SIAM J. OPTIM.
Vol. 8, No. 2, pp. 561–582, May 1998

014

# COST APPROXIMATION: A UNIFIED FRAMEWORK OF DESCENT ALGORITHMS FOR NONLINEAR PROGRAMS*

MICHAEL PATRIKSSON†

**Abstract.** This paper describes and analyzes the cost approximation algorithm. This class of iterative descent algorithms for nonlinear programs and variational inequalities places a large number of algorithms within a common framework and provides a means for analyzing relationships among seemingly unrelated methods. A common property of the methods included in the framework is that their subproblems may be characterized by monotone mappings, which replace an additive part of the original cost mapping in an iterative manner; alternately, a step is taken in the direction obtained in order to reduce the value of a merit function for the original problem. The generality of the framework is illustrated through examples, and the convergence characteristics of the algorithm are analyzed for applications to nondifferentiable optimization. The convergence results are applied to some example methods, demonstrating the strength of the analysis compared to existing results.

**Key words.** nondifferentiable optimization, cost approximation, partial linearization, descent algorithms, convergence analysis

**AMS subject classifications.** 90C30, 49M37

**PII.** S105262349427577X

## 1. Introduction.

Let $u : \Re^n \mapsto \Re \cup \{+\infty\}$ be lower semicontinuous (l.s.c.), proper, and convex and $f : \Re^n \mapsto \Re \cup \{+\infty\}$ be continuously differentiable on an open neighborhood of dom $u$. Consider the *nondifferentiable optimization problem*

$$\text{[NDP]} \qquad \min_{x \in \Re^n} T(x) := f(x) + u(x).$$

This problem is generic in mathematical programming and encompasses the problem of minimizing a convex and/or continuously differentiable real-valued function over a nonempty, closed, and convex set in $\Re^n$.

The most common approach to solving [NDP] is to construct a sequence $\{x^t\}$ of iterates in dom $u$ such that the sequence $\{T(x^t)\}$ is strictly monotonically decreasing; typically, the sequence $\{x^t\}$ converges to a solution $x^*$ to the *generalized equation* [Rob79]

$$\text{[GE]} \qquad \nabla f(x^*) + \partial u(x^*) \ni 0,$$

which, under the mild assumption below (cf. [Roc66]), constitutes the *necessary* optimality conditions of $x^*$ in [NDP] and further is *sufficient* for the global optimality of $x^*$ in [NDP] whenever either $f$ is convex or $u$ is an indicator function of a nonempty, closed, and convex set and $f$ is pseudoconvex.

The directional derivative of $u$ in the direction of $d$ is

$$u'(x; d) := \lim_{\lambda \downarrow 0} \frac{1}{\lambda}[u(x + \lambda d) - u(x)].$$

*Received by the editors October 17, 1994; accepted for publication (in revised form) February 22, 1997. This research was supported in part by grants from the Swedish Institute (303 GH/MLH) and the Swedish Research Council for Engineering Sciences (TFR) (282-93-1195).
http://www.siam.org/journals/siopt/8-2/27577.html

†Department of Mathematics, Chalmers University of Technology, S-412 96 Gothenburg, Sweden (mipat@math.chalmers.se). Formerly with Department of Mathematics, Linköping Institute of Technology, S-581 83 Linköping, Sweden.

*Assumption* 1.1 (lower semicontinuity of $u'(x;\cdot)$). For any $x \in \operatorname{dom} u$ and $d \in \Re^n$,

$$u'(x;d) = \liminf_{e \to d} u'(x;e). \tag{1}$$

Note that (1) is fulfilled for any $x \in \operatorname{rint}(\operatorname{dom} u)$ [Roc70a, Thm. 23.4]. A direct consequence of the assumption is that for any $x \in \operatorname{dom} u$ and $d \in \Re^n$ the directional derivative may be calculated as ([Roc66] and [Roc70a, Thm. 23.2])

$$u'(x;d) = \sup\{\, \xi_u(x)^{\mathrm{T}} d \mid \xi_u(x) \in \partial u(x) \,\}. \tag{2}$$

We let $\Omega$ denote the set of solutions to [GE], and assume that it is nonempty; a sufficient condition for this to be the case is that $T$ is *coercive*, that is, $\operatorname{dom} T$ is bounded or

$$\lim_{\|x\| \to \infty} \{T(x)/\|x\|\} = +\infty. \tag{3}$$

Moreover, $\Omega$ is a singleton set if it is nonempty and if $T$ is strictly convex on its effective domain. Note that *strong convexity* of $T$, that is, there exists a constant $m_T > 0$ such that for any $\xi_T(x) \in \partial T(x)$ and $\xi_T(y) \in \partial T(y)$, $x, y \in \operatorname{dom} T$,

$$[\xi_T(x) - \xi_T(y)]^{\mathrm{T}}(x - y) \geq m_T \|x - y\|^2, \tag{4}$$

implies both strict convexity and coercivity of $T$ and hence ensures the existence of a unique solution to [NDP]. (The expression (4) actually states that the subdifferential mapping $\partial T$ is strongly monotone on $\operatorname{dom} T$; we may use it as a definition of convexity of $T$ by allowing $m_T = 0$.) Introducing some additional terms, a mapping $\Pi$ is *maximal monotone* [Min62, Bro68] if it is monotone and its graph is not properly contained in the graph of any other monotone operator; its inverse mapping is denoted by $\Pi^{-1}$. Further, a mapping $\Pi$ is *Lipschitz continuous* on $\operatorname{dom} \Pi$ (with modulus $M_\Pi \geq 0$) if

$$\|\Pi(x) - \Pi(y)\| \leq M_\Pi \|x - y\| \qquad \forall x, y \in \operatorname{dom} \Pi.$$

An interesting application of [NDP] results from letting $u$ be the *indicator function* of a nonempty, closed, and convex subset $X$ of $\Re^n$, that is, $u \equiv \delta_X$, where

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

The subdifferential mapping of $\delta_X$ is the *normal cone* operator associated with the set $X$,

$$N_X(x) := \begin{cases} \{z \in \Re^n \mid z^{\mathrm{T}}(y - x) \leq 0 \quad \forall y \in X\}, & x \in X, \\ \emptyset, & x \notin X. \end{cases} \tag{5}$$

(This is an example of a maximal monotone operator.) The problem [GE] then reduces to the problem of finding a vector $x^*$ such that

$$\nabla f(x^*) + N_X(x^*) \ni 0,$$

which (without the need for Assumption 1.1) describes the first-order necessary conditions for the optimality of $x^* \in X$ in the constrained differentiable program

$$\min_{x \in X} f(x). \tag{CDP}$$

We also note that the problem [NDP] encompasses *dual* formulations of convex programs (e.g., [FHN96]).

A large amount of literature has been devoted to the study of computational methods for [NDP]. The present paper provides a framework of algorithms for [NDP] which includes many methods previously analyzed and provides a convergence analysis which unifies and improves upon existing ones. The analysis includes investigations of inexact solutions of the auxiliary problems, different step length rules, and convergence rate results.

**2. The cost approximation algorithm.** The term *cost approximation* (CA) was coined in the author's Ph.D. thesis [Pat93a] to describe a framework of descent algorithms for nonlinear programs and variational inequality problems. In this section we present the algorithm framework for the solution of [NDP]. The main idea is to iteratively approximate the cost mapping $\nabla f$ in order to obtain a problem which is more easily solved in the sense that a (possibly) nonmonotone mapping is replaced by a monotone one. (The term cost mapping is taken from applications of [GE] in equilibrium problems, where $\nabla f$ usually represents a cost vector.) The solution to this problem defines a search direction in which a step is taken to decrease the value of $T$.

**2.1. The subproblem phase.** Let $x \in \operatorname{dom} u$. We introduce a continuous and monotone *cost approximating mapping* $\Phi : \operatorname{dom} u \mapsto \Re^n$. If the mapping $\nabla f$ is replaced by $\Phi$, then the error made in the approximation is obviously $\Phi - \nabla f$. This error is taken into account by subtracting from $\Phi$ the fixed error term $\Phi(x) - \nabla f(x)$. The approximation may alternatively be interpreted as a fixation of the second term of the original cost $[\Phi + \partial u] + [\nabla f - \Phi]$ at $x$.

Thus, we arrive at a generalized equation in which a point $y$ is sought such that

$$[\mathrm{GE}_\Phi] \qquad \Phi(y) + \partial u(y) + \nabla f(x) - \Phi(x) \ni 0.$$

We let $Y(x)$ denote the set of solutions $y$ to $[\mathrm{GE}_\Phi]$. (Note that a fixed term may be added to $\Phi$ without affecting this subproblem.)

By construction, the approximation is exact at $x$. This fact is important since it provides a termination criterion for the algorithm: if $x$ solves the subproblem $[\mathrm{GE}_\Phi]$, then it immediately follows that $x$ also solves [GE]. The reverse is also true, and the subproblem thus provides a reformulation of [GE] as a fixed point problem in the (possibly point–to–set) mapping $x \mapsto Y(x)$.

*Remark* 2.1. The descent property of the search direction requires the mapping $\Phi$ to be monotone; this crucial property is, seemingly, not possible to relax within this framework.

In several known instances of CA algorithms, the corresponding mapping $\Phi$ may be identified as an approximation of $\nabla f$ (cf. Newton-type methods); the properties of $\nabla f$ in a given application may hence, implicitly through the requirements on $\Phi$, restrict the possible choices of CA methods.

The convergence analysis provided in this paper concentrates on approximations made of the differentiable (and possibly nonconvex) function $f$ and does not cover approximations of the convex function $u$, although such algorithms are easily devised within the framework; examples are subgradient optimization methods (e.g., [Sho85]) and the methods in [CoZ84]. Note that in such methods, however, the corresponding subproblems need not yield directions of descent with respect to $T$ without proper modifications. □

If $\Phi$ is chosen as the gradient mapping of a function $\varphi : \text{dom } \mapsto \Re$, which then is convex and continuously differentiable on dom $u$, then the subproblem [$\text{GE}_\Phi$] reduces to solving the convex subproblem

$$[\text{NDP}_\varphi] \quad \min_{y \in \Re^n} T_\varphi(y, x) := \varphi(y) + u(y) + f(x) - \varphi(x) + [\nabla f(x) - \nabla \varphi(x)]^{\text{T}}(y - x).$$

To give an example, assume that $f$ is a convex function in $C^2$ on dom $u$ and choose $\varphi(y, x) := \frac{1}{2}(y-x)^{\text{T}}\nabla^2 f(x)(y-x)$, from which one obtains $T_\varphi(y, x) := u(y) + \{f(x) + \nabla f(x)^{\text{T}}(y-x) + \frac{1}{2}(y-x)^{\text{T}}\nabla^2 f(x)(y-x)\}$; an extension of Newton's method to nondifferentiable optimization is obtained, in which only the differentiable part of the objective is approximated. The choice $\varphi(y, x) := \frac{1}{2\gamma}\|y - x\|^2$, $\gamma > 0$, leads to extensions of steepest descent, gradient projection, and proximal point algorithms.

