Shomesova, 1980, "Subdifferentials of functions on sets", *Cybernetics*, 16/1:24–31.

[6] V.F. Dem'yanov and L.V. Vasil'ev, 1985, "Nondifferentiable Optimization," Optimization Software, New York, NY.

[7] J.C. Dunn, 1973, "On recursive averaging processes and Hilbert space extensions of the contraction mapping principle", *Journal of the Franklin Institute*, 295:117–133.

[8] Yu.M. Ermol'ev, 1966, "Methods for solving nonlinear extremal problems", *Cybernetics*, 2/4:1–14.

[9] M.L. Fisher, 1981, "The Lagrangian relaxation method for solving integer programming problems", *Management Science*, 27:1–18.

[10] M.L. Fisher, 1985, "An applications oriented guide to Lagrangian relaxation", *Interfaces*, 15:10–21.

[11] M. Frank and P. Wolfe, 1956, An algorithm for quadratic programming, *Naval Research Logistics Quarterly*, 3:95–110.

[12] J.-B. Hiriart-Urruty and C. Lemaréchal, 1993a, "Convex Analysis and Minimization Algorithms I: Fundamentals," Springer-Verlag, Berlin.

[13] J.-B. Hiriart-Urruty and C. Lemaréchal, 1993b, "Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods," Springer-Verlag, Berlin.

[14] J. Kornai and T. Lipták, 1965, "Two level planning", *Econometrica*, 33:141–169.

[15] T. Larsson, M. Patriksson, and A.-B. Strömberg, 1992, "Conditional subgradient optimization—theory and applications", Report LiTH-MAT-R-92-24, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden. Revised 1994. *European Journal of Operational Research* (to appear).

[16] T. Larsson, M. Patriksson, and A.-B. Strömberg, 1995, "Ergodic results and bounds on the optimal value in subgradient optimization", Paper presented at the Symposium on Operations Research (SOR'95), Annual Conference of the DGOR, GMÖOR, and ÖGOR, Universität Passau, September 13–15, Passau, Germany.

[17] B.T. Polyak, 1967, "A general method of solving extremum problems", *Soviet Mathematics Doklady*, 8/3:593–597.

[18] B.T. Polyak, 1969, "Minimization of unsmooth functionals", *USSR Computational Mathematics and Mathematical Physics*, 9:14–29.

[19] W.B. Powell and Y. Sheffi, 1982, "The convergence of equilibrium algorithms with predetermined step sizes", *Transportation Science*, 16:45–55.

[20] R.T. Rockafellar, 1970a, "Convex Analysis," Princeton University Press, Princeton, NJ.

[21] R.T. Rockafellar, 1970b, "On the maximal monotonicity of subdifferential mappings", *Pacific Journal of Mathematics*, 33:209–216.

[22] R.T. Rockafellar, 1970c, "On the maximality of sums of nonlinear monotone operators", *Transactions of the American Mathematical Society*, 149:75–88.

[23] N.Z. Shor, 1985, "Minimization Methods for Non-Differentiable Functions," Springer-Verlag, Berlin.

[24] N.Z. Shor, 1991, "The development of numerical methods for nonsmooth optimization in the USSR", *in*: "History of Mathematical Programming: A Collection of Personal Reminiscences," J.K. Lenstra, A.H.G. Rinnoy Kan and A. Schrijver, eds., CWI and North-Holland, Amsterdam, pp. 135–139.

[25] E.H. Zarantonello, 1971, "Projections on convex sets in Hilbert space and spectral theory", *in*: "Contributions to Nonlinear Functional Analysis," E.H. Zarantonello, ed., Proceedings of a symposium conducted by the Mathematics research center, University of Wisconsin, Madison, April 1971, Academic Press, New York, NY, pp. 237–424.