Optimization

A Journal of Mathematical Programming and Operations Research

ISSN: 0233-1934 (Print) 1029-4945 (Online) Journal homepage: https://www.tandfonline.com/loi/gopt20

Worst-case evaluation complexity of non-monotone gradient-related algorithms for unconstrained optimization

C. Cartis, Ph.R. Sampaio & Ph.L. Toint

To cite this article: C. Cartis, Ph.R. Sampaio & Ph.L. Toint (2015) Worst-case evaluation complexity of non-monotone gradient-related algorithms for unconstrained optimization, Optimization, 64:5, 1349-1361, DOI: 10.1080/02331934.2013.869809

To link to this article: https://doi.org/10.1080/02331934.2013.869809

Published online: 13 Jan 2014.

Optimization, 2015
Vol. 64, No. 5, 1349–1361, http://dx.doi.org/10.1080/02331934.2013.869809

Taylor & Francis
Taylor & Francis Group

Worst-case evaluation complexity of non-monotone gradient-related algorithms for unconstrained optimization

C. Cartis[a], Ph.R. Sampaio[b] and Ph.L. Toint[b*]

[a] *School of Mathematics, University of Edinburgh, Edinburgh, UK;* [b] *Namur Center for Complex Systems (naXys) and Department of Mathematics, University of Namur, Namur, Belgium*

(*Received 12 March 2013; accepted 5 November 2013*)

The worst-case evaluation complexity of finding an approximate first-order critical point using gradient-related non-monotone methods for smooth non-convex and unconstrained problems is investigated. The analysis covers a practical linesearch implementation of these popular methods, allowing for an unknown number of evaluations of the objective function (and its gradient) per iteration. It is shown that this class of methods shares the known complexity properties of a simple steepest-descent scheme and that an approximate first-order critical point can be computed in at most $O(\epsilon^{-2})$ function and gradient evaluations, where $\epsilon > 0$ is the user-defined accuracy threshold on the gradient norm.

Keywords: non-linear optimization; evaluation complexity; worst-case analysis; linesearch algorithms; non-monotone methods

1. Introduction

The worst-case evaluation complexity of optimization algorithms applied to non-linear and potentially non-convex problems has been studied in a sequence of recent papers, both for the unconstrained case [1–4] and for the constrained one [5,6]. Of particular interest here are the results of Nesterov [1, p.29], in which this author analyses the worst-case behaviour of the steepest-descent method for unconstrained minimization (both for exact and approximate linesearches) and shows that an approximate first-order critical point, that is a point at which the norm of the gradient of the objective function is less than $\epsilon > 0$, must be obtained in at most $O(\epsilon^{-2})$ iterations. Nesterov's analysis of the steepest-descent variants therefore effectively assumes that a single objective function value per iteration is computed, or at least that the number of such evaluations in the course of a single iteration is bounded. His bounds thus specify iteration-complexity rather than evaluation complexity. At variance, more typical implementations use a linesearch (which makes no explicit use of the Lipschitz constant) to compute a suitable steplength, with the possible drawback that an unknown number of additional function evaluations may be required during the course of a single iteration. The question of the worst-case objective-function evaluation complexity of linesearch implementations of this type has not yet been considered specifically. Interestingly, a worst-case complexity analysis is available for other

*Corresponding author. Email: philippe.toint@unamur.be

first-order algorithms, such as first-order trust-region methods [2] and first-order regularization algorithms [7].

In parallel, it has long been known that 'gradient related' minimization methods share a number of their convergence properties with the steepest-descent algorithm (see Ortega and Rheinboldt [8] for an early reference). In these methods, a linesearch is performed along a direction whose angle with the negative gradient is bounded away from orthogonality. This class covers a wide range of practical algorithms, including for instance variable-metric techniques or finite-difference schemes when Hessian approximations have bounded conditioning (see Nocedal and Wright [9,p.40] for instance). Despite their close connection with steepest descent, their worst-case analysis remains so far an open question.

Standard linesearch methods are usually defined in a way which ensures monotonically decreasing objective-function values as the iterations proceed. However, 'non-monotone' generalizations of these algorithms, where this monotonicity property is abandoned, have gained respect in practice because of their often better performance. We refer the reader to Grippo et al. [10,11], or Toint [12] for more details on these methods. Grippo and co-authors also provided a convergence analysis supporting their proposals. However, the worst-case performance of this interesting class of algorithms is so far unexplored.