The construction of this subproblem may be given an alternative interpretation as a *partial linearization* [Pat93a] of $T$; expressing it as $[\varphi + u] + [f - \varphi]$, the function $T_\varphi$ is obtained from a first-order approximation (or linearization) of the second term at $x$.

Thus, by applying the CA concept, the problem [NDP] can be solved through a sequence of convex optimization problems. Indeed, in most of the iterative methods for the solution of [NDP] that one may identify as special cases from the class of CA algorithms, subproblems of the form [$\text{NDP}_\varphi$] are solved. (Examples where the subproblem [$\text{GE}_\Phi$] does not reduce to the convex optimization problem [$\text{NDP}_\varphi$] include some matrix splitting methods for quadratic programs [CPS92] and Gauss–Seidel/SOR methods [OrR70].)

**2.2. The line search phase.** In general, we cannot expect the original problem to be solved by the subproblem solution $y$. An improved solution is therefore defined through a step taken in the direction of $d := y - x$ such that the value of a merit function for [NDP] is reduced sufficiently. Although any merit function with sufficient continuity properties will do (such as the Euclidean distance to $\Omega$), we choose $T$ as the merit function. (In the solution of asymmetric variational inequality problems, *gap functions* associated with the CA subproblem [$\text{NDP}_\varphi$] are available; see [Pat93c, LaP94, Pat94a, Pat94b, Pat97, Pat98] for further details.)

At the new point, the original mapping is again approximated—perhaps using another mapping $\Phi$—and the algorithm proceeds until some stopping criterion is fulfilled.

**2.3. The conceptual algorithm.** A description of the CA algorithm is given in Table 1. A sequence $\{\Phi^t\}$ of monotone mappings is assumed to be given. (Note, however, that each mapping may also be constructed adaptively, given $x^s$, $s = 1, 2, \ldots, t$, instead of being chosen a priori.)

It will be made clear in the rest of the paper how accurately the steps **1** and **3** need to be performed.

**2.4. Conventions and assumptions.** In the convergence analysis of the CA algorithm, three forms of the cost approximating mapping will be used. In the most general case, a sequence $\{\Phi^t\}$ of continuous and monotone mappings on dom $u$ is used; the mappings in this sequence are neither assumed to be dependent on each other nor to form a limit function (that is, equicontinuity is not assumed). In the second case, it is assumed that there exists a known continuous mapping $\Phi : \text{dom } u \times \text{dom } u \mapsto \Re^n$ which is monotone on dom $u$ in its first argument; each individual cost approximating mapping $\Phi^t$ in the sequence $\{\Phi^t\}$ then has the form $\Phi(\cdot, x^t)$. (This form will be

TABLE 1
*The CA algorithm.*

**0**. (Initialization). Choose an initial point $x^0 \in \text{dom } u$, and let $t := 0$.
**1**. (Search direction generation). Find a solution $y^t$ to

$[\text{GE}_{\Phi^t}]$

$$\Phi^t(y^t) + \partial u(y^t) + \nabla f(x^t) - \Phi^t(x^t) \ni 0. \tag{6}$$

The resulting search direction is $d^t := y^t - x^t$.
**2**. (Termination criterion). If $x^t$ solves $[\text{GE}_{\Phi^t}] \rightarrow$ Stop ($x^t \in \Omega$). Otherwise, continue.
**3**. (Line search). Choose a step length, $\ell_t$, such that $x^t + \ell_t d^t \in \text{dom } u$ and the value of $T$ is reduced sufficiently.
**4**. (Update). Let $x^{t+1} := x^t + \ell_t d^t$ and $t := t + 1$.
**5**. (Termination criterion). If $x^t$ is acceptable $\rightarrow$ Stop. Otherwise, go to Step 1.

required whenever we do not wish to impose strong monotonicity conditions on $\Phi$.) The third, and most simple, form of mapping to be used is the iteration independent form, that is, $\Phi^t \equiv \Phi$ for all $t$ for a given continuous and monotone mapping $\Phi : \text{dom } u \mapsto \Re^n$. (The latter form will only be used in the linear convergence Theorem 4.6 and its Corollary 5.1.) Whenever the iteration dependency of $\Phi^t$ is insignificant in the context, the superscript will be suppressed. (This applies especially to the technical lemmas in section 3.)

In general, the CA algorithm is only asymptotically convergent. (Finite convergence is obtained when the problem enjoys a *sharpness* property; see [Pat93c, Pat98] for further details.) Therefore, we shall implicitly presume that the sequence of iterates is infinite. Further, we do not study the behavior of the algorithm when a solution does not exist.

The remainder of the paper is organized as follows. In the next section, we provide some properties of the search directions and the step length rules considered for different assumptions on the mappings $\Phi^t$. In section 4, a convergence analysis is made. Some consequences of this analysis are described in section 5, demonstrating its strength compared to previous analyses of algorithms included in the framework.

## 3. Preliminaries.

### 3.1. Analysis of the search directions.

LEMMA 3.1 (characterizations of the search directions). *Let $x \in \text{dom } u$, $Y(x)$ be the (possibly empty) set of solutions to $[\text{GE}_\Phi]$, $y \in Y(x)$, and $d := y - x$.*

(a) (A fixed point characterization of $\Omega$).

(1) $x \in \Omega \Longleftrightarrow x \in Y(x)$.

(2) *Let $\Phi \equiv \nabla\varphi$. Then, $x \in \Omega \Longleftrightarrow x \in Y(x) \Longleftrightarrow T_\varphi(y, x) = T_\varphi(x, x)$.*

(b) (A characterization of $Y(x)$). *Let $\Phi$ be maximal monotone. Then, $\Phi + \partial u$ is maximal monotone and*

$$Y(x) = [\Phi + \partial u]^{-1}[\Phi - \nabla f](x). \tag{7}$$

(c) (A well-posedness characterization).

(1) *Let $\Phi$ be maximal monotone. Then, the set $Y(x)$ is nonempty and locally bounded if*

$$\Phi(x) - \nabla f(x) \in \text{int } (\text{dom } ([\Phi + \partial u]^{-1})). \tag{8}$$

(2) *Let* $\Phi \equiv \nabla\varphi$. *Then, the set* $Y(x)$ *is nonempty and bounded if and only if*

$$\nabla\varphi(x) - \nabla f(x) \in \text{int } (\text{dom } ([\varphi + u]^\circ)), \tag{9}$$

*where* $[\varphi + u]^\circ$ *is the conjugate function of* $\varphi + u$.

(d) *Let* $\Phi : \text{dom } u \times \text{dom } u \mapsto \Re^n$ *be a continuous mapping on* dom $u \times$ dom $u$ *of the form* $\Phi(y, x)$ *and monotone on dom* $u$ *with respect to* $y$. *Let* $S \subseteq \text{dom } u$ *be an arbitrary, closed set in* $\Re^n$. *Then, the mapping* $x \mapsto D(x) := Y(x) - x$ *is closed on* $S$.

(e) *Let* $u$ *be strictly convex on* dom $u$ *or* $\Phi$ *strictly monotone on* dom $u$. *If* $Y(x)$ *is nonempty, then it is a singleton.*

*Proof.*

(a) (1) Follows immediately from the construction of [$\text{GE}_\Phi$].
(2) The second equivalence follows immediately from the construction of [$\text{NDP}_\varphi$].

(b) The first result follows from [Roc70b, Roc70c]; the second result then follows by direct calculation.

(c) (1) The set $Y(x)$ is characterized by (7). Let $\Pi := \Phi + \partial u$. Since, by definition, range $\Pi = \text{dom } \Pi^{-1}$, the set $Y(x)$ is nonempty if and only if $[\Phi - \nabla f](x) \in \text{dom } \Pi^{-1}$; further, it is locally bounded if $[\Phi - \nabla f](x) \in \text{int } (\text{dom } \Pi^{-1})$ [Roc69].
(2) From [Roc70a, Cor. 23.5.1 and Thm. 27.1.b], the set $Y(x)$ is characterized by $Y(x) = \partial([\varphi + u]^\circ)(\nabla\varphi(x) - \nabla f(x))$. By [Roc70a, Thms. 23.4 and 27.1.d], $Y(x)$ is nonempty and bounded if and only if (9) holds.

(d) Let $\{x^t\} \subseteq S$ and $\{y^t\}$ be sequences in dom $u$ with $y^t \in Y(x^t)$, $\{x^t\} \to x$ and $\{y^t\} \to y$. By the continuity of $\Phi$ and $\nabla f$ and the convexity of $u$, Theorem 24.4 of [Roc70a] yields that $y \in Y(x)$. Hence, the mapping $x \mapsto Y(x)$ is closed on $S$. The graph of the mapping $D$ is obtained from that of $Y$ through an affine transformation; it then follows from [Ber63, p. 111] that $x \mapsto D(x)$ is closed on $S$, as desired.

(e) Follows from the strict monotonicity of $\Phi + \partial u$. □

*Remark* 3.1. The fixed point property in (a) validates the termination criterion of Step 2 of the CA algorithm; the result (a)(2) shows that it is also easily checked whenever $\Phi$ is a gradient mapping.

The characterization (7) of the set $Y(x)$ made in (b) bears resemblance to iterative formulas describing proximal point methods [Mar70, Roc76] and forward–backward splitting methods [Bre73, EcB92] for finding a zero of the sum of two maximal monotone mappings. The class of CA algorithms is actually far more general; for example, the mapping $\Phi$ is not necessarily affine or strongly monotone and can be chosen to adapt to problem structures (such as separability in $u$ and $f$), and a (possibly inexact) line search is embedded in the algorithm. We also note that a monotone and single-valued mapping on $\Re^n$ is automatically maximal monotone.

The properties of $Y(x)$ characterized in the result (c) are natural well-posedness requirements on the subproblem [$\text{GE}_\Phi$]. A sufficient condition for (8) to hold is that $\Phi + \partial u$ is *weakly coercive* [Zei90, Def. 32.34.b], since this condition implies that dom $([\Phi + \partial u]^{-1}) = \Re^n$ (see [Zei90, Cor. 32.35] and [Bro68]). The corresponding sufficient condition in the case where $\Phi$ is a gradient is that $\varphi + u$ is *coercive* (cf. (3)), since $\varphi + u$ then is *cofinite* [Roc70a, pp. 116 and 259] and dom $([\varphi + u]^\circ) = \Re^n$ holds [Roc70a, Cor. 13.3.1]. All these conditions are satisfied if dom $u$ is bounded; an

important special case is when $u$ includes, as an additive term, an indicator function of a compact set.

