Cost approximation algorithms with nonmonotone line searches for a general class of nonlinear programs[*]

Michael Patriksson[†]

June 16, 2011

Abstract. When solving ill-conditioned nonlinear programs by descent algorithms, the descent requirement may induce the step lengths to become very small, thus resulting in very poor performances. Recently, suggestions have been made to circumvent this problem, among which is a class of approaches in which the objective value may be allowed to increase temporarily. Grippo *et al.* [GLL91] introduce nonmonotone line searches in the class of deflected gradient methods in unconstrained differentiable optimization; this technique allows for longer steps (typically of unit length) to be taken, and is successfully applied to some ill-conditioned problems. This paper extends their nonmonotone approach and convergence results to the large class of cost approximation algorithms of Patriksson [Pat93b], and to optimization problems with both convex constraints and nondifferentiable objective functions.

Key Words. Nondifferentiable optimization, cost approximation, nonmonotone algorithms

Abbreviated Title. Cost approximation with nonmonotone line searches

1 Introduction

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

[NDP]

$$\min_{x \in \Re^n} T(x) = f(x) + u(x).$$

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 $\mathtt{dom}\, u$ 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])

[*] This research was sponsored by a grant from the Swedish Research Council for Engineering Sciences (TFR) (282-93-1195)

[†] Department of Mathematics, Linköping University, S-581 83 Linköping, Sweden

[GE]

$$0 \in \nabla f(x^*) + \partial u(x^*),$$

which, under a mild assumption (see Assumption 2.1.1 below), 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 $X \subseteq \Re^n$ ($u(x) = 0$, $x \in X$; $u(x) = +\infty$, $x \notin X$) and f is pseudoconvex.

We let Ω denote the set of solutions to [GE], and assume that it is nonempty.

The descent requirement on $\{T(x^t)\}$ in iterative methods for [NDP] has been observed to sometimes lead to severe efficiency losses. In connection with penalty algorithms, Maratos [Mar78] observes that a descent requirement on the merit function employed may lead to very short steps being taken in the line search. A similar behaviour is observed in applications of Newton-type methods (as with most other descent methods) to ill-conditioned problems, such as when the objective is highly nonlinear with deep, narrow curved alleys—a characteristic of penalized problems—at which bottom a descent algorithm is trapped (e.g., [GLL86]).

The ability of an algorithm to attain long steps is, however, often associated with a high convergence rate; for Newton-type methods, the attainment of unit steps after at most a finite number of iterations characterizes their superlinear convergence rate (e.g., [DeM74, DeM77]). Further, if a line search can be avoided then it will result in substantial time savings when the objective function is difficult to evaluate, as in applications to problems in control theory.

It therefore seems natural to construct *nonmonotone* techniques, that is, methods based on rules which will allow long steps to be taken even if the objective value increases, as long as there is no indication of divergence.

A fundamental work in the development of such schemes is that of Grippo *et al.* [GLL91]. Based on some observations on the behaviour of Newton's method, they are able to state reasonable conditions for the realization of gradient-related methods in unconstrained, differentiable optimization which allows for a nonmonotone sequence of objective function values.

Nonmonotone extensions of several classical algorithms have been devised, many of which are recent; these include methods based on sequential quadratic programming (SQP) ([CPLP82, PaT91, BPTZ92]), Newton-type methods ([GLL86, GLL89, GLL90, PHR91, FeL94, DiF95, FLR96]) and (the more general) deflected gradient algorithms in unconstrained optimization ([GLL91, Toi96]), bundle methods in nondifferentiable optimization ([FaL93]), and trust region algorithms ([DXZ93, Toi94]). Other nonmonotone algorithms, although not constructed with the objective of obtaining long steps, include subgradient optimization algorithms in nondifferentiable optimization (e.g., [Sho85]), and heuristics in combinatorial optimization (e.g., [Ree93]).

Most of these works are concerned with extensions of Newton-type methods. There is however every reason to believe that the introduction of nonmonotone techniques can be beneficial also in areas of application where Newton-like methods may not be possible, or realistic, to apply. (For example, the objective function may not enjoy sufficient differentiability properties, or the problem may have a structure amenable to decomposition techniques which excludes the use of a Newton-like approach. See Section 6 for more discussions on this topic.) The paper presents a general class of algorithms for the solution of [NDP]. It is a combination of two approaches: (1) the search direction finding subproblems are based on the *cost approximation* (CA) framework of Patriksson [Pat93b]; (2) the line search is an extension of the nonmonotone technique of Grippo *et al.* [GLL91] to nondifferentiable and/or constrained optimization problems. The resulting algorithm is a class of CA algorithms in which larger steps are allowed, and is simultaneously an extension of the nonmonotone algorithm in [GLL91] to both more general problems and more general search direction finding schemes and line search techniques.

The rest of the paper is organized as follows. In the next section, we outline the CA algorithm for [NDP] and collect some of its properties. The nonmonotone line search technique in [GLL91] and the combined scheme is presented in Section 3. Convergence is established in Section 4 for the general problem [NDP], and in Section 5 for a specialization to differentiable optimization with convex constraints. We conclude in Section 6 with discussions on possible extensions and directions for further research.