The purpose of this paper is to bring together these three questions (standard linesearch, gradient-related directions and non-monotonicity) and to provide an analysis which covers them all. We therefore consider non-monotone gradient-related linesearch optimization methods and show that, as for steepest-descent, their objective-function evaluation complexity is $O(\epsilon^{-2})$. Note that standard monotone variants are also covered by this analysis.

Section 2 states the problem and describes the class of algorithms considered, while Section 3 provides an upper bound on their worst-case evaluation complexity. Some comments and perspectives are finally presented in Section 4.

2. The problem and algorithm

We consider the non-linear and possibly non-convex smooth unconstrained minimization problem

$$\min_{x \in \Re^n} f(x) \tag{2.1}$$

for which we assume the following.

$\boxed{\textbf{AF0}}$ $f(x)$ is bounded below on $\Re^n$, that is there exists a constant[1] $\kappa_{\rm lbf}$ such that, for all $x \in \Re^n$, $f(x) \geq \kappa_{\rm lbf}$.

$\boxed{\textbf{AF1}}$ $f(x)$ is continuously differentiable on $\Re^n$.

As stated in the introduction, we consider a class of algorithms in which the search directions are 'gradient-related' (see Ortega and Rhienboldt [8] and Bertsekas [13, p.35]). This terminology means that, at iteration k, an approximate unidimensional minimization of the objective function is performed along a direction d_k whose angle with the steepest descent is controlled by the condition

$$\langle g_k, d_k \rangle \leq -\kappa_1 \|g_k\|^2 \text{ and } \|d_k\| \leq \kappa_2 \|g_k\|, \tag{2.2}$$

where $g_k \overset{\rm def}{=} g(x_k) \overset{\rm def}{=} \nabla_x f(x_k)$, x_k is the k-th iterate, $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ are the Euclidean inner product and norm, respectively, and κ_1 and κ_2 are positive constants independent of k.

Once the direction is fixed, it is then used in a non-monotone linesearch. We choose here a Goldstein-Armijo variant (see Grippo et al. [10] or Nocedal and Wright [9, p.33–37]), in which a stepsize t_k (yielding a new iterate $x_{k+1} = x_k + t_k d_k$) is accepted whenever the conditions

$$f(x_k + t_k d_k) \leq \max_{0 \leq j \leq M} \left[f(x_{k-j}) \right] + \alpha\, t_k \langle g_k, d_k \rangle, \tag{2.3}$$

and

$$f(x_k + t_k d_k) \geq \max_{0 \leq j \leq M} \left[f(x_{k-j}) \right] + \beta\, t_k \langle g_k, d_k \rangle \tag{2.4}$$

hold, where $M \geq 0$, $\alpha \in (0, 1)$ and $\beta \in (\alpha, 1)$ are constants independent of k, and where, by convention, $x_{-M} = \ldots x_{-1} = x_0$. Note that $M = 0$ corresponds to the monotone case.

The class of algorithms of interest may now be stated formally as Algorithm 2.1.

Algorithm 2.1 A gradient-related non-monotone linesearch algorithm.

Step 0: **Initialization.** An initial point x_0 is given, as well as an accuracy level $\epsilon > 0$. The constants t_{ini}, M, α and β are also given, satisfying $t_{\text{ini}} > 0$, $M \geq 0$ and $0 < \alpha < \beta < 1$. Compute $f(x_0)$, g_0 and set $k = 0$.

Step 1: **Test for termination.** If $\|g_k\| \leq \epsilon$, terminate.

Step 2: **Select a search direction.** Choose d_k such that (2.2) holds.

Step 3: **Linesearch: test initial stepsize.**

1. Set $t_k = t_{\text{ini}} > 0$, $t_{\text{low}} = 0$ and compute $f(x_k + t_k d_k)$.
2. If (2.3) fails, go to Step 4.
3. If (2.4) fails, go to Step 5.
4. Else go to Step 7.

Step 4: **Linesearch: backtracking.**

1. While (2.3) fails, set $t_{\text{up}} \longleftarrow t_k$, $t_k \longleftarrow \frac{1}{2} t_k$ and compute $f(x_k + t_k d_k)$.
2. If (2.4) holds, go to Step 7, or set $t_{\text{low}} \longleftarrow t_k$ and go to Step 6 otherwise.