A function $\varphi : \mathrm{dom}\ u \times \mathrm{dom}\ u \mapsto \Re$ of the form $\varphi(y, x)$, continuous on $\mathrm{dom}\ u \times \mathrm{dom}\ u$ and convex and differentiable on $\mathrm{dom}\ u$ with respect to $y$, automatically has the continuity property of the mapping $\Phi = \nabla_y \varphi$ required in the result (d). Note that $\mathrm{dom}\ u$ is not closed in general [Roc70a, p. 52], in which case it cannot take the role of the set $S$. □

LEMMA 3.2 (descent properties). *Let $x \in \mathrm{dom}\ u \setminus \Omega$, $Y(x)$ be the (possibly empty) set of solutions to $[\mathrm{GE}_\Phi]$, $y \in Y(x)$, and $d := y - x$.*

(a) *If either $u$ is strictly convex on $\mathrm{dom}\ u$ and $\partial u(x)$ is bounded, or $\Phi$ is strictly monotone on $\mathrm{dom}\ u$, then $T'(x; d) < 0$.*

(b) *Let $\Phi \equiv \nabla\varphi$, $\overline{y} \in \Re^n$ be any point such that $T_\varphi(\overline{y}, x) < T_\varphi(x, x)$, and $\overline{d} := \overline{y} - x$. Then, $T'(x; \overline{d}) < 0$. Especially, $T'(x; d) < 0$.*

(c) *Let $u$ be strongly convex on $\mathrm{dom}\ u$ or $\Phi$ strongly monotone on $\mathrm{dom}\ u$. Let $\overline{y} \in \Re^n$ be an approximate solution to $[\mathrm{GE}_\Phi]$ in the sense that for some vector $r \in \Re^n$,*

$$\Phi(\overline{y}) + \partial u(\overline{y}) + \nabla f(x) - \Phi(x) \ni r, \tag{10}$$

*and $\overline{d} := \overline{y} - x$. If $\|r\| < (m_u + m_\Phi)\|\overline{d}\|$, then $T'(x; \overline{d}) < 0$. Especially, $Y(x)$ is nonempty and singleton and*

$$T'(x; d) \le -(m_u + m_\Phi)\|d\|^2. \tag{11}$$

(d) *Let $\Phi$ be strongly monotone on $\mathrm{dom}\ u$ and Lipschitz continuous on $\mathrm{dom}\ u$. Then, there exists a $\xi_u(y) \in \partial u(y)$ such that*

$$-\frac{T'(x; d)}{\|\nabla f(x) + \xi_u(y)\| \cdot \|d\|} \ge \rho, \tag{12}$$

*where $\rho := m_\Phi / M_\Phi$.*

*Remark* 3.2. The reader should note here that in (11) we make use of the convention that if a mapping $\Pi$ is monotone, then it satisfies (4) with $m_\Pi = 0$. □

*Proof.*

(a) If $u$ is strictly convex, then for all $\xi_u(x) \in \partial u(x)$, $[-\Phi(y) - \nabla f(x) + \Phi(x) - \xi_u(x)]^{\mathrm{T}} d > 0$ holds, which, from the monotonicity of $\Phi$, yields

$$[\nabla f + \xi_u(x)]^{\mathrm{T}} d < 0 \qquad \forall \xi_u(x) \in \partial u(x). \tag{13}$$

If $\partial u(x)$ is bounded, then the supremum of the left-hand side of the above inequality over $\xi_u(x) \in \partial u(x)$ is attained, and the result follows.

If $\Phi$ is strictly monotone, then the monotonicity of $\partial u$ [Roc70a, Cor. 31.5.2] yields $[\nabla f + \xi_u(x)]^{\mathrm{T}} d \le [\Phi(x) - \Phi(y)]^{\mathrm{T}} d < 0$, which implies the desired result.

(b) By the convexity of $\varphi$, $T_\varphi(\overline{y}, x) < T_\varphi(x, x)$ implies that $\nabla f(x)^{\mathrm{T}} \overline{d} < u(x) - u(\overline{y})$. Then, from (2),

$$\begin{aligned} T'(x; \overline{d}) &= \nabla f(x)^{\mathrm{T}} \overline{d} + \sup\{\, \xi_u(x)^{\mathrm{T}} (\overline{y} - x) \mid \xi_u(x) \in \partial u(x) \,\} \\ &< u(x) - u(\overline{y}) + \sup\{\, \xi_u(x)^{\mathrm{T}} \overline{d} \mid \xi_u(x) \in \partial u(x) \,\}, \end{aligned} \tag{14}$$

and the right-hand side of (14) is nonpositive since $u$ is convex. The point $y$ is just a special case of such a $\overline{y}$.

(c) We obtain from (10) that

$$[\nabla f(x) + \xi_u(x)]^{\mathrm{T}} \overline{d} \leq -(m_u + m_\Phi)\|\overline{d}\|^2 + \|r\| \cdot \|\overline{d}\|$$

holds for all $\xi_u(x) \in \partial u(x)$, from which the result easily follows.

(d) By $[\mathrm{GE}_{\Phi^t}]$ and the Lipschitz continuity of $\Phi$, there exists a $\xi_u(y) \in \partial u(y)$ such that $\|\nabla f(x) + \xi_u(y)\| \leq M_\Phi \|d\|$. Together with (11), with $m_u = 0$, (12) follows. ∎

*Remark* 3.3. The second result of (a) requires that $\Phi$ is strictly monotone if it is not a gradient mapping (cf. the result (b)); that nonstrict monotonicity is not sufficient is clear from the following example: take $u := \delta_X$, where $X := \{\, x \in \Re^2 \mid 0 \leq x \leq 2 \,\}$, and $f(x) := \frac{1}{2}(x_1^2 + x_2^2)$. Let further $x := (2,2)^{\mathrm{T}}$, which clearly is nonoptimal, and

$$\Phi(x) := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} x,$$

which is a nonstrictly monotone and asymmetric mapping. By simple calculations, $y = (0,4)^{\mathrm{T}}$ can be shown to solve $[\mathrm{GE}_\Phi]$, but $\nabla f(x)^{\mathrm{T}}(y - x) = 0$.

The result (b) establishes that descent properties hold under very mild assumptions if $\Phi$ is a gradient mapping and validates inexact solutions of $[\mathrm{NDP}_\varphi]$. (In the above example, choose, for example, $\Phi := 0$; then $y = (0,0)^{\mathrm{T}}$ solves $[\mathrm{GE}_\Phi]$, and $\nabla f(x)^{\mathrm{T}}(y - x) = -8$.)

The result (c) establishes the descent property under conditions on the accuracy of the solution of $[\mathrm{GE}_\Phi]$. While solving $[\mathrm{GE}_\Phi]$ with some iterative procedure, the validity of (10) for a given value of $r$ can often be checked easily.

The vector $r$ appearing in (10) need not be associated only with a measure of the distance from $\overline{y}$ to $Y(x)$. The vector can include (or be equal to) an error in the calculation of $\nabla f(x)$; the result (c) thus provides conditions for obtaining descent when the value of $\nabla f(x)$ is observed in the presence of noise. See [Pol87, Chap. 4] for discussions on such errors and their effect on gradient methods.

The inequality (12) in (d) generalizes the property of *gradient relatedness* of search directions used in unconstrained differentiable optimization [OrR70, Def. 14.3.1]. ∎

**3.2. Step length rules.** This subsection presents and validates the step length rules considered in the realization of Step 3 of the CA algorithm. Four different step length rules are used; the first is a conceptual one.

DEFINITION 3.3 (exact line search). *Let $x \in \mathrm{dom}\, u$, $y \in Y(x)$, and $d := y - x$.*

(E) *Choose $\ell$ as a solution to* $\min\{\, T(x + \ell d) \mid \ell \geq 0 \,\}$.

We next define a new step length rule, which generalizes the Armijo [Arm66] rule from differentiable optimization. The original statement of the Armijo rule cannot be applied here due to the nondifferentiability of $T$; the fact that the generalized gradient cannot be Lipschitz continuous makes it necessary to replace the directional derivative used in the original Armijo rule with a monotonicity measure in $\Phi$. Also, the unit step is the largest for which the modified Armijo rule can be validated (see Lemma 3.5 below); we therefore choose a unit first trial step length. Further, $\Phi$ must be chosen strictly monotone.

DEFINITION 3.4 (Armijo step length). *Let $\Phi$ be strictly monotone on* $\mathrm{dom}\, u$. *Let $x \in \mathrm{dom}\, u \setminus \Omega$, $Y(x) := \{y\}$, and $d := y - x$. Further, let $\alpha, \beta \in (0,1)$.*

(A′) *Let $\ell := \beta^{\bar{\imath}}$, where $\bar{\imath}$ is the smallest nonnegative integer $i$ such that*

$$T(x + \beta^i d) - T(x) \leq \alpha \beta^i [\Phi(x) - \Phi(y)]^{\mathrm{T}} d. \tag{15}$$

To see that this step length rule generalizes the Armijo rule, let $u \equiv 0$ in [NDP]. Then a subproblem solution $y$ is characterized by the system of nonlinear equations $\nabla f(x) = \Phi(x) - \Phi(y)$ (cf. (6)); using this relationship in (15) yields that Rule A′ reduces to the original Armijo rule, with a unit first trial step. (Note that Rule A′ uses an over-estimating model of $T$, since $T'(x;d) \leq [\Phi(x) - \Phi(y)]^{\mathrm{T}} d$ holds.)

The Armijo Rule A′ is validated below.

LEMMA 3.5 (validity of the Armijo rule). *Let $\Phi$ be strictly monotone on* dom $u$. *Let $x \in \operatorname{dom} u \setminus \Omega$, $Y(x) := \{y\}$, and $d := y - x$. Further, let $\alpha, \beta \in (0,1)$.*

(a) *There exists a nonnegative integer $\bar{\imath}$ satisfying* (15).

(b) *Assume further that $\nabla f$ is Lipschitz continuous on* dom $u$. *Then, either a unit step satisfies* (15) *or*

$$\beta^{\bar{\imath}} > 2\beta(\alpha - 1)\frac{[\Phi(x) - \Phi(y)]^{\mathrm{T}} d}{M_{\nabla f}\|d\|^2}. \tag{16}$$

*Proof.*

(a) Taylor's formula yields, for $\ell \in [0,1]$,

$$f(x + \ell d) - f(x) = \ell \nabla f(x)^{\mathrm{T}} d + o(\ell). \tag{17}$$

Further, for $\ell \in [0,1]$ (cf. [Roc70a, Cor. 24.2.1 and Thm. 24.1], [Roc66]),

$$\begin{aligned} u(x + \ell d) - u(x) &= \int_0^\ell u'(x + sd; d)\, ds \leq -\ell u'(x + d; -d) \\ &= \ell \inf\{\, \xi_u(y)^{\mathrm{T}} d \mid \xi_u(y) \in \partial u(y) \,\} \\ &\leq \ell \xi_u(y)^{\mathrm{T}} d \qquad \forall \xi_u(y) \in \partial u(y). \end{aligned} \tag{18}$$

In (18), let $\xi_u(y) \in \partial u(y)$ be given by (6). Then, combining (17) and (18),

$$T(x + \ell d) - T(x) \leq \ell [\Phi(x) - \Phi(y)]^{\mathrm{T}} d + o(\ell).$$

For any value of $\alpha \in (0,1)$, there must therefore be an $\bar{\ell} > 0$ such that, for all $\ell \in (0, \bar{\ell})$, $T(x + \ell d) - T(x) \leq \alpha \ell [\Phi(x) - \Phi(y)]^{\mathrm{T}} d$ holds, from which we get the result.

(b) We begin by showing that for $\ell \in [0,1]$,

$$T(x + \ell d) - T(x) \leq \ell [\Phi(x) - \Phi(y)]^{\mathrm{T}} d + \frac{M_{\nabla f}}{2} \ell^2 \|d\|^2 \tag{19}$$

holds. Taylor's formula and the Lipschitz continuity of $\nabla f$ yields, for $\ell \in [0,1]$,

$$\begin{aligned} f(x + \ell d) - f(x) &= \int_0^\ell \nabla f(x + sd)^{\mathrm{T}} d\, ds \leq \ell \nabla f(x)^{\mathrm{T}} d + \int_0^\ell M_{\nabla f} \|d\|^2 s\, ds \\ &= \ell \nabla f(x)^{\mathrm{T}} d + \frac{M_{\nabla f}}{2} \ell^2 \|d\|^2. \end{aligned} \tag{20}$$

Combining (20) with (18), the latter obtained as in (a) from [$\mathrm{GE}_{\Phi^t}$], we obtain (19). Using (19) and the strict monotonicity of $\Phi$, by choosing $\ell \leq 2(\alpha - 1)([\Phi(x) - \Phi(y)]^{\mathrm{T}} d)/(M_{\nabla f}\|d\|^2)$, we obtain

$$T(x + \ell d) - T(x) \leq \alpha \ell [\Phi(x) - \Phi(y)]^{\mathrm{T}} d. \tag{21}$$

Replacing $\ell$ with $\beta^{\bar{\imath}}$, we see that $\bar{\imath}$ is the smallest integer $i$ to satisfy

$$\beta^i \leq 2(\alpha-1)\frac{[\Phi(x)-\Phi(y)]^{\mathrm{T}}d}{M_{\nabla f}\|d\|^2}. \tag{22}$$

From (22) then follows the desired result. □

*Remark* 3.4. From (22) we may find conditions under which the unit step, that is, $\bar{\imath}=0$, is accepted by Rule A′. Assume that $\Phi$ is strongly monotone on dom $u$. Then, a unit step is obtained if

$$\frac{2(1-\alpha)m_\Phi}{M_{\nabla f}} \geq 1. \tag{23}$$

Note that, from (16), the step length given by Rule A′ always satisfies $\ell \geq \min\{1, 2\beta(1-\alpha)([\Phi(y)-\Phi(x)]^{\mathrm{T}}d)/(M_{\nabla f}\|d\|^2)\}$, which reduces to

$$\ell \geq \min\left\{1, \frac{2\beta(1-\alpha)m_\Phi}{M_{\nabla f}}\right\} \tag{24}$$

whenever $\Phi$ is strongly monotone. □

If the calculation of $T$ is expensive, then a step length formula which does not require its calculation is of interest. Such a rule is described next.

DEFINITION 3.6 (relaxation step length). *Assume that $\nabla f$ is Lipschitz continuous on* dom $u$*, and let $\Phi$ be strongly monotone on* dom $u$*. Further, let $0<\varepsilon_i<m_\Phi/M_{\nabla f}$, $i=1,2$.*

(R) *Choose $\ell$ in the closed interval $(\varepsilon_1, \min\{1, 2m_\Phi/M_{\nabla f}-\varepsilon_2\})$.*

*Remark* 3.5. The inverse of the quotient $m_\Phi/M_{\nabla f}$ appearing in the description of the maximal step length is in some cases an approximate condition number of the Hessian of $f$; consider, for example, Newton's method. □

The step length rule is validated below.

LEMMA 3.7 (validity of the relaxation step). *Assume that $\nabla f$ is Lipschitz continuous on* dom $u$*, and let $\Phi$ be strongly monotone on* dom $u$*. Let $x \in \text{dom } u \setminus \Omega$, $Y(x):=\{y\}$, and $d:=y-x$. If $\ell$ is chosen according to Rule R, then $T(x+\ell d)<T(x)$.*

*Proof.* The inequality (19), together with the strong monotonicity of $\Phi$, yields, for all $\ell \in [0,1]$, that

$$T(x+\ell d)-T(x) \leq \ell\left(-m_\Phi + \frac{M_{\nabla f}}{2}\ell\right)\|d\|^2, \tag{25}$$

from which the result easily follows. □

Note that if

$$\frac{2m_\Phi}{M_{\nabla f}} > 1 \tag{26}$$

holds, then a unit step yields descent. (This condition is implied by (23).)

The relaxation step length has the advantage of being very simple, but it requires the knowledge of (an estimate of) the Lipschitz constant of $\nabla f$. (In certain cases, such as when $f$ is quadratic or separable, estimates of the Lipschitz constant $M_{\nabla f}$ are available; see, for example, [WoZ96]. The maximal step length is usually determined through experiments.) The last step length rule introduced yields descent *eventually* and does not require the knowledge (or estimation) of any problem or method parameters.

DEFINITION 3.8 (divergent series step length).
(D) *Let the sequence* $\{\ell_t\}$ *satisfy*

$$[0,1] \supset \{\ell_t\} \to 0, \qquad \sum_{t=0}^{\infty} \ell_t = \infty. \tag{27}$$

*Remark* 3.6. The requirements of the relaxation and divergent series step length rules need not be viewed as defining step length formulas per se, but can instead be viewed as conditions that some arbitrary step length rule must satisfy in order to yield convergence. The *almost complete relaxation* strategy described in [DeV85, Sec. 3.4], for example, allows for the utilization of an arbitrary step length selection rule through the (almost never used) upwards or downwards rounding of a tentative step length in order to fall between two step lengths (very small and very large, respectively) defined by preselected formulas both satisfying (27). □

## 4. Convergence results.

**4.1. The conceptual method.** In our first convergence result, we will assume that the mappings $\Phi^t$ are *monotone* only. In order to ensure the convergence of the CA algorithm, this will require that the sequence $\{\Phi^t\}$ is constructed by a given function $\varphi$ on dom $u \times$ dom $u$ (that is, that $\Phi^t \equiv \nabla\varphi(\cdot, x^t)$ for each $t$), and that the line search is performed exactly.

THEOREM 4.1 (convergence under Rule E). *Assume that* $u$ *is continuous on* dom $u$. *Let the sequence* $\{\Phi^t\}$ *of cost approximating mappings be constructed such that, for each* $t$, $\Phi^t \equiv \nabla\varphi(\cdot, x^t)$ *for a given function* $\varphi$ : dom $u \times$ dom $u \mapsto \Re$ *of the form* $\varphi(y, x)$, *continuous on* dom $u \times$ dom $u$, *and convex and in* $C^1$ *on* dom $u$ *with respect to* $y$. *Assume that* $x^0 \in$ dom $u$ *is such that the lower level set* $L(x^0) := \{\, x \in \text{dom } u \mid T(x) \le T(x^0) \,\}$ *is bounded, and further that the problem* [NDP$_\varphi$] *is well defined in the sense that* (9) *holds for every* $x \in L(x^0)$. *Let Rule E be used. Then,* $\{T(x^t)\} \to T(\overline{x})$ *for some* $\overline{x} \in \Omega$, *any accumulation point of the sequence* $\{x^t\}$ *(at least one such point exists) lies in* $\Omega$, *and*

$$\left\{ \inf_{x \in \Omega \cap L(x^0)} \|x^t - x\| \right\} \to 0. \tag{28}$$

*Proof.* In order to apply Zangwill's Theorem A [Zan69, Sec. 4.5], we first identify the *solution set* with $\Omega$, the *descent function* with $T$, and the *algorithmic map* with the composite mapping $A := ED$, where $D$ is the direction finding mapping and $E$ is the exact line search map. To fulfill the assumptions of Theorem A, we next show that (1) the sequence $\{x^t\}$ lies in a compact set, (2) $T(x^{t+1}) < T(x^t)$ if $x^t \notin \Omega$ and the algorithm terminates if $x^t \in \Omega$, and (3) the mapping A is closed at all points in $L(x^0) \setminus \Omega$.

(1) From the boundedness assumption and [StW70, p. 134], the set $L(x^0)$ is compact; the descent property then ensures that the sequence $\{x^t\}$ lies in a compact set.
(2) The descent property follows from Lemma 3.2(b). The termination property ($x^t \in \Omega$) follows from Lemma 3.1(a)(2).
(3) The mapping $D$ is closed on $L(x^0)$ by Lemma 3.1(d) (identifying $S$ with $L(x^0)$), and the same is true for the mapping $E$, since $T$ is continuous on dom $u$ [Zan69, Lem. 5.1]. The mapping $\partial([\varphi(\cdot, x) + u]^\circ)$ is monotone, and therefore, by (9) and [Zei90, Prop. 32.33], locally bounded at $\nabla\varphi(x, x) -$

$\nabla f(x)$. From (7), the mapping $Y$ is therefore locally bounded on $L(x^0)$. From (1), the sequence $\{x^t\}$ is bounded. The local boundedness of $Y$ implies that the sequence $\{y^t\}$ is also bounded. This sequence then has an accumulation point. Then, [Zan69, Lem. 4.2] ensures that the composite map $A$ is closed on $L(x^0)$.

The assumptions of Theorem A are thus fulfilled, and we conclude that $\{T(x^t)\}$ converges to $T(\overline{x})$ for some $\overline{x} \in \Omega$, and any accumulation point of $\{x^t\}$ lies in $\Omega$. (The existence of at least one such point follows from the boundedness of $\{x^t\}$.) The last statement follows from [OrR70, Thm. 14.1.4]. □

Note that it is essential that $u$ is continuous on $L(x^0)$, since otherwise the line search mapping $E$ need not be closed. (In the special case where $u \equiv \delta_X$, the line search is made with respect to $f$, and the continuity assumption on $u$ may be removed.)

**4.2. A truncated algorithm.** In this section we provide one example of the many possibilities for constructing realizations of *truncated* CA algorithms. The basis for the algorithm described and validated here is Lemma 3.2(b). (Another truncated CA algorithm, based on Lemma 3.2(c), is presented in section 4.3.) The idea behind a truncated CA method is to reduce the work performed on [NDP$_\varphi$] by limiting the number of iterations performed when solving it using a descent algorithm. This strategy introduces a trade-off between the computational effort spent on solving the subproblem and the quality of the search direction obtained. (Examples of such methods are extensions of truncated Newton methods for systems of nonlinear equations, in which a limited number of steps of an iterative scheme is performed on the linear systems.)

The following assumption is made on the algorithm used for solving [NDP$_\varphi$].

*Assumption* 4.1 (properties of a truncated algorithm). Given an $x \in \operatorname{dom} u$, the problem [NDP$_\varphi$] is solved, starting from $x$, with an iterative algorithm where one iteration may be described by a mapping $G : \operatorname{dom} u \times \operatorname{dom} u \mapsto 2^{\operatorname{dom} u} \times \operatorname{dom} u$ having the following properties.

(1) *Fixed point.* $z \in Y(x) \Longleftrightarrow (z, x) \in G(z, x)$, in which case the algorithm terminates.
(2) *Descent.* If $x \notin \Omega$ and $z \notin Y(x)$, then $T_\varphi(y, x) < T_\varphi(z, x)$ for all $y$ with $(y, x) \in G(z, x)$.
(3) *Closedness.* The mapping $G$ is upper semicontinuous, closed, and compact valued on $\operatorname{dom} u \times L(x^0)$.
(4) *Finite termination.* The number of iterations performed is bounded from below by 1 and from above by a positive integer $\overline{k}$.

*Remark* 4.1. The condition (1) ensures that the fixed point property of the original algorithm is preserved and that the algorithm terminates whenever a solution to [NDP$_\varphi$] has been obtained. The condition (2), together with the lower bound in (4) on the number of iterations performed and Lemma 3.2(b), ensures that the resulting approximate solution $\overline{y}$ defines a direction of descent with respect to $T$. Condition (3) ensures the closedness of the mapping describing the overall algorithm. Whenever $\operatorname{dom} u$ is bounded, it can be replaced by the assumption that the mapping $G$ is closed on $\operatorname{dom} u \times \operatorname{dom} u$. Condition (4), finally, limits the work done in each main iteration of the CA algorithm, and, together with (3), it ensures that any sequence $\{\overline{y}\}$ of approximate solutions generated over a compact set is bounded.

If $\operatorname{dom} u$ is bounded, then the mapping $G$ describing *any* truncated CA algorithm based on a function $\hat{\varphi} : \operatorname{dom} u \times \operatorname{dom} u \mapsto \Re$ of the form $\hat{\varphi}(y, x)$, together with *any* closed updating step which guarantees descent (such as Rule E or a fixed relaxation step), satisfies (1)–(3). □

The truncated CA algorithm is obtained from replacing Step 1 of the CA algorithm (see Table 1) by the following.

1′. *Search direction generation.* Apply $1 \le k_t \le \overline{k}$ iterations of the algorithm described by $G$, starting from $x^t$; that is, let

$$(\overline{y}^t, x^t) \in \underbrace{G \circ \cdots \circ G}_{k_t \text{ times } G}(x^t, x^t).$$

The resulting search direction is $d^t := \overline{y}^t - x^t$.

Note that the value of $k_t$ need not be determined a priori and could instead be regarded as a product of the algorithm and the termination criteria chosen for [NDP$_\varphi$].

THEOREM 4.2 (convergence of a truncated algorithm). *Replace Step* 1 *by Step* 1′ *in the CA algorithm. In Theorem* 4.1, *replace the assumption that* (9) *holds for every* $x \in L(x^0)$ *with Assumption* 4.1. *Then, the conclusions of Theorem* 4.1 *hold for the truncated CA algorithm.*

*Proof.* The proof utilizes Zangwill's theorem and the Spacer Step theorem [Lue84, p. 231]. We begin by identifying the algorithmic mapping $A$. We assume that $\{x^t\}$ is infinite (otherwise, by Assumption 4.1(1), with $z = x$, the algorithm is terminated at a point in $\Omega$). By Assumption 4.1(4), there must be at least one positive integer, say $k$, that occurs an infinite number of times in the sequence $\{k_t\}$. We define

$$\overline{Y}(x) := \{\, \overline{y} \in \Re^n \mid (\overline{y}, x) \in \underbrace{G \circ \cdots \circ G}_{k \text{ times } G}(x, x) \,\}.$$

The algorithmic mapping is $A := E \circ \overline{D}$, where $\overline{D} := \overline{Y} - \cdot$ and $E$ is the exact line search map.

We next establish that this mapping has the properties desired (boundedness, adaption, closedness). We first note that Assumption 4.1(2), together with Lemma 3.2(b), yields that the mapping $\overline{D}$ has the descent property. From the use of Rule E, it then follows that $T(\overline{x}) < T(x)$ for any $\overline{x} \in E \circ \overline{D}(x)$, and therefore $A$ satisfies the second condition (adaption). The boundedness assumption on the set $L(x^0)$ then implies that the sequence $\{x^t\}$ is bounded, thereby satisfying the first condition of Zangwill (boundedness).

We now establish the third condition (closedness). The mapping $G$ is closed on dom $u \times L(x^0)$ (Assumption 4.1(3)). By the upper semicontinuity and compactness of $G$ (Assumption 4.1(3) again), the image $\cup_{x \in L(x^0)} G(x, x)$ of $G$ on the compact set $L(x^0)$ is compact [Nik68, Lem. 4.5]. Lemma 12 of [Mey79] then implies that if $\overline{y}^\infty$ is an accumulation point of a sequence $\{\overline{y}^t\}$ given by $(\overline{y}^{t+1}, x) \in G(\overline{y}^t, x)$, then the set $\{\, y \mid (y, x) \in G \circ G(\overline{y}^\infty, x) \,\}$ contains an accumulation point of $\{\overline{y}^t\}$. We are then in the position of using Lemma 4.2 of [Zan69] to conclude that the composite mapping $G \circ G$ is closed on dom $u \times L(x^0)$. This result used repeatedly then establishes that the mapping $\overline{Y}$, and therefore also $\overline{D}$, is closed on $L(x^0)$. Applying Lemma 4.2 of [Zan69] again, we may conclude that the mapping $E \circ \overline{D}$ is closed on $L(x^0)$. Moreover, the boundedness of $L(x^0)$ and the descent property of $E \circ \overline{D}$ ensures that the range of this mapping over $L(x^0)$ is a compact set. We invoke Corollary 4.2.1 of [Zan69] to conclude that the algorithmic mapping $A$ is closed on $L(x^0)$.

The algorithmic mapping identified by extracting the value $k$ from the choices in the sequence $\{k_t\}$ is clearly of the form

$$C(x) = \{\, y \in \text{dom } u \mid T(y) \le T(x) \,\}, \qquad x \in \text{dom } u.$$

We may therefore invoke the Spacer Step theorem [Lue84, p. 231], which guarantees

that the result holds, thanks to the properties of the mapping $A$. The rest of the proof follows that of Theorem 4.1. □

A natural sufficient condition for the boundedness of $\{\overline{y}^t\}$ is that dom $u$ is bounded; the special case $u \equiv \delta_X$ for a nonempty, compact, and convex set $X$ is covered in [Pat93b].

**4.3. Implementable step length rules.**

THEOREM 4.3 (convergence under Rule A′). *Let* $\Phi : \text{dom } u \times \text{dom } u \mapsto \Re^n$ *be a continuous mapping on* dom $u \times$ dom $u$ *of the form* $\Phi(y, x)$*, maximal and strictly monotone on* dom $u$ *with respect to* $y$*. Assume that the point* $x^0$ *is chosen so that the lower level set* $L(x^0)$ *is bounded, and assume further that the problem* [$\text{GE}_\Phi$] *is well defined in the sense that* (8) *holds for every* $x \in L(x^0)$*. Let Rule A′ be used. Then, any accumulation point of the sequence* $\{x^t\}$ *(at least one such point exists) lies in* $\Omega$*, and* (28) *holds.*

*Proof.* Let $x^\infty$ be any accumulation point of the sequence $\{x^t\}$; the existence of such a point follows from the boundedness of $L(x^0)$ and the descent property. Likewise, the sequence $\{y^t\}$ must be bounded by the well-posedness assumption (8); let $y^\infty$ be an arbitrary accumulation point. Hence, $\{d^t\}$ is bounded.

We will next show that $\{d^t\} \to 0$ must hold; the proof is by contradiction.

By Lemma 3.5(a),

$$T(x^{t+1}) - T(x^t) \leq \alpha \ell_t [\Phi(x^t, x^t) - \Phi(y^t, x^t)]^{\mathrm{T}} d^t < 0, \qquad t = 0, 1, \ldots.$$

If $\{d^t\} \not\to 0$, then as $\{T(x^{t+1}) - T(x^t)\} \to 0$ holds, there must be a subsequence $\mathcal{T}$ such that $\{\ell_t\}_{\mathcal{T}} \to 0$. There must then be an index $\bar{t}$ such that for every $t \geq \bar{t}$ in $\mathcal{T}$, the step length produced by Rule A′ is less than the initial trial step; that is,

$$T(x^t + (\ell_t/\beta)d^t) - T(x^t) > \alpha(\ell_t/\beta)[\Phi(x^t, x^t) - \Phi(y^t, x^t)]^{\mathrm{T}} d^t, \quad t \geq \bar{t}, \quad t \in \mathcal{T}. \tag{29}$$

By the convexity of $u$, we obtain in the limit of $\mathcal{T}$ in the characterization of $y^t$ that $[\Phi(x^\infty, x^\infty) - \Phi(y^\infty, x^\infty)]^{\mathrm{T}} d^\infty \geq [\nabla f(x^\infty) + \xi_u(x^\infty)]^{\mathrm{T}} d^\infty$ holds for any subgradient $\xi_u(x^\infty)$ of $u$ at $x^\infty$. Hence, $[\Phi(x^\infty, x^\infty) - \Phi(y^\infty, x^\infty)]^{\mathrm{T}} d^\infty \geq T'(x^\infty; d^\infty)$ holds. Dividing the inequality (29) by $\ell_t/\beta$, and taking the limit of $\mathcal{T}$, we obtain, taking into account the above inequality, Assumption 1.1, and [Roc81, Prop. 3G], that $T'(x^\infty; d^\infty) \geq 0$ holds.

A contradiction is now reached in the limit of the inequality (cf. Lemma 3.2(a)) $T'(x^t; d^t) \leq [\nabla f(x^t) + \xi_u(y^t)]^{\mathrm{T}} d^t = [\Phi(x^t, x^t) - \Phi(y^t, x^t)]^{\mathrm{T}} d^t < 0$. We conclude that $\{d^t\} \to 0$. Using this result in the characterization of $y^\infty$ then yields that $x^\infty \in \Omega$.

The result (28) follows as in Theorem 4.1. □

In order to introduce simpler step length rules, we assume that the mappings $\Phi^t$ are *strongly* monotone; we then automatically reach a large freedom of choosing their form; for example, they are not necessarily determined by a continuous mapping on dom $u \times$ dom $u$.

THEOREM 4.4 (convergence under strongly monotone mappings $\Phi^t$). *Assume that* $T$ *is bounded from below on* $\Re^n$ *and* $\nabla f$ *is Lipschitz continuous on* dom $u$*. For each* $t$*, let* $\Phi^t$ *be strongly monotone and Lipschitz continuous on* dom $u$*. Assume further that* $m_\Phi := \liminf_{t\to\infty}\{m_{\Phi^t}\} > 0$ *and* $M_\Phi := \limsup_{t\to\infty}\{M_{\Phi^t}\} < +\infty$.

(a) *From any starting point* $x^0 \in$ dom $u$ *and under Rule E, A′, or R, any accumulation point of* $\{x^t\}$ *lies in* $\Omega$.

(b) *If further* $\{x^t\}$ *is bounded and* $\Omega$ *is nonempty, then under Rules E, A′, or R,* (28) *holds. Further, under Rule D, at least one accumulation point of* $\{x^t\}$ *lies in* $\Omega$.

(c) *If further $\Omega$ is finite, then under Rule A′ or R, $\{x^t\}$ converges to one point in $\Omega$. The same conclusion holds for Rule E whenever $\{\ell_t\}$ is bounded above.*

*Proof.*

(a) We begin by showing that for each of the three step length rules, $\{d^t\} \to 0$ must hold. From (25) and the use of Rule E follows that, for $\ell \in [0,1]$,

$$T(x^{t+1}) \le T(x^t + \ell d^t) \le T(x^t) + \ell \left( -m_{\Phi^t} + \frac{M_{\nabla f}}{2} \ell \right) \|d^t\|^2, \qquad t = 0, 1, \ldots .$$

Minimizing the right-hand side of this inequality, we obtain

$$\ell_t^* = \min\{1, m_{\Phi^t}/M_{\nabla f}\}.$$

Hence,

$$T(x^{t+1}) - T(x^t) \le -\rho \|d^t\|^2, \qquad t = 0, 1, \ldots , \tag{30}$$

with $\rho := \min\{m_\Phi^2/(2M_{\nabla f}), M_{\nabla f}/2\}$. The Armijo rule satisfies (30) with $\rho := (2\beta\alpha(1-\alpha)m_\Phi^2)/M_{\nabla f}$, so does the relaxation rule, with $\rho := \varepsilon_1\varepsilon_2 M_{\nabla f}/2$.

Since $T$ is lower bounded and, from the above, $\{T(x^t)\}$ is decreasing, $\{T(x^{t+1}) - T(x^t)\} \to 0$. Thus, by (30), in each of the three rules we obtain that $\{d^t\} \to 0$.

From (6) and the upper bound on $M_{\Phi^t}$, as $\{d^t\} \to 0$,

$$\{\nabla f(x^t) + \xi_u(y^t)\} = \{\Phi^t(x^t) - \Phi^t(y^t)\} \to 0. \tag{31}$$

Assume now that $\{x^t\}$ has an accumulation point, $x^\infty$, corresponding to a convergent subsequence $\{x^t\}_{t\in\mathcal{T}}$. Since $\{y^t - x^t\} = \{d^t\} \to 0$, we have that $\{y^t\}_{t\in\mathcal{T}} \to x^\infty$. From the closedness of $\partial u$ [Roc70a, Thm. 24.4] and the continuity of $\nabla f$, (31) yields that $\{\xi_u(y^t)\}_{t\in\mathcal{T}} \to \xi_u(x^\infty) \in \partial u(x^\infty)$, and, again appealing to (31), $\nabla f(x^\infty) + \xi_u(x^\infty) = 0$. Thus, $x^\infty \in \Omega$.

(b) The result (28) follows from (a) and [OrR70, Thm. 14.1.4]. Since $\{\ell_t\} \to 0$ in Rule D, it holds for all $t \ge \bar{t}$ (cf. (25)) that, for some $\sigma > 0$, $T(x^{t+1}) - T(x^t) \le -\ell_t\sigma\|d^t\|^2$. Hence, eventually $\{T(x^t)\}$ is decreasing, and by the lower boundedness assumption, it converges to a finite value. We therefore must have that $\sigma\sum_{t=\bar{t}}^\infty \ell_t\|d^t\|^2 < \infty$. But $\sum_{t=0}^\infty \ell_t = \infty$ holds, from which it follows that $\liminf_{t\to\infty} \|d^t\| = 0$ must hold. The corresponding accumulation point of $\{x^t\}$ must then, by the construction of $[\mathrm{GE}_\Phi]$, lie in $\Omega$.

(c) The result follows from the equality $\|x^{t+1} - x^t\| = \ell_t\|d^t\|$, the result of (a) that $\{d^t\} \to 0$, the upper bound on $\ell_t$, and [OrR70, Thm. 14.1.5]. □

The proof for Rule A′ shows that convergence is guaranteed for *any* step length rule that results in a larger reduction of $T$ at each iteration than the Armijo rule.

The above result can be improved to require Lipschitz continuity only for Rules R and D [Pat98].

It is to be observed that although the strong monotonicity property of $\Phi$ implies that (8) holds for every $x \in \operatorname{dom} u$, neither $\{x^t\}$ nor $\{y^t\}$ are necessarily bounded. Since $\{T(x^t)\}$ converges and $\{d^t\} \to 0$, however, boundedness of $\{x^t\}$ and $\{y^t\}$ is ensured if the lower boundedness of $T$ is replaced by the stronger condition of *weak coercivity* ($\operatorname{dom} T$ is bounded or $\lim_{\|x\|\to\infty}\{T(x)\} = +\infty$).

The maximal allowed step length in Rule R is bounded by the constant $2m_{\Phi^t}/M_{\nabla f}$. Now, suppose that $f$ may be written as $f := f_1 + f_2$, where $f_1$ and $f_2$ are convex with Lipschitz continuous gradients, and that $f_1$ is strongly convex. For the sake

of this example, we further assume that $f_1$ is quadratic. To enhance the speed of convergence of the CA algorithm, we consider redefining $\varphi^t$ as $\varphi^t := \varphi^t + f_1$. (This operation corresponds to applying a cost approximation to the function $[u + f_1] + f_2$, instead of to $u + f$.) Indeed, this operation yields an increased maximal allowed step length since

$$\frac{2m_{(\varphi^t+f_1)}}{M_{\nabla f_2}} = \frac{2(m_{\varphi^t} + m_{f_1})}{M_{\nabla(f-f_1)}} > \frac{2m_{\varphi^t}}{M_{\nabla f}}.$$

It also yields a faster convergence for the other step length rules; for Rule A′, observe the change in the linear convergence ratio in Theorem 4.6 below.

From this example, we are lead to conclude that, in order to achieve the best possible rate of convergence, the function $\varphi^t := \varphi^t + f_1$ should be constructed such that $f_2 := f - f_1$ is *not* strongly convex. (The desire to obtain a high rate of convergence must of course be weighed against the computational difficulty of the subproblems.)

Chen and Rockafellar [ChR92] study splitting methods for the problem of finding a zero of the sum of two maximal monotone operators. They argue that all the strong monotonicity inherent in the problem mapping should be kept in the mapping defining the *forward step*; their result corresponds to the above for this special case of CA method.

We conclude this section by combining inexact solutions of $[\mathrm{GE}_{\Phi^t}]$ and inexact line searches into an implementable algorithm.

THEOREM 4.5 (an implementable algorithm). *Let the assumptions of Theorem* 4.4 *hold. In the CA algorithm, let Step* 1 *be performed inexactly, in the sense that, for each* $t$*, the approximate solution* $\overline{y}^t$ *satisfies* (10)*, with* $\|r^t\| \le \omega\|\overline{d}^t\|$ *for some* $\omega < m_\Phi$*. Then, all the conclusions of Theorem* 4.4 *hold, with the exception that in step length Rule R, the constant* $m_\Phi$ *should be replaced by* $m_\Phi - \omega$*.*

The proof of this result is very similar to that of Theorem 4.4; its basis is a modification of the inequality (25) resulting from (10), in which $m_\Phi$ is replaced by $m_\Phi - \omega$.

A slightly weaker result can be established wherein the condition $\|r^t\| \le \omega\|\overline{d}^t\|$ is replaced by the requirement that $\{\|r^t\|\} \to 0$ holds; the argument is, roughly, that if $\|r^t\| \ge m_\Phi\|\overline{d}^t\|$ holds for infinitely many $t$, then $\{\|\overline{d}^t\|\} \to 0$ must hold; if not, then Lemma 3.2(c) establishes that $\overline{d}^t$ eventually is a direction of descent.

**4.4. Linear convergence.** For affine mappings $\Phi$, it is possible to show that the CA algorithm converges *Q-linearly* to an element $x^*$ of $\Omega$, that is, that

$$\limsup_{t\to\infty} \frac{\|x^{t+1} - x^*\|}{\|x^t - x^*\|} =: q < 1. \tag{32}$$

Convergence rests on a nonexpansiveness result for a part of the mapping $x \mapsto Y(x)$.

Let $\Phi(x) := Bx$ for some positive definite matrix $B \in \Re^{n\times n}$. Then, from the characterization $[\mathrm{GE}_{\Phi^t}]$ of the set $Y(x)$, for any $x \in \operatorname{dom} u$,

$$y(x) = [B + \partial u]^{-1}[B - \nabla f](x). \tag{33}$$

We then rewrite (33) as $y(x) := VW(x)$, where

$$V := \left[\frac{1}{m_\Phi}(B + \partial u)\right]^{-1}, \qquad W := \left[\frac{1}{m_\Phi}(B - \nabla f)\right],$$

and $m_\Phi$ is the modulus of strong monotonicity of $\Phi$.

This rewriting of the characterization of $y(x)$ is done because the operator $V$ can be shown to be *nonexpansive* on dom $u$; that is,

$$\|V(x) - V(y)\| \leq \|x - y\|, \qquad \forall x, y \in \text{dom } u.$$

Indeed, from the monotonicity of $\partial u$ and the positive definiteness of $B$ we have, for all $x, y \in \text{dom } u$, and $\xi_u(x) \in \partial u(x)$, $\xi_u(y) \in \partial u(y)$,

$$[\xi_u(x) - \xi_u(y)]^{\mathrm{T}}(x - y) \geq 0,$$
$$(x - y)^{\mathrm{T}} B(x - y) \geq m_\Phi \|x - y\|^2.$$

Dividing both the above inequalities by $m_\Phi$ and adding them, we obtain

$$\left(\left[\frac{1}{m_\Phi}(B + \xi_u)\right](x) - \left[\frac{1}{m_\Phi}(B + \xi_u)\right](y)\right)^{\mathrm{T}} (x - y) \geq \|x - y\|^2.$$

But this inequality yields that the operator $V$ is firmly nonexpansive and therefore nonexpansive [Roc76, EcB92].

The importance of this result is that the convergence analysis is simplified since, using Lemma 3.1(a)(1), we immediately have that for any $x^* \in \Omega$,

$$\|y(x) - x^*\| = \|VW(x) - VW(x^*)\| \leq \|W(x) - W(x^*)\|,$$

and it suffices to study the operator norm of $W$ in the vicinity of $x^*$.

THEOREM 4.6 (linear convergence). *Assume that $\nabla f$ is Lipschitz continuous on* dom $u$. *Let $\Phi(x) := Bx$, where $B \in \Re^{n \times n}$ is positive definite. Assume that, using Rule A′ or Rule R, $\{x^t\} \to x^* \in \Omega$, where $f$ is twice continuously differentiable and $\nabla^2 f(x^*)$ is positive definite. If*

$$\sigma := \frac{M_\Phi}{m_\Phi} \|I - B^{-1} \nabla^2 f(x^*)\| < 1, \tag{34}$$

*then $\{x^t\}$ converges Q-linearly with the ratio*

$$q \leq \max\left\{\sigma, 1 - \tfrac{2\beta(1-\alpha)m_\Phi}{M_{\nabla f}}(1 - \sigma)\right\}, \quad \textit{for Rule A′, and}$$

$$q \leq \max\{\sigma, 1 - \underline{\ell}(1 - \sigma)\} \qquad \textit{for Rule R,}$$

*where $\underline{\ell} := \inf_t\{\ell_t\} \geq \varepsilon_1$.*

*Proof.* From Lemma 3.1(a)(1) and the nonexpansiveness property of $V$, we have $\|y^t - x^*\| = \|VW(x^t) - VW(x^*)\| \leq \|W(x^t) - W(x^*) - \nabla W(x^*)(x^t - x^*)\| + \|\nabla W(x^*)(x^t - x^*)\|$, where $\nabla W(x^*) := (1/m_\Phi)(B - \nabla^2 f(x^*))$.

Let $\varepsilon > 0$ be arbitrary. Since $\{x^t\} \to x^*$, we have for a sufficiently large $t$ that $\|W(x^t) - W(x^*) - \nabla W(x^*)(x^t - x^*)\| \leq \varepsilon \|x^t - x^*\|$. We also have that $\|\nabla W(x^*)\| \leq (M_\Phi/m_\Phi)\|I - B^{-1}\nabla^2 f(x^*)\| =: \sigma$, so that $\|y^t - x^*\| \leq (\sigma + \varepsilon)\|x^t - x^*\|$. Then, $\|x^{t+1} - x^*\| = \|x^t + \ell_t d^t - x^*\| \leq \ell_t \|y^t - x^*\| + (1 - \ell_t)\|x^t - x^*\| \leq [1 - \ell_t(1 - \sigma - \varepsilon)]\|x^t - x^*\|$.

For Rule A′, from (24), we then have, for a large enough $t$,

$$\|x^{t+1} - x^*\| \leq q_\varepsilon \|x^t - x^*\|, \tag{35}$$

where $q_\varepsilon := \max\{\sigma + \varepsilon, 1 - 2\beta(1 - \alpha)m_\Phi / M_{\nabla f}(1 - \sigma - \varepsilon)\}$, while, for Rule R, (35) holds for large enough $t$ with $q_\varepsilon := \max\{\sigma + \varepsilon, 1 - \underline{\ell}(1 - \sigma - \varepsilon)\}$. Since $\varepsilon$ was arbitrary we have asymptotically that

$$q := \limsup_{t \to \infty} \frac{\|x^{t+1} - x^*\|}{\|x^t - x^*\|} \leq q_0,$$

where $q_0 := \max\{\sigma, 1 - 2\beta(1-\alpha)m_\varphi/M_{\nabla f}(1-\sigma)\}$ in the case of Rule A′, and $q_0 := \max\{\sigma, 1 - \underline{\ell}(1-\sigma)\}$ for Rule R. If $\sigma < 1$, then $q_0 < 1$, and the theorem is proved. □

If (26) holds, then $\ell_t \equiv 1$ is a valid step length in Rule R, and $q \leq \sigma$ holds. Hence, if $B := \nabla^2 f(x^*)$, then *superlinear* convergence is obtained; that is, $q = 0$ in (32). For Rule A′, the same conclusion may be drawn, provided that (23) holds.

The result of Theorem 4.4 suggests a means to choosing a matrix $B$ to obtain linear convergence in practice; choosing $\Phi^t$ of the form $x \mapsto B^t x$ for some positive definite matrix $B^t$ which, at least asymptotically, approaches the Hessian $\nabla^2 f(x^t)$, eventually the requirement (34) will be fulfilled. Any matrix $B^t$ generated after this occurrence will be an appropriate choice for the fixed matrix $B$ in the above theorem. Further, the longer this choice is postponed, the closer the value of $\sigma$ will be to zero; therefore, a convergence rate close to superlinear may be obtained. (In fact, it is possible to extend the above result to incorporate general iteration-dependent mappings $\Phi^t$ [Pat98].)

We have not been able to establish linear convergence for Rule E or D, since no lower bound on the respective step length is available.

**5. Applications of the convergence analysis.** We illustrate in this section further examples of the generality of the framework and demonstrate the strength of the convergence analysis. We demonstrate that the results strengthen those previously obtained for special cases of CA methods and introduce a flexibility in the realizations of existing methods under the same convergence conditions.

**5.1. On the linear convergence of CA algorithms.** Despite the fact that the linear convergence Theorem 4.6 is here established only for iteration independent and affine cost approximation mappings, it is strong enough to reproduce some well-known linear convergence results in differentiable optimization, as we illustrate below.

COROLLARY 5.1 (linear convergence). *Let the assumptions of Theorem* 4.6 *hold, and let* $B := (1/\gamma)I$, $\gamma > 0$.

(a) *Let Rule R be applied, using unit steps. If the method converges, then it converges with a linear rate. Further, the convergence ratio is*

$$q \leq \max\{|1 - \gamma m|, |1 - \gamma M|\}, \tag{36}$$

*where* $m$ *and* $M$ *are the smallest and largest eigenvalues of* $\nabla^2 f(x^*)$*, respectively.*

(b) *Let further* $u \equiv \delta_X$*, where* $X$ *is a nonempty, closed, and convex set in* $\Re^n$*. Then, the algorithm is equivalent to the Goldstein–Levitin–Polyak gradient projection algorithm* [Gol64, LeP66]*. Moreover, the convergence criterion* (26) *and the linear convergence ratio* (36) *coincide with those of this algorithm.*

(c) *Let further* $X := \Re^n$*. Then, the algorithm is equivalent to the steepest descent method with fixed step length* $\gamma$ [Pol63]*, with the corresponding convergence criterion and linear convergence ratio.*

*Proof.*

(a) We first note that $m_\Phi = M_\Phi = 1/\gamma$. The condition (26) for convergence using unit steps then reduces to $\gamma < 2/M_{\nabla f}$, and since $\|\nabla^2 f(x^*)\| \leq M_{\nabla f}$ and for any symmetric matrix $A$, $\|I - \gamma A\| < 1$ if $\gamma < 2/\|A\|$, (34) follows.

With $\ell := 1$ we have that $q \leq \sigma := \|I - \gamma\nabla^2 f(x^*)\| = \max\{|1-m|, |1-M|\}$, where $m$ and $M$ are, respectively, the smallest and the largest eigenvalues of $\gamma\nabla^2 f(x^*)$.

(b) Straightforward calculations yield $y^t = P_X(x^t - \gamma \nabla f(x^t))$, where $P_X(z)$ denotes the Euclidean projection of $z$ onto $X$. Since $\ell_t \equiv 1$, $x^{t+1} = y^t$ follows, and we obtain the Goldstein–Levitin–Polyak gradient projection algorithm. The convergence criterion ($\gamma < 2/M_{\nabla f}$) and linear convergence ratio (36) can be found in [LeP66].

(c) Follows from (b), with $X = \Re^n$. □

**5.2. Improvements of existing results.** We provide a few examples of previous convergence analyses of special cases of CA algorithms and demonstrate that the analysis in this paper improves upon them.

**Application 1.** In the CA algorithm, choose the following: $\varphi^t := 0$; each subproblem [NDP$_\varphi$] is solved exactly; and the conceptual step length Rule E is used. We then obtain the method of Mine and Fukushima [MiF81]. Comparing their convergence results to that of Theorem 4.1, their convergence conditions include an unnecessary strict convexity assumption on $u$; further, they do not recognize that $u$ needs to be continuous. Needless to say, our convergence analysis also allows for a larger flexibility in the realization of this method under the same convergence conditions; for example, using inexact solutions of [NDP$_\varphi$] is optional (see Theorem 4.2).

When $u \equiv \delta_X$ for a closed, convex set $X$ in $\Re^n$, this algorithm reduces to the Frank–Wolfe method. For this algorithm, the validation of truncated subproblems (Theorem 4.2) is new, and its successful use in several applications (where it is often referred to as an heuristic procedure; see, e.g., [LHB85]) is thus supported in theory. Further, in this special case of [NDP], the original Armijo step length rule can be used in place of Rule A$'$, and convergence is established without the need for $\Phi$ to be strictly monotone (see [Pat93c, Pat98]). As far as the author is aware, previous convergence results for Armijo step length rules within the Frank–Wolfe algorithm have always assumed that $\nabla f$ is Lipschitz continuous (e.g., [PsD78, Chap. III.3]), which, however, is unnecessary (cf. Theorem 4.3).

**Application 2.** In the CA algorithm, choose the following: $\alpha > 0$; $\varphi^t(x) := 1/(2\gamma_t)\|x\|^2$, where $\{\gamma_t\}$ is chosen so that $\inf_t\{\gamma_t\} > 0$ and $\sup_t\{\gamma_t\} < 1/\alpha$ holds; each subproblem [NDP$_{\varphi^t}$] is solved exactly; and the Armijo step length rule A$'$ is used. We then obtain the method of Fukushima and Mine [FuM81]. The condition that $\sup_t\{\gamma_t\} < 1/\alpha$ holds is unnecessary in the presence of the standard assumption that the acceptance parameter satisfies $\alpha \in (0, 1)$, as evidenced by Lemma 3.5 and Theorem 4.4. Further, the enforced dependency of $\alpha$ on the possible choices of $\gamma_t$ in their method has detrimental effects both in terms of allowing less flexibility in the realization of the method, and in terms of the convergence rate, which becomes arbitrarily poor if $\gamma_t$ is chosen close to $1/\alpha$. Optional realizations of this algorithm include the use of inexact solutions of [NDP$_{\varphi^t}$] as well as the use of the step length Rules R and D (cf. Theorems 4.4 and 4.5). Moreover, the larger flexibility with which the function $\varphi$ can be chosen introduces possibilities to improve the convergence rate of the method, as the next application establishes.

**Application 3.** Assume that $f \equiv 0$; replacing Rule A$'$ with Rule R using unit steps reduces the above method to the *proximal point algorithm* [Mar70, Roc76]. If $u$ is also differentiable, then by the results of Rockafellar [Roc76] the proximal point algorithm can be made superlinearly convergent by letting $\{\gamma_t\} \to +\infty$. Superlinear convergence is not possible to obtain in the method of [FuM81] in the nondifferentiable case, since letting $\{\gamma_t\} \to +\infty$ would destroy the validity of their step length rule. The discussion following Theorem 4.6, however, demonstrates that also in the

nondifferentiable case, a convergence rate as close to superlinear as desired can be obtained in the CA algorithm by introducing second-order information from $f$ into $\Phi$; we conjecture that the method which is obtained if the matrix $B^t$ is never fixed is actually superlinearly convergent. We note that this result is consistent with the results for gradient related methods in unconstrained differentiable optimization, where superlinear convergence is characterized by the convergence of the search directions toward Newton directions.

**Application 4.** In the CA algorithm, choose the following: $\varphi^t$ is strongly convex with a Lipschitz continuous gradient; and Rule R is used. We then obtain the *auxiliary problem principle* of Cohen [Coh78, Coh80] (see, e.g., [Pat93c, LaP94, Pat98] for detailed comparisons). The convergence results of this scheme correspond essentially to that of Theorem 4.4(b) for this special choice of CA algorithm. There are several optional realizations of the algorithm with the same convergence conditions, including the use of step length Rule A$'$, the use of mappings $\Phi^t$ that are not gradients, and the use of truncated subproblem solutions.

In the case where $u \equiv \delta_X$ for a closed, convex set in $\Re^n$, we may identify several other algorithm frameworks that are included in that defined by the CA method, and the convergence analyses are, in all cases, more limited and weaker than ours; examples of such frameworks include those in [DFL86, LaM90, Tse91, Mig94, ZhM95]. Applications of the convergence analysis of the CA method to differentiable optimization can be found in [Pat93b, Pat93c, Pat98].

**5.3. Concluding remarks.** The class of CA algorithms is very general; some examples of methods that are included have been mentioned. What we have not mentioned is that the generality present allows one to adapt the algorithm to problem structures (such as choosing $\Phi$ to adapt to a separability in $u$ or $f$); further, the analysis provides convergence results for new methods.

In this paper a convergence analysis was made for the class of CA algorithms when applied to a nondifferentiable optimization program; consequences of the analysis were illustrated, which resulted in several improvements of previous analyses. A convergence analysis is performed for the special case of differentiable, constrained, and unconstrained optimization problems in the author's Ph.D. thesis [Pat93c]; in several instances, in particular concerning the requirements on $\Phi^t$ and the convergence rate, the results reached in this paper are improved upon. These, and new developments, will be reported elsewhere in the future [Pat98].

**Acknowledgments.** The author wishes to acknowledge the many constructive remarks made by the reviewers, the associate editor, and associate professor Torbjörn Larsson.

REFERENCES

[Arm66] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[Ber63] C. Berge, *Topological Spaces*, Oliver & Boyd, Edinburgh, 1963.

[Bre73] H. Brézis, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North–Holland, Amsterdam, 1973.

[Bro68] F. E. Browder, *Nonlinear maximal monotone mappings in Banach spaces*, Math. Ann., 175 (1968), pp. 81–113.

[ChR92] G. H.-G. Chen and R. T. Rockafellar, *Forward-Backward Splitting Methods in Lagrangian Optimization*, Report, Department of Applied Mathematics, University of Washington, Seattle, WA, 1992.

[Coh78] G. Cohen, *Optimization by decomposition and coordination: A unified approach*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 222–232.

[Coh80] G. Cohen, *Auxiliary problem principle and decomposition of optimization problems*, J. Optim. Theory Appl., 32 (1980), pp. 277–305.

[CoZ84] G. Cohen and D. L. Zhu, *Decomposition coordination methods in large scale optimization problems: The nondifferentiable case and the use of augmented Lagrangians*, in Advances in Large Scale Systems, Vol. 1, J. B. Cruz, ed., JAI Press, Greenwich, CT, 1984, pp. 203–266.

[CPS92] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, San Diego, CA, 1992.

[DeV85] V. F. Dem'yanov and L. V. Vasil'ev, *Nondifferentiable Optimization*, Optimization Software, New York, 1985.

[DFL86] J.-P. Dussault, J. A. Ferland, and B. Lemaire, *Convex quadratic programming with one constraint and bounded variables*, Math. Programming, 36 (1986), pp. 90–104.

[EcB92] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Programming, 55 (1992), pp. 293–318.

[FHN96] M. Fukushima, M. Haddou, V. H. Nguyen, J.-J. Strodiot, T. Sugimoto, and E. Yamakawa, *A parallel descent algorithm for convex programming*, Comput. Optim. Appl., 5 (1996), pp. 5–37.

[FuM81] M. Fukushima and H. Mine, *A generalized proximal point algorithm for certain nonconvex minimization problems*, Internat. J. Systems Sci., 12 (1981), pp. 989–1000.

[Gol64] A. A. Goldstein, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.

[LaM90] T. Larsson and A. Migdalas, *An algorithm for nonlinear programs over Cartesian product sets*, Optimization, 21 (1990), pp. 535–542.

[LaP94] T. Larsson and M. Patriksson, *A class of gap functions for variational inequalities*, Math. Programming, 64 (1994), pp. 53–79.

[LHB85] L. J. LeBlanc, R. V. Helgason, and D. E. Boyce, *Improved efficiency of the Frank–Wolfe algorithm for convex network programs*, Transportation Sci., 19 (1985), pp. 445–462.

[LeP66] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), pp. 1–50.

[Lue84] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed., Addison–Wesley, Reading, MA, 1984.

[Mar70] B. Martinet, *Regularisation d'inéquations variationnelles par approximations successives*, Revue Française d'Informatique et de Recherche Opérationelle, R-3 (1970), pp. 154–158.

[Mey79] G. G. L. Meyer, *Asymptotic properties of sequences iteratively generated by point–to–set maps*, Math. Programming Study, 10 (1979), pp. 115–127.

[Mig94] A. Migdalas, *A regularization of the Frank–Wolfe method and unification of certain nonlinear programming methods*, Math. Programming, 65 (1994), pp. 331–345.

[MiF81] H. Mine and M. Fukushima, *A minimization method for the sum of a convex function and a continuously differentiable function*, J. Optim. Theory Appl., 33 (1981), pp. 9–23.

[Min62] G. J. Minty, *Monotone (nonlinear) operators in Hilbert space*, Duke Math. J., 29 (1962), pp. 341–346.

[Nik68] H. Nikaido, *Convex Structures and Economic Theory*, Academic Press, New York, 1968.

[OrR70] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[Pat93a] M. Patriksson, *Partial linearization methods in nonlinear programming*, J. Optim. Theory Appl., 78 (1993), pp. 227–246.

[Pat93b] M. Patriksson, *A unified description of iterative algorithms for traffic equilibria*, European J. Oper. Res., 71 (1993), pp. 154–176.

[Pat93c] M. Patriksson, *A unified framework of descent algorithms for nonlinear programs and variational inequalities*, Ph.D. dissertation, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1993.

[Pat94a] M. Patriksson, *On the convergence of descent methods for monotone variational inequalities*, Oper. Res. Lett., 16 (1994), pp. 265–269.

[Pat94b] M. Patriksson, *The Traffic Assignment Problem: Models and Methods*, VSP, Utrecht, 1994.

[Pat97] M. Patriksson, *Merit functions and descent algorithms for a class of variational inequality problems*, Optimization, 41 (1997), pp. 37–55.

[Pat98] M. Patriksson, *Nonlinear Programming and Variational Inequalities: A Unified Approach*, Kluwer Academic Publishers, Norwell, MA, to appear.

[Pol63] B. T. Polyak, *Gradient methods for the minimisation of functionals*, U.S.S.R. Comput. Math. and Math. Phys., 3 (1963), pp. 864–878.

[Pol87] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.

[PsD78] B. N. Pshenichny and Yu. M. Danilin, *Numerical Methods in Extremal Problems*, Mir Publishers, Moscow, 1978.

[Rob79] S. M. Robinson, *Generalized equations and their solutions, part* I: *Basic theory*, Math. Programming Study, 10 (1979), pp. 128–141.

[Roc66] R. T. Rockafellar, *Characterization of the subdifferentials of convex functions*, Pacific J. Math., 17 (1966), pp. 497–510.

[Roc69] R. T. Rockafellar, *Local boundedness of nonlinear, monotone operators*, Michigan Math. J., 16 (1969), pp. 397–407.

[Roc70a] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[Roc70b] R. T. Rockafellar, *On the maximal monotonicity of subdifferential mappings*, Pacific J. Math., 33 (1970), pp. 209–216.

[Roc70c] R. T. Rockafellar, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 75–88.

[Roc76] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[Roc81] R. T. Rockafellar, *The Theory of Subgradients and its Applications to Problems of Optimization: Convex and Nonconvex Functions*, Heldermann-Verlag, Berlin, 1981.

[Sho85] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag, Berlin, 1985.

[StW70] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions* I, Springer-Verlag, Berlin, 1970.

[Tse91] P. Tseng, *Decomposition algorithm for convex differentiable minimization*, J. Optim. Theory Appl., 70 (1991), pp. 109–135.

[WoZ96] G. R. Wood and B. P. Zhang, *Estimation of the Lipschitz constant of a function*, J. Global Optim., 8 (1996), pp. 91–103.

[Zan69] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*, Prentice–Hall, Englewood Cliffs, NJ, 1969.

[Zei90] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* II/B: *Nonlinear Monotone Operators*, Springer-Verlag, New York, 1990.

[ZhM95] D. Zhu and P. Marcotte, *Coupling the auxiliary problem principle with descent methods of pseudoconvex programming*, European J. Oper. Res., 83 (1995), pp. 670–685.