2 The cost approximation algorithm

The term *cost approximation* (CA for short) was coined by Patriksson [Pat93b] to describe a framework of descent algorithms for nonlinear programs and variational inequality problems. The CA algorithm is derived as follows. Given an iterate $x^t \in \mathtt{dom}\, u$, we introduce a strongly monotone and Lipschitz continuous *cost approximating mapping* $\Phi^t : \mathtt{dom}\, u \mapsto \Re^n$; *strong monotonicity* of Φ^t is equivalent to the existence of a positive constant m_{Φ^t} such that

$$[\Phi^t(x) - \Phi^t(y)]^{\mathrm{T}}(x-y) \geq m_{\Phi^t}\|x-y\|^2, \qquad \forall x, y \in \mathtt{dom}\, u,$$

and *Lipschitz continuity* to the existence of a positive constant M_{Φ^t} such that

$$\|\Phi^t(x) - \Phi^t(y)\| \leq M_{\Phi^t}\|x-y\|, \qquad \forall x, y \in \mathtt{dom}\, u.$$

(The mapping Φ^t need in general not fulfill as strong conditions in the CA algorithm, but they are required in the nonmonotone version presented in this paper.) Replacing ∇f by Φ^t in [GE] introduces the error $\nabla f - \Phi^t$; this error is approximated by the fixed term $\nabla f(x^t) - \Phi^t(x^t)$. Thus, we arrive at the subproblem of finding y^t such that

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

$$0 \in \Phi^t(y^t) + \partial u(y^t) + \nabla f(x^t) - \Phi^t(x^t).$$

The strong monotonicity assumption on Φ^t ensures that [GE$_{\Phi^t}$] has a unique solution; note that $y^t = x^t$ if and only if $x^t \in \Omega$. If Φ^t is a gradient mapping, then $\Phi^t \equiv \nabla\varphi^t$ for a strongly convex function $\varphi^t : \mathtt{dom}\, u \mapsto \Re$ in C^1 on $\mathtt{dom}\, u$, and [GE$_{\Phi^t}$] reduces to

[NDP$_{\varphi^t}$]

$$\min_{y \in \Re^n} T_{\varphi^t}(y) = \varphi^t(y) + u(y) + [\nabla f(x^t) - \nabla\varphi^t(x^t)]^{\mathrm{T}} y.$$

This problem has the further interpretation of a partial linearization (a linearization of the second term) of the function $T = [\varphi^t + u] + [f - \varphi^t]$ (see also [Pat93c]).

If $y^t \neq x^t$, then $d^t = y^t - x^t$ determines a search direction. A step is taken in the direction of d^t such that a merit function for [GE] is reduced sufficiently, defining the next iterate x^{t+1}; although any real-valued function whose set of (local) minimizers coincide with Ω and enjoying strong enough continuity properties suffices, we shall use $T = f + u$ as the merit function. (If ∇f in [GE] is replaced by a more general function $F : \mathtt{dom}\, u \mapsto \Re^n$, then a merit function for the resulting variational inequality problem with the properties required may be constructed from [NDP$_{\varphi^t}$]; see, e.g., [Pat93a, Pat93b, LaP94].) The step length is often determined through an inexact line search, such as the Armijo step length rule; in some circumstances, the merit function may be too expensive to evaluate, and in such cases a fixed (relaxation) step is used. The CA algorithm is summarized in Table 2.1.

Table 2.1: The cost approximation algorithm

0 (*Initialization*): Choose an initial point $x^0 \in \mathtt{dom}\, u$, and let $t = 0$.

1 (*Search direction generation*): Find the solution y^t to [GE$_{\Phi^t}$]. The resulting search direction is $d^t = y^t - x^t$.

2 (*Termination criterion*): If $d^t = 0 \rightarrow$ Stop ($x^t \in \Omega$). Otherwise, continue.

3 (*Line search*): Choose a step length, l_t, such that $x^t + l_t d^t \in \mathtt{dom}\, u$ and the value of T is reduced sufficiently.

4 (*Update*): Let $x^{t+1} = x^t + l_t d^t$, and $t := t + 1$.

5 (*Termination criterion*): If x^t is acceptable $\rightarrow$ Stop. Otherwise, go to Step 1.

The class of cost approximation algorithms includes many well-known iterative algorithms for [NDP] and its special cases, among others steepest descent, Newton and Gauss–Seidel algorithms for unconstrained optimization, the Frank–Wolfe, gradient projection and regularization algorithms for constrained optimization problems, and multiplier and sequential programming methods for dual formulations of convex programs and saddle point problems. A small sample of special cases is given in Table 2.2, characterized by their respective cost approximating mappings Φ^t. Further examples are found in [Pat93b, Pat94b]. In the table, δ_X denotes the indicator function of a nonempty, closed and convex set X, $\gamma_t > 0$, $r^t : \mathtt{dom}\, u \mapsto \Re^n$ is strongly monotone and Lipschitz continuous on $\mathtt{dom}\, u$, $Q^t \in \Re^{n \times n}$ is positive definite, and $\cdot$ denotes the identity mapping on the appropriate space. Further, we let $\mathcal{C}$ denote an index set and $u_i : \Re^{n_i} \mapsto \Re \cup \{+\infty\}$ be l.s.c., proper and convex functions on $\Re^{n_i}$, with $\sum_{i \in \mathcal{C}} n_i = n$.

The reader should note that in cases where the mapping Φ^t depends on entities associated with the function f (such as in Newton-type methods), its convexity and differentiability requirements introduce *implicit* conditions on f. In such cases, the properties of f

Table 2.2: Instances of the cost approximation algorithm

Problem	Algorithm	Φ^t
$u \equiv \delta_X$	Steepest descent ($X = \Re^n$)/Gradient projection	$(1/\gamma_t)\cdot$
$u \equiv \delta_X$	Deflected gradient ($X = \Re^n$)/Scaled gradient projection	$Q^t\cdot$
$u \equiv \delta_X$	Newton	$\nabla^2 f(x^t)\cdot$
$u \equiv \sum_{i\in\mathcal{C}} u_i$	Jacobi	$\nabla_i f(x^t_{i_-}, \cdot, x^t_{i_+})$, $i \in \mathcal{C}$
$u \equiv \sum_{i\in\mathcal{C}} u_i$	Gauss–Seidel	$\nabla_i f(x^{t+1}_{i_-}, \cdot, x^t_{i_+})$, $i \in \mathcal{C}$
$u \equiv \delta_X$	Frank–Wolfe	0
	Proximal point	$\nabla f + (1/\gamma_t)\cdot$
	Regularization	$\nabla f + (1/\gamma_t) r^t$

in a given problem [NDP] therefore induce restrictions on the possible choices of mappings Φ^t.

The basic properties of the CA algorithms presented in this section are established under the following assumption, in addition to those given in Section 1.

Assumption 2.1.

(1) The directional derivative $u'(x;\cdot)$ is l.s.c., that is, for any $x \in \texttt{dom}\, u$ and $d \in \Re^n$, $u'(x;d) = \liminf_{e\to d} u'(x;e)$.

(2) u is continuous on $\texttt{dom}\, u$.

(3) ∇f is Lipschitz continuous on $\texttt{dom}\, u$.

(4) Let x^0 be the initial solution. Then, the level set $\{\, x \in \Re^n \,|\, T(x) \leq T(x^0) \,\}$ is bounded.

(5) Let $\{\Phi^t\}$ be a sequence of cost approximating mappings adopted in a CA algorithm. Each mapping $\Phi^t : \texttt{dom}\, u \mapsto \Re^n$ is strongly monotone and Lipschitz continuous on $\texttt{dom}\, u$ with modulus m_{Φ^t} and M_{Φ^t}, respectively. Further, $m_\Phi = \liminf_t\{m_{\Phi^t}\} > 0$ and $M_\Phi = \limsup_t\{M_{\Phi^t}\} < +\infty$.

We note that whenever u is the indicator function of a nonempty, closed and convex set, or $\texttt{dom}\, u = \Re^n$, parts (1) and (2) are superfluous. Parts (3) are (5) ensure that the search directions and range of possible step lengths are well-defined, while part (4) ensures that the sequence $\{x^t\}$ is bounded.

The first result establishes the descent properties of the search directions. Throughout this section, it is assumed that $x^t \in \texttt{dom}\, u$, y^t is the unique solution to [GE$_{\Phi^t}$] defined at x^t, and $d^t = y^t - x^t$. By Assumption 2.1.1, the directional derivative of T at x in the direction of d satisfies ([Roc66], [Roc70, Thm. 23.2])

$$T'(x;d) = \nabla f(x)^{\mathrm{T}} d + \sup_{\xi_u \in \partial u(x)} \xi_u^{\mathrm{T}} d.$$

Proposition 2.1 [Pat93b, Pat94a]**.**

- **(a)** $x^t \in \Omega \iff d^t = 0.$
- **(b)** $T'(x^t; d^t) \leq -m_{\Phi^t}\|d^t\|^2.$

We next introduce an inexact line search which may be used in Step 3 of the CA algorithm.

Definition 2.1 (Rule A'). Let $\alpha, \beta \in (0,1)$ and the step length $l_t = \beta^{\bar{\imath}}$, where $\bar{\imath}$ is the smallest nonnegative integer i such that

$$T(x^t + \beta^i d^t) \leq T(x^t) + \alpha\beta^i[\Phi^t(x^t) - \Phi^t(y^t)]^{\mathrm{T}} d^t. \tag{2.1}$$

We note here that Rule A' is a generalization of the classical Armijo rule in unconstrained differentiable optimization; indeed, when $u \equiv 0$, the form of (2.1) becomes (cf. [GE$_{\Phi^t}$])

$$f(x^t + \beta^i d^t) \leq f(x^t) + \alpha\beta^i \nabla f(x^t)^{\mathrm{T}} d^t, \tag{2.2}$$

which is the Armijo rule with a unit initial step length.

Proposition 2.2 [Pat93b, Pat94a] (Validity of Rule A'). *Assume that $x^t \notin \Omega$. Then there exists a finite integer $\bar{\imath}$ satisfying (2.1). Further, $\bar{\imath} = 0$ whenever $2(1-\alpha)m_{\Phi^t}/M_{\nabla f} \geq 1$, and $\beta^{\bar{\imath}} \geq \min\{1, 2\beta(1-\alpha)m_{\Phi^t}/M_{\nabla f}\}$ always holds.*

We note the interesting implication of this result that a proper scaling of the mapping Φ^t ensures that a unit step yields descent.

The basic convergence result for the CA algorithm using Rule A' is given below.

Theorem 2.1 [Pat93b, Pat94a] (Convergence of the CA algorithm). *Let $\{x^t\}$ be a sequence generated by the CA algorithm using Rule A' in Step 3, and $\{d^t\}$ the associated sequence of search directions.*

- **(a)** $\{d^t\} \to 0.$
- **(b)** *The sequence $\{x^t\}$ is bounded, and every accumulation point lies in Ω.*
- **(c)** *If Ω is finite, then $\{x^t\}$ converges.*

The above result implicitly assumes (as will be done henceforth) that the sequence $\{x^t\}$ is infinite; finite termination at some iteration t occurs if and only if $x^t \in \Omega$, according to Proposition 2.1.

An advantage with being able to utilize long steps in the CA algorithm is that the convergence rate improves with the minimal step length used ([Pat93b, Pat94a]); moreover, if unit steps can be taken eventually, then in the presence of explicit, convex constraints, the sequence of iterates finitely identifies the set of active constraints at the limit point, thus eventually reducing the search to a manifold of smaller dimension ([Pat93b]).

3 The nonmonotone CA algorithm

A nonmonotone line search for Newton-type algorithms in unconstrained, differentiable optimization was proposed in [GLL86]. Given a nonnegative integer m, the step length acceptance criterion is of the form

$$f(x^t + \beta^i d^t) \leq \max_{0 \leq j \leq m} \{ f(x^{t-j}) \} + \alpha \beta^i \nabla f(x^t)^{\mathrm{T}} d^t \tag{3.1}$$

(which reduces to (2.2) when $m = 0$). Although this step length rule is less restrictive than (2.2) when $m > 0$, it still requires the iterates x^t to lie in the level sets $\{ x \in \Re^n \mid f(x) \leq \max_{0 \leq j \leq m} \{ f(x^{t-j}) \} \}$, and may still be too restrictive for highly nonlinear problems. The strategy taken in [GLL91] is to allow a unit step length without checking the objective value at all but in occasional *control* points. The convergence test in the remainder of the iterations (the *standard* points) is based on the magnitude of $\|d^t\|$. At a standard point x^t, a unit step is accepted if $\|d^t\|$ is smaller than a prescribed bound, Δ; otherwise, a nonmonotone line search, such as (3.1), is performed. At a control point x^t, the objective value is compared to an adjustable reference value, W. If the objective value is smaller, then we proceed as if it was a standard point; otherwise, a backtracking is made, whereby we restore the variable vector x^t to the last point accepted based on its objective value and perform a nonmonotone line search.

We summarize the nonmonotone CA algorithm in Table 3.1.

In the table, ℓ refers to the latest point accepted based on its objective value. We let $x^{\ell(j)}$ denote these points, and $\{W_j\}$ the sequence of reference values. Initially, $j = 0$, and $j := j + 1$ is set whenever we set $\ell = t$. The reference value W_j is initially set to $T(x^0)$; whenever a point $x^{\ell(j)}$ is generated such that $T(x^{\ell(j)}) < W_j$, this value is updated, taking into account a prefixed number $m(j) \leq M$ of previous objective values. We require that the reference value W_{j+1} satisfies the following assumption.

Assumption 3.1 (Conditions on the reference value)**.** Given $M \geq 0$, let $m(j+1)$ satisfy

$$m(j+1) \leq \min\{m(j) + 1, M\},$$

and define

$$F_{j+1} = \max_{0 \leq i \leq m(j+1)} \{ T(x^{\ell(j+1-i)}) \}.$$

The condition on the reference value W_{j+1} is that

$$T(x^{\ell(j+1)}) \leq W_{j+1} \leq F_{j+1}. \tag{3.3}$$

Several rules for choosing W_{j+1} satisfying (3.3) are given in [GLL91, Toi96]. Further details and some possible extensions are also mentioned.

With the choice $W_j = F_j$ the nonmonotone line search of [GLL86] (generalized to incorporate the function u as in Rule A') is obtained as Δ tends to zero; with $M = 0$, the nonmonotone line search reduces to Rule A'. The method of [GLL91] is obtained when $u \equiv 0$.

Table 3.1: The nonmonotone cost approximation algorithm

0 (*Initialization*): Choose an initial point $x^0 \in \mathtt{dom}\, u$, and let $t = 0$. Let $\Delta_0 > 0$, $0 < \alpha, \beta, \gamma < 1$, and $N \geq 1$. Set $\Delta = \Delta_0$, $\ell = 0$, and $W = T(x^0)$.

1 (*Search direction generation*): Find the solution y^t to [GE$_{\Phi^t}$]. The resulting search direction is $d^t = y^t - x^t$.

2 (*Termination criterion*): If $d^t = 0 \rightarrow$ Stop ($x^t \in \Omega$). If $t = \ell + N$, then go to Step 3.a; otherwise, go to Step 3.b.

3.a (*Control point*): If $T(x^t) \geq W$, then replace x^t by x^ℓ, set $t = \ell$, and go to Step 4. If $T(x^t) < W$, then set $\ell = t$, and update W. If $\|d^t\| \leq \Delta$, then set $x^{t+1} = x^t + d^t$, $t := t + 1$, $\Delta := \gamma\Delta$, and go to Step 1; otherwise, go to Step 4.

3.b (*Standard point*): If $\|d^t\| \leq \Delta$, then set $x^{t+1} = x^t + d^t$, $t := t + 1$, $\Delta := \gamma\Delta$, and go to Step 1. If $\|d^t\| > \Delta$, then: if $T(x^t) \geq W$, then replace x^t by x^ℓ, and set $t = \ell$; otherwise, set $\ell = t$, and update W.

4 (*Nonmonotone line search*): Let $\bar{\imath}$ be the smallest nonnegative integer i such that

$$T(x^t + \beta^i d^t) \leq W + \alpha\beta^i[\Phi^t(x^t) - \Phi^t(y^t)]^{\mathrm{T}} d^t. \tag{3.2}$$

Let $l_t = \beta^{\bar{\imath}}$, $x^{t+1} = x^t + l_t d^t$, $t := t + 1$, $\ell = t$, update W, and go to Step 1.

4 Convergence results

The first two lemmas are direct consequences of the use of the nonmonotone technique.

Lemma 4.1. *Let $\{x^t\}$ be a sequence generated by the nonmonotone CA algorithm.*

(a) *The sequence $\{F_j\}$ is non-increasing and has a limit F^*.*

(b) *For any index j, $F_i < F_j$ for all $i \geq j + M + 1$.*

(c) *The sequence $\{x^t\}$ is bounded.*

Proof The proof is exactly the same as that of Lemma 1 of [GLL91]. □

Lemma 4.2. *Let $\{x^{\ell(j)}\}$ be the sequence of points generated in the nonmonotone CA algorithm where the objective value is evaluated, and let $q(t)$ be an index defined by*

$$q(t) = \max\{j \,|\, \ell(j) \leq t\}.$$

Then there exists a sequence $\{x^{s(j)}\}$ satisfying the following conditions:

(a) *$F_j = T(x^{s(j)})$ for $j = 0, 1, \ldots$.*

(b) *For any integer t there exist indices h_t and j_t such that*

$$0 < h_t - t \le N(M+1), \qquad h_t = s(j_t), \qquad F_{j_t} = T(x^{h_t}) < F_{q(t)}.$$

Proof The proof is exactly the same as that of Lemma 2 of [GLL91]. □

Lemma 4.3. *Let $\{x^t\}$ be a sequence generated by the nonmonotone CA algorithm.*

(a) *$\{T(x^t)\} \to F^*$.*

(b) *$\{\|x^{t+1} - x^t\|\} \to 0$.*

Proof Let $\{x^t\}_{\mathcal{T}}$ denote the (possibly empty) set of points satisfying the norm test at Step 3.a or 3.b, so that

$$\|d^t\| \le \Delta_0 \gamma^k, \qquad t \in \mathcal{T}, \tag{4.1}$$

where the integer k increases with t; when $t \in \mathcal{T}$ we set $l_t = 1$. If $\mathcal{T}$ is infinite then (4.1) yields that

$$\{l_t \|d^t\|\}_{\mathcal{T}} \to 0. \tag{4.2}$$

Let $s(j)$ and $q(t)$ be the indices defined in Lemma 4.2. We show by induction that, for any fixed positive integer i,

$$\{l_{s(j)-i} \|d^{s(j)-i}\|\} \to 0 \tag{4.3}$$

and

$$\{T(x^{s(j)-i})\} \to F^*. \tag{4.4}$$

(We assume that j is large enough so that no negative indices occur.) Assume first that $i = 1$. If $s(j) - 1 \in \mathcal{T}$, then (4.3) holds with $t = s(j) - 1$. Otherwise, by Lemma 4.2 and (3.2),

$$\begin{aligned} F_j &= T(x^{s(j)}) = T(x^{s(j)-1} + l_{s(j)-1} d^{s(j)-1}) \\ &\le F_{q(s(j)-1)} + \alpha l_{s(j)-1} [\Phi^{s(j)-1}(x^{s(j)-1}) - \Phi^{s(j)-1}(y^{s(j)-1})]^{\mathrm{T}} d^{s(j)-1}, \end{aligned}$$

from which it follows that

$$F_{q(s(j)-1)} - F_j \ge \alpha l_{s(j)-1} [\Phi^{s(j)-1}(y^{s(j)-1}) - \Phi^{s(j)-1}(x^{s(j)-1})]^{\mathrm{T}} d^{s(j)-1}. \tag{4.5}$$

Therefore, if $s(j) - 1 \notin \mathcal{T}$ for an infinite subsequence, then from Assumption 2.1.5, Lemma 4.1.a and (4.5) we obtain that $\{\alpha l_{s(j)-1} \|d^{s(j)-1}\|\} \to 0$ in this subsequence. Hence, (4.3) holds for $i = 1$. Moreover, since $T(x^{s(j)}) = T(x^{s(j)-1} + l_{s(j)-1} d^{s(j)-1})$, by (4.3) and the uniform continuity of T on the compact set containing $\{x^t\}$, (4.4) holds for $i = 1$.

Assume that (4.3) and (4.4) holds for a given i and consider the point $x^{s(j)-i-1}$. Reasoning as before, we can again distinguish between the case $s(j) - i - 1 \in \mathcal{T}$, whence (4.1) holds with $t = s(j) - i - 1$, and the case $s(j) - i - 1 \notin \mathcal{T}$, whence $T(x^{s(j)-i}) \leq F_{q(s(j)-i-1)} + \alpha l_{q(s(j)-i-1)}[\Phi^{s(j)-i-1}(x^{s(j)-i-1}) - \Phi^{s(j)-i-1}(y^{s(j)-i-1})]^{\mathrm{T}} d^{s(j)-i-1}$, and hence

$$\begin{aligned} &F_{q(s(j)-i-1)} - T(x^{s(j)-i}) \geq \\ &\alpha l_{q(s(j)-i-1)}[\Phi^{s(j)-i-1}(y^{s(j)-i-1}) - \Phi^{s(j)-i-1}(x^{s(j)-i-1})]^{\mathrm{T}} d^{s(j)-i-1}. \end{aligned} \tag{4.6}$$

Then, using (4.2), (4.4), (4.6) and Assumption 2.1.5, (4.3) holds with i replaced by $i+1$. By (4.3) and the uniform continuity of T, it follows that also (4.4) holds with i replaced by $i+1$. This completes the induction.

Let x^t be any iteration point. By Lemma 4.2 there is an $x^{h_t} \in \{x^{s(j)}\}$ such that

$$0 < h_t - t \leq (M+1)N. \tag{4.7}$$

But

$$x^t = x^{h_t} - \sum_{i=1}^{h_t - t} l_{h_t - i} d^{h_t - i}$$

implies, by (4.3) and (4.7), that

$$\{\|x^t - x^{h_t}\|\} \to 0.$$

From the uniform continuity of T it follows that

$$\lim_{t\to\infty} T(x^t) = \lim_{t\to\infty} T(x^{h_t}) = \lim_{j\to\infty} F_j, \tag{4.8}$$

which establishes (a).

If $t \notin \mathcal{T}$, then $T(x^{t+1}) \leq F_{q(t)} + \alpha l_t[\Phi^t(x^t) - \Phi^t(y^t)]^{\mathrm{T}} d^t$, whence

$$F_{q(t)} - T(x^{t+1}) \geq \alpha l_t[\Phi^t(y^t) - \Phi^t(x^t)]^{\mathrm{T}} d^t \tag{4.9}$$

follows. Therefore, by (4.2), (4.8), (4.9) and Assumption 2.1.5, we may conclude that

$$\{l_t\|d^t\|\} \to 0,$$

which establishes (b). □

The following is the main result of this paper.

Theorem 4.1 (Convergence of the nonmonotone CA algorithm)**.** *Let $\{x^t\}$ be a sequence generated by the nonmonotone CA algorithm.*

- **(a)** *The sequence $\{x^t\}$ is bounded, and every accumulation point lies in Ω.*
- **(b)** *No accumulation point is a local maximizer of T.*
- **(c)** *If Ω is finite, then $\{x^t\}$ converges.*

Proof The main effort is to show that $\{d^t\} \to 0$; this we prove by contradiction.

Let x^∞ be any accumulation point of the sequence $\{x^t\}$; the existence of such a point follows from the boundedness of $\{x^t\}$ (cf. Lemma 4.1.c). Arguing by contradiction, Lemma 4.3.b yields that there must be some subsequence $\mathcal{T}$ for which $\{l_t\}_\mathcal{T} \to 0$ and $\{d^t\}_\mathcal{T} \to d^\infty$. For a sufficiently large index $\bar{t}$, it must then be the case that for every $t \geq \bar{t}$, the point x^{t+1} is produced through the nonmonotone line search with a resulting step length less than one, and hence

$$T\left(x^t + \frac{l_t}{\beta}d^t\right) > W_{q(t)} + \alpha\frac{l_t}{\beta}[\Phi^t(x^t) - \Phi^t(y^t)]^\mathrm{T}d^t \geq T(x^t) + \alpha\frac{l_t}{\beta}[\Phi^t(x^t) - \Phi^t(y^t)]^\mathrm{T}d^t.$$

The subproblem $[\mathrm{GE}_{\Phi^t}]$ is characterized by the inclusion

$$-\Phi^t(y^t) - \nabla f(x^t) + \Phi^t(x^t) \in \partial u(y^t);$$

letting $\xi_u(y^t)$ be an element of $\partial u(y^t)$ such that

$$\xi_u(y^t) = -\Phi^t(y^t) - \nabla f(x^t) + \Phi^t(x^t), \tag{4.10}$$

we then have that

$$[\Phi^t(x^t) - \Phi^t(y^t)]^\mathrm{T}d^t = [\nabla f(x^t) + \xi_u(y^t)]^\mathrm{T}d^t;$$

further, using the monotonicity of ∂u for convex functions u, we obtain that

$$[\Phi^t(x^t) - \Phi^t(y^t)]^\mathrm{T}d^t \geq [\nabla f(x^t) + \xi_u(x^t)]^\mathrm{T}d^t$$

holds for any subgradient $\xi_u(x^t)$ of u at x^t. After dividing the resulting inequality by l_t/β, we obtain, for all sufficiently large t, that

$$\frac{\beta}{l_t}\left[T\left(x^t + \frac{l_t}{\beta}d^t\right) - T(x^t)\right] \geq \alpha[\nabla f(x^t) + \xi_u(x^t)]^\mathrm{T}d^t.$$

As t tends to infinity in $\mathcal{T}$, the left–hand–side of this inequality tends to a limit, which by Assumption 2.1.1 and [Roc81, Prop. 3G] equals $T'(x^\infty; d^\infty)$. By the closedness of the mapping ∂u ([Roc70, Thm. 24.4]) we may choose the sequence of subgradients $\xi_u(x^t)$ in the right–hand–side such that $\{\xi_u(x^t)\}_\mathcal{T} \to \xi_u^\infty \in \partial u(x^\infty)$, and, by again appealing to Assumption 2.1.1, such that $T'(x^\infty; d^\infty) = \nabla f(x^\infty) + \xi_u^\infty$. With this choice, in the limit of $\mathcal{T}$, the right–hand–side of the above inequality equals $\alpha T'(x^\infty; d^\infty)$, and since $\alpha < 1$, the inequality reduces to

$$T'(x^\infty; d^\infty) \geq 0.$$

To reach a contradiction we utilize the fact that, in the limit of $\mathcal{T}$, Proposition 2.1.b yields that

$$T(x^\infty; d^\infty) \leq -m_\Phi \|d^\infty\|^2,$$

where $m_\Phi > 0$ by Assumption 2.1.5. Hence, it must be the case that

$$\{d^t\} \to 0. \tag{4.11}$$

For all t, let $\xi_u(y^t)$ be a subgradient of u at y^t such that (4.10) holds. By (4.11) and Assumption 2.1.5,

$$\{\nabla f(x^t) + \xi_u(y^t)\} \to 0.$$

In the limit of the subsequence corresponding to the accumulation point x^∞, $\{\xi_u(y^t)\} \to \xi_u^*$, which, by (4.11) and the closedness of ∂u, belongs to $\partial u(x^\infty)$. Since, by the above, we must have that $\nabla f(x^\infty) + \xi_u^* = 0$, $x^\infty \in \Omega$ follows. This establishes (a).

The result (b) follows by identical arguments as in the proof of Theorem 1.b of [GLL91], and (c) from known results for sequences satisfying the result of Lemma 4.3.b (see [OrR70, Thm. 14.1.5]). □

We remark that the proof of (4.11) can be made also using Proposition 2.2: since the nonmonotone line search (3.2) is less restrictive than Rule A' $[W \geq T(x^t)]$, it follows that the step length resulting from the nonmonotone line search must be at least as great as that given by Rule A'. It then follows from Proposition 2.2 that the step length in (3.2) satisfies $l_t \geq \min\{1, 2\beta(1-\alpha)m_{\Phi^t}/M_{\nabla f}\}$. By Assumption 2.1.5,

$$\inf_t \{l_t\} \geq \min\{1, 2\beta(1-\alpha)m_\Phi/M_{\nabla f}\} > 0,$$

and hence, by Lemma 4.3.b, (4.11) must hold.

5 A specialization to differentiable optimization

Assume that the function u is the indicator of a nonempty, closed and convex set $X \subseteq \Re^n$. We briefly mention here the corresponding differences in the nonmonotone CA algorithm and the convergence results presented above.

With this choice of function u, the problem [NDP] reduces to the solution of the constrained differentiable optimization problem

[CDP]

$$\min_{x \in X} f(x),$$

and the optimality conditions [GE] reduces to

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

where $N_X(x^*)$ is the normal cone to X at x^*. In the CA algorithm and its nonmonotone version, the merit function is f; as a result, Assumption 2.1.1 and 2.1.2 are no longer necessary. Continuing with the list of assumptions, Assumption 2.1.3 can here be written as the condition that the set $\{\, x \in X \mid f(x) \leq f(x^0) \,\}$ is bounded; this is in particular true when X is bounded. Proposition 2.1.b is replaced by the result that $\nabla f(x^t)^{\mathrm{T}} d^t \leq -m_{\Phi^t}\|d^t\|^2$. Further, Rule A' may be replaced by the Armijo rule (2.2) with no change in the results of Proposition 2.2 or Theorem 2.1 (cf. [Pat93b, Le. 4.3, and Le. 4.4 and Thm. 4.3.b]); the criterion (3.2) of the nonmonotone line search is replaced by

$$f(x^t + \beta^i d^t) \leq W + \alpha\beta^i \nabla f(x^t)^{\mathrm{T}} d^t.$$

No other changes in the algorithm are necessary, and the results of Section 4 applies with T replaced by f.

6 Extensions and further research

The promising results obtained in experiments reported in [GLL91, Toi96] motivate a further study and application of nonmonotone CA algorithms in the more general framework considered in this paper. We note especially that (possibly ill-conditioned) problems of the form [NDP] arise naturally in applications of exact penalty functions, and in more general penalty methods where only a subset of the complete set of constraints are penalized.

While a unit step is the natural choice of default step size in Newton-type methods, whether such a choice is natural in the more general setting of this paper depends largely on the properties of Φ^t (as is, for example, seen from the result of Prop. 2.2). We then observe that the results obtained above still hold if the update $x^{t+1} = x^t + d^t$ in Step 3 of the nonmonotone CA algorithm is replaced by the more general update $x^{t+1} = x^t + \gamma_t d^t$, with $0 < \varepsilon \leq \gamma_t \leq 1$; the value of the default step length γ_t can, for example, be an upper bound on the maximal descent step length given in Proposition 2.2.

A particularly interesting type of problem for applications of the nonmonotone methods of this paper is large-scale differentiable optimization problems over Cartesian product sets. Such problems are amenable to parallel decomposition (e.g., [Pat97]), but Newton-like methods may not be applicable in such a setting because the quadratic subproblem would not be separable with respect to the partition of $\Re^n$ defined by the Cartesian product. Choosing the subproblems to be separable, on the other hand, lessens the quality of the subproblem solutions, because the coupling among the variable components in the objective function is not taken into account, and this may result in rather short steps in a line search; this is illustrated in the convergence analysis of decomposition versions of the CA algorithm in [Pat97]. In the context of parallel decomposition, a nonmonotone technique can be beneficial for two reasons: (1) it allows for longer steps to be taken, thereby enabling an increased speed of convergence; and (2) it lessens the need to perform line searches; since these constitute serial operations, the result is an improved parallelism.

An interesting subject for future research is the validation and application of nonmonotone CA algorithms for more general problems, especially for generalizations of [GE] to problems with non-gradient mappings F, that is, for variational inequality problems (see Section 2); this is of special interest since the merit functions employed in descent algorithms for these problems (see [LaP94, Pat93b]) are very expensive to evaluate; a relaxation of the descent requirement would thus substantially reduce the computational burden associated with, for example, an Armijo-type backtracking line search.

The large freedom of choice of both problem and method instances makes it difficult to select a small test sample for inclusion in the present paper; we have decided to relegate a more substantial test to a future paper.

References

[BPTZ92] J. F. Bonnans, E. R. Panier, A. L. Tits, and J. L. Zhou, *Avoiding the Maratos effect by means of a nonmonotone line search, II. Inequality constrained problems—feasible iterates*, SIAM Journal on Numerical Analysis, 29 (1992), pp. 1187–1202.

[CPLP82] R. M. Chamberlain, M. J. D. Powell, C. Lemaréchal, and H. C. Pedersen, *The watchdog technique for forcing convergence in algorithms for constrained optimization*, Mathematical Programming Study, 16 (1982), pp. 1–17.

[DXZ93] N. Y. Deng, Y. Xiao, and F. J. Zhou, *Nonmonotonic trust region algorithm*, Journal of

Optimization Theory and Applications, 76 (1993), pp. 259–285.

[DeM74] J. E. Dennis and J. J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Mathematics of Computation, 28 (1974), pp. 549–560.

[DeM77] J. E. Dennis and J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM Review, 19 (1977), pp. 46–89.

[DiF95] S. P. Dirkse and M. C. Ferris, *The PATH solver: A non-monotone stabilization scheme for mixed complementarity problems*, Optimization Methods & Software, 5 (1995), pp. 123–156.

[FaL93] F. Facchinei and S. Lucidi, *Nonmonotone bundle-type scheme for convex nonsmooth minimization*, Journal of Optimization Theory and Applications, 76 (1993), pp. 241–257.

[FeL94] M. C. Ferris and S. Lucidi, *Nonmonotone stabilization methods for nonlinear equations*, Journal of Optimization Theory and Applications, 81 (1994), pp. 53–71.

[FLR96] M. C. Ferris, S. Lucidi, and M. Roma, *Nonmonotone curvilinear line search methods for unconstrained optimization*, Computational Optimization and Applications, 6 (1996), pp. 117–136.

[GLL86] L. Grippo, F. Lampariello, and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM Journal on Numerical Analysis, 23 (1986), pp. 707–716.

[GLL89] L. Grippo, F. Lampariello, and S. Lucidi, *A truncated Newton method with nonmonotone line search for unconstrained optimization*, Journal of Optimization Theory and Applications, 60 (1989), pp. 401–419.

[GLL90] L. Grippo, F. Lampariello, and S. Lucidi, *A quasi-discrete Newton algorithm with a nonmonotone stabilization technique*, Journal of Optimization Theory and Applications, 64 (1990), pp. 495–510.

[GLL91] L. Grippo, F. Lampariello, and S. Lucidi, *A class of nonmonotone stabilization methods in unconstrained optimization*, Numerische Mathematik, 59 (1991), pp. 779–805.

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

[Mar78] N. Maratos, *Exact penalty function algorithms for finite dimensional and control optimization problems*, Doctoral dissertation, Imperial College of Science and Technology, University of London, London, 1978.

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

[PHR91] J.-S. Pang, S. P. Han, and N. Rangaraj, *Minimization of locally Lipschitz functions*, SIAM Journal on Optimization, 1 (1991), pp. 57–82.

[PaT91] E. R. Panier and A. L. Tits, *Avoiding the Maratos effect by means of a nonmonotone line search, I. General constrained problems*, SIAM Journal on Numerical Analysis, 28 (1991), pp. 1183–1195.

[Pat93a] M. Patriksson, *A descent algorithm for a class of generalized variational inequalities*, Report LiTH-MAT-R-93-35, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1993. Revised 1996 for possible publication in Optimization.

[Pat93b] M. Patriksson, *A unified framework of descent algorithms for nonlinear programs and variational inequalities*, Doctoral dissertation, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1993.

[Pat93c] M. Patriksson, *Partial linearization methods in nonlinear programming*, Journal of Optimization Theory and Applications, 78 (1993), pp. 227–246.

[Pat94a] M. Patriksson, *Cost approximation: A unified framework of descent algorithms for nonlinear programs*, Report, Department of Mathematics, Box 354350, University of Washington, Seattle, WA 98195-4350, 1994. Revised 1996 for possible publication in SIAM Journal on Optimization.

[Pat94b] M. Patriksson, *A taxonomy of descent algorithms for nonlinear programs and variational inequalities*, Report, Department of Mathematics, Box 354350, University of Washington, Seattle, WA 98195-4350, 1994.

[Pat97] M. Patriksson, *Decomposition methods for differentiable optimization problems on Cartesian product sets*, Computational Optimization and Applications (1997) (to appear).

[Ree93] C.R. Reeves, *Modern Heuristic Techniques for Combinatorial Problems*, Blackwell Scientific Publications, Oxford, 1993.

[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 Journal of Mathematics, 17 (1966), pp. 497–510.

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[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.

[Toi94] Ph. L. Toint, *A non-monotone trust-region algorithm for nonlinear optimization subject to convex constraints*, Report 94/24, Department of Mathematics, Facultés Universitaires ND de la Paix, Namur, Belgium, 1994.

[Toi96] Ph. L. Toint, *An assessment of non-monotone linesearch techniques for unconstrained optimization*, SIAM Journal on Scientific Computing, 17 (1996), pp. 725–739.