Step 5: **Linesearch: look ahead.**

1. While (2.4) fails, set $t_{\text{low}} \longleftarrow t_k$, $t_k \longleftarrow 2t_k$ and compute $f(x_k + t_k d_k)$.
2. If (2.3) holds, go to Step 7, or set $t_{\text{up}} \longleftarrow t_k$ and go to Step 6 otherwise.

Step 6: **Linesearch: bisect inside bracket.**

1. Set $t_k \longleftarrow \frac{1}{2}(t_{\text{low}} + t_{\text{up}})$ and compute $f(x_k + t_k d_k)$.
2. If (2.3) fails, set $t_{\text{up}} \longleftarrow t_k$ and return to Step 6.
3. If (2.4) fails, set $t_{\text{low}} \longleftarrow t_k$ and return to Step 6.

Step 7: **Compute the new iterate and gradient.** Set $x_{k+1} = x_k + t_k d_k$ and compute $g_{k+1} = g(x_{k+1})$. Increment k by one and return to Step 1.

Note that the successive phases of the Goldstein-Armijo technique are apparent in the algorithm's description: a bracket containing the desired step is first identified by bactracking (Step 4) or look-ahead (Step 5), and the final step is then computed by bisection (Step 6).

3. Worst-case evaluation complexity analysis

We now analyse the worst-case behaviour of Algorithm 2.1. A first step in this analysis is to specify our assumptions.

$\boxed{\textbf{AF2}}$ $g(x)$ is Lipschitz continuous on $\Re^n$, that is there exists a constant $L_g > 0$ such that, for all $x, y \in \Re^n$,

$$\|g(x) - g(y)\| \le L_g \|x - y\|.$$

The first simple but crucial property that can be deduced from these assumptions is that the stepsize is bounded below by a constant inversely proportional to the Lipschitz constant L_g.

Lemma 3.1 *Suppose that* AF0–AF2 *hold. Then any value of* $t > 0$ *such that* (2.4) *holds for* $t_k = t$ *also satisfies the inequality*

$$t \ge \frac{2(1-\beta)\kappa_1}{L_g \kappa_2^2}. \tag{3.1}$$

Proof We successively use the mean value theorem, the Cauchy-Schwarz inequality and AF2 to obtain that

$$\begin{aligned} f(x_k + t d_k) &= f(x_k) + t\langle g_k, d_k\rangle + \int_0^1 \langle g(x_k + \tau t d_k) - g_k, t d_k\rangle d\tau \\ &\le f(x_k) + t\langle g_k, d_k\rangle + t\|d_k\| \int_0^1 \|g(x_k + \tau t d_k) - g_k\| d\tau \\ &\le f(x_k) + t\langle g_k, d_k\rangle + \tfrac{1}{2} t^2 L_g \|d_k\|^2 \\ &\le \max_{0\le j\le M} [f(x_{k-j})] + t\langle g_k, d_k\rangle + \frac{1}{2} t^2 L_g \|d_k\|^2. \end{aligned} \tag{3.2}$$

Combining this relation with (2.4) and (2.2), we have that

$$t \ge \frac{2\langle g_k, d_k\rangle(\beta - 1)}{L_g \|d_k\|^2} \ge \frac{2(1-\beta)\kappa_1 \|g_k\|^2}{L_g \|g_k\|^2 \kappa_2^2} = \frac{2(1-\beta)\kappa_1}{L_g \kappa_2^2}. \tag{3.3}$$

□

We now prove that there is a finite and non-empty interval of acceptable stepsizes.

Lemma 3.2 *Suppose that* AF0–AF1 *hold and that* $g_k \neq 0$. *Then there exists an interval* $\left[t_k^\beta, t_k^\alpha\right]$ *such that*

$$0 < t_k^\beta < t_k^\alpha < +\infty \tag{3.4}$$

and (2.3)–(2.4) *hold for every value of* $t_k \in \left[t_k^\beta, t_k^\alpha\right]$.

Proof Observe first that the slope of $f(x_k + td_k)$ is steeper than that of the straight lines $f(x_k) + \alpha t \langle g_k, d_k \rangle$ and $f(x_k) + \beta t \langle g_k, d_k \rangle$, $(t \geq 0)$, since $\alpha < 1$ and $\beta < 1$. Thus, for all $t > 0$ sufficiently small,

$$f(x_k + td_k) < f(x_k) + \alpha t \langle g_k, d_k \rangle \leq \max_{0 \leq j \leq M} f(x_{k-j}) + \alpha t \langle g_k, d_k \rangle \tag{3.5}$$

and

$$f(x_k + td_k) < f(x_k) + \beta t \langle g_k, d_k \rangle \leq \max_{0 \leq j \leq M} f\left(x_{k-j}\right) + \beta t \langle g_k, d_k \rangle. \tag{3.6}$$

It follows from (3.5) that (2.3) holds for all t_k sufficiently small. Furthermore, (2.3) does not hold in the limit as $t_k = t \to \infty$ since $f(x_k) + \alpha t \langle g_k, d_k \rangle \leq f(x_k) - \alpha t \kappa_1 \|g_k\|^2 \to -\infty$ (because of (2.2)), while $f(x_k + td_k) \geq \kappa_{\text{lbf}}$ for all t due to AF0. Thus there exists a value $0 < t_k^\alpha < \infty$ such that

$$f(x_k + t_k^\alpha d_k) = \max_{0 \leq j \leq M} f\left(x_{k-j}\right) + \alpha t_k^\alpha \langle g_k, d_k \rangle. \tag{3.7}$$

For simplicity, let us choose the smallest t_k^α that satisfies (3.7) so that (3.5) holds for all $t \in (0, t_k^\alpha)$. Since $\alpha < \beta < 1$, we note that

$$\max_{0 \leq j \leq M} f\left(x_{k-j}\right) + \beta t \langle g_k, d_k \rangle < \max_{0 \leq j \leq M} f\left(x_{k-j}\right) + \alpha t \langle g_k, d_k \rangle \quad \text{for all} \quad t > 0.$$

Letting $t = t_k^\alpha$ in this inequality and using (3.7), we deduce that (2.4) must continue to hold for $0 < t_k = t < t_k^\alpha$ sufficiently close to t_k^α. However, (3.6) implies that (2.4) must fail for sufficiently small $t > 0$, and using again AF1, we conclude that there exists $0 < t_k^\beta < t_k^\alpha$ such that

$$f(x_k + t_k^\beta d_k) = \max_{0 \leq j \leq M} f(x_{k-j}) + \beta t_k^\beta \langle g_k, d_k \rangle, \tag{3.8}$$

and (2.4) holds for all $t_k \in [t_k^\beta, t_k^\alpha]$. (Clearly, t_k^β must be distinct from $t_k^\alpha < \infty$ due to (3.7), (3.8) and $\alpha < \beta$). This concludes the proof since (2.3) holds for t_k in the same interval due to (3.5) and the definition of t_k^α. □

Having proved the existence of an interval of acceptable stepsizes, we now verify that the measure of this interval is bounded below by some positive constant.

LEMMA 3.3 *Suppose that* AF0–AF2 *hold, and define* t_k^α *and* t_k^β *to be any solutions of* (3.7) *and* (3.8)*, respectively, such that* (2.3) *and* (2.4) *hold for each* $t \in [t_k^\beta, t_k^\alpha]$. *Then the interval* $[t_k^\beta, t_k^\alpha]$ *has a strictly positive measure in the sense that there exists a constant* $\kappa_{\text{int}} > 0$ *only depending on* α, β, κ_1, κ_2 *and* L_g *such that*

$$t_k^\alpha - t_k^\beta \geq \kappa_{\text{int}}. \tag{3.9}$$

Proof Assume first that $f(x_k + t_k^\alpha d_k) \leq f(x_k + t_k^\beta d_k)$. Then (3.7) and (3.8) imply that $\alpha t_k^\alpha > \beta t_k^\beta$, and so, using also Lemma 3.1,

$$t_k^\alpha - t_k^\beta \geq \frac{\beta - \alpha}{\alpha} t_k^\beta \geq \frac{2(\beta - \alpha)(1 - \beta)\kappa_1}{\alpha L_g \kappa_2^2}. \tag{3.10}$$

Suppose now that $f(x_k + t_k^\alpha d_k) > f(x_k + t_k^\beta d_k)$. Applying the mean value theorem to $f(x+td)$ on $[t_k^\beta, t_k^\alpha]$ yields that

$$\begin{aligned} f\left(x_k + t_k^\alpha d_k\right) - f\left(x_k + t_k^\beta d_k\right) &= \left(t_k^\alpha - t_k^\beta\right) \langle g\left(x_k + t_\xi d_k\right), d_k\rangle \\ &\le \left(t_k^\alpha - t_k^\beta\right) \|g\left(x_k + t_\xi d_k\right)\| \, \|d_k\| \\ &\le \left(t_k^\alpha - t_k^\beta\right) \kappa_2 \|g\left(x_k + t_\xi d_k\right)\| \, \|g_k\|, \end{aligned}$$

where $t_\xi \in \left(t_k^\beta, t_k^\alpha\right)$ and where the first inequality follows from the Cauchy-Schwarz inequality and the second from (2.2). Furthermore, the Lipschitz continuity of g (AF2), (2.2) and the bound $t_\xi < t_k^\alpha$ give that

$$\begin{aligned} \|g\left(x_k + t_\xi d_k\right)\| &\le \|g\left(x_k + t_\xi d_k\right) - g_k\| + \|g_k\| \\ &\le L_g t_\xi \|d_k\| + \|g_k\| \\ &\le L_g t_\xi \kappa_2 \|g_k\| + \|g_k\| \\ &\le \left(L_g t_k^\alpha \kappa_2 + 1\right) \|g_k\|. \end{aligned}$$

Thus

$$f(x_k + t_k^\alpha d_k) - f(x_k + t_k^\beta d_k) \le \kappa_2 (t_k^\alpha - t_k^\beta)(L_g t_k^\alpha \kappa_2 + 1)\|g_k\|^2. \tag{3.11}$$

The definition of t_k^α and t_k^β in Lemma 3.2 then give that (3.7) and (3.8) both hold, and so

$$\begin{aligned} f\left(x_k + t_k^\alpha d_k\right) - f\left(x_k + t_k^\beta d_k\right) &= \alpha t_k^\alpha \langle g_k, d_k\rangle - \beta t_k^\beta \langle g_k, d_k\rangle \\ &= \left(\alpha t_k^\alpha - \beta t_k^\beta\right) \langle g_k, d_k\rangle \\ &\ge \left(\beta t_k^\beta - \alpha t_k^\alpha\right) \kappa_1 \|g_k\|^2, \end{aligned} \tag{3.12}$$

where we again used the Cauchy-Schwarz inequality and (2.2) to deduce the last inequality. From (3.11), we now deduce that

$$\left(\beta t_k^\beta - \alpha t_k^\alpha\right) \kappa_1 \le \kappa_2 \left(t_k^\alpha - t_k^\beta\right) \left(L_g t_k^\alpha \kappa_2 + 1\right). \tag{3.13}$$

This inequality is equivalent to

$$\kappa_2^2 L_g \left(t_k^\alpha\right)^2 + \left(\kappa_2 + \alpha\kappa_1 - \kappa_2^2 L_g t_k^\beta\right) t_k^\alpha - \left(\kappa_2 + \beta\kappa_1\right) t_k^\beta \ge 0, \tag{3.14}$$

and so, since $t_k^\alpha > 0$, we deduce that

$$t_k^\alpha \ge \frac{\kappa_2^2 L_g t_k^\beta - \kappa_2 - \alpha\kappa_1 + \sqrt{\left(\kappa_2 + \alpha\kappa_1 - \kappa_2^2 L_g t_k^\beta\right)^2 + 4\left(\kappa_2 + \beta\kappa_1\right)\kappa_2^2 L_g t_k^\beta}}{2\kappa_2^2 L_g}, \tag{3.15}$$

and therefore that

$$\begin{aligned} &2\kappa_2^2 L_g \left(t_k^\alpha - t_k^\beta\right) \\ &\quad \ge -\left(\kappa_2 + \alpha\kappa_1 + \kappa_2^2 L_g t_k^\beta\right) + \sqrt{\left(\kappa_2 + \alpha\kappa_1 - \kappa_2^2 L_g t_k^\beta\right)^2 + 4\left(\kappa_2 + \beta\kappa_1\right)\kappa_2^2 L_g t_k^\beta} \end{aligned}$$

$$= \frac{-\left(\kappa_2+\alpha\kappa_1+\kappa_2^2L_gt_k^\beta\right)^2+\left(\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta\right)^2+4\left(\kappa_2+\beta\kappa_1\right)\kappa_2^2L_gt_k^\beta}{\kappa_2+\alpha\kappa_1+\kappa_2^2L_gt_k^\beta+\sqrt{\left(\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta\right)^2+4\left(\kappa_2+\beta\kappa_1\right)\kappa_2^2L_gt_k^\beta}}$$

$$= \frac{4\left(\beta-\alpha\right)\kappa_1\kappa_2^2L_gt_k^\beta}{\kappa_2+\alpha\kappa_1+\kappa_2^2L_gt_k^\beta+\sqrt{\left(\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta\right)^2+4\left(\kappa_2+\beta\kappa_1\right)\kappa_2^2L_gt_k^\beta}}. \tag{3.16}$$

As a consequence, we obtain that

$$\frac{\left(t_k^\alpha-t_k^\beta\right)}{2\left(\beta-\alpha\right)\kappa_1} \geq \frac{t_k^\beta}{\kappa_2+\alpha\kappa_1+\kappa_2^2L_gt_k^\beta+\sqrt{\left(\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta\right)^2+4\left(\kappa_2+\beta\kappa_1\right)\kappa_2^2L_gt_k^\beta}} \overset{\text{def}}{=} E\left(t_k^\beta\right). \tag{3.17}$$

Defining $S(t_k^\beta)\overset{\text{def}}{=}\sqrt{\left(\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta\right)^2+4\left(\kappa_2+\beta\kappa_1\right)\kappa_2^2L_gt_k^\beta}$, differentiating $E\left(t_k^\beta\right)$ with respect to t_k^β then gives that

$$\begin{aligned}
&E'\left(t_k^\beta\right)\\
&= \frac{\kappa_2+\alpha\kappa_1+\kappa_2^2L_gt_k^\beta+S\left(t_k^\beta\right)-t_k^\beta\left[\kappa_2^2L_g+\frac{-\left(\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta\right)\kappa_2^2L_g+2(\kappa_2+\beta\kappa_1)\kappa_2^2L_g}{S\left(t_k^\beta\right)}\right]}{\left[\kappa_2+\alpha\kappa_1+\kappa_2^2L_gt_k^\beta+S\left(t_k^\beta\right)\right]^2}\\
&= \frac{(\kappa_2+\alpha\kappa_1)\,S\left(t_k^\beta\right)+\left(\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta\right)^2+\left(\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta\right)\kappa_2^2L_gt_k^\beta+2\left(\kappa_2+\beta\kappa_1\right)\kappa_2^2L_gt_k^\beta}{\left[\kappa_2+\alpha\kappa_1+\kappa_2^2L_gt_k^\beta+S\left(t_k^\beta\right)\right]^2S\left(t_k^\beta\right)}\\
&= \frac{(\kappa_2+\alpha\kappa_1)\left[S\left(t_k^\beta\right)+\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta\right]+2\left(\kappa_2+\beta\kappa_1\right)\kappa_2^2L_gt_k^\beta}{\left[\kappa_2+\alpha\kappa_1+\kappa_2^2L_gt_k^\beta+S\left(t_k^\beta\right)\right]^2S\left(t_k^\beta\right)}.
\end{aligned} \tag{3.18}$$

It then follows that

$$E'\left(t_k^\beta\right)>0 \quad \text{for all } t_k^\beta>0$$

since

$$S\left(t_k^\beta\right)+\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta > \left|\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta\right|+\kappa_2+\alpha\kappa_1-\kappa_2^2L_gt_k^\beta \geq 0$$

and each constant and variable in $E'\left(t_k^\beta\right)$ is positive. Thus $E\left(t_k^\beta\right)$ is increasing as a function of t_k^β, and we obtain, because of Lemma 3.1 and the fact that (2.4) holds at t_k^β by construction, that

$$E\left(t_k^\beta\right) \geq E\left(\frac{2\left(1-\beta\right)\kappa_1}{L_g\kappa_2^2}\right),$$

and we finally deduce from (3.18) that

$$t_k^\alpha - t_k^\beta \geq 2(\beta-\alpha)\kappa_1\, E\left(\frac{2(1-\beta)\kappa_1}{L_g\kappa_2^2}\right).$$

Combining this with (3.10), we deduce that (3.9) holds with

$$\kappa_{\rm int} \overset{\rm def}{=} 2(\beta-\alpha)\kappa_1 \min\left[\frac{(1-\beta)}{\alpha L_g\kappa_2^2}, E\left(\frac{2(1-\beta)\kappa_1}{L_g\kappa_2^2}\right)\right],$$

where this lower bound only depends on α, β, κ_1, κ_2 and L_g, as desired. □

We now turn to estimating the worst-case evaluation complexity of Algorithm 2.1 for the task of finding an ϵ-first-order critical point.

THEOREM 3.4 *Suppose that* AF0–AF2 *hold. Then there exists a constant* $\kappa_{\rm GNL} > 0$ *such that, for any* $\epsilon \in (0,1)$*, Algorithm* 2.1 *needs at most*

$$k_{\max} \overset{\rm def}{=} \left\lfloor \frac{\kappa_{\rm GNL}(f(x_0)-\kappa_{\rm lbf})}{\epsilon^2} + M \right\rfloor \tag{3.19}$$

iterations to produce an iterate x_k *such that* $\|g_k\| \leq \epsilon$*, where* $\kappa_{\rm lbf}$ *and* M *are defined in* AF0 *and* (2.3)–(2.4)*, respectively, and where*

$$\kappa_{\rm GNL} \overset{\rm def}{=} \frac{(M+1)}{\alpha\kappa_1}\max\left[\frac{L_g\kappa_2^2\max[n_1,n_2]}{2(1-\beta)\kappa_1}, \frac{2(n_2+1)}{t_{\rm ini}}\right]$$

with

$$n_1 \overset{\rm def}{=} \left|\log_2\left(\frac{(1-\beta)\kappa_1}{t_{\rm ini}L_g\kappa_2^2}\right)\right| \quad\text{and}\quad n_2 \overset{\rm def}{=} \left|\log_2\left(\frac{\kappa_{\rm int}}{t_{\rm ini}}\right)\right|.$$

Proof The proof proceeds by first establishing the minimum achieved decrease in the objective function between iterate x_{k+1} and its ‘predecessor’ $x_{\pi(k+1)}$, where

$$\pi(k+1) = k - \arg\max_{0\leq j\leq M} f(x_{k-j}) \tag{3.20}$$

when using Algorithm 2.1.

- Assume first that both (2.3) and (2.4) hold for $t_k = t_{\rm ini}$ (in Step 3). Then we obtain a decrease

$$f(x_{\pi(k+1)}) - f(x_{k+1}) \geq -\alpha\, t_{\rm ini}\langle g_k, d_k\rangle \geq \alpha\, t_{\rm ini}\kappa_1\|g_k\|^2, \tag{3.21}$$

because of (2.3) and (2.2), and this decrease is obtained for a single additional function evaluation.
- Assume now that (2.3) fails at Step 3.2, and Step 4 is therefore entered. Assume furthermore that $j_3 \geq 1$ backtracking steps are performed in Step 4.1. The j_3 is the smallest non-negative integer such that (2.3) holds for $t_k = t_{\rm ini}2^{-j_3}$, which means that j_3 is the largest integer for which this inequality is violated for $t = t_{\rm ini}2^{-j_3+1}$.

Because $\alpha < \beta$, we deduce that (2.4) must hold for this value of t_k. Using now Lemma 3.1, we obtain that

$$t = 2^{-j_3+1} t_{\text{ini}} \geq \frac{2(1-\beta)\kappa_1}{L_g \kappa_2^2},$$

which in turn implies that

$$j_3 \leq \left| \log_2 \left(\frac{(1-\beta)\kappa_1}{t_{\text{ini}} L_g \kappa_2^2} \right) \right| \overset{\text{def}}{=} n_1. \tag{3.22}$$

Step 4 therefore requires at most n_1 function evaluations. If the linesearch is terminated in Step 4.2 (i.e. branching occurs to Step 7), we obtain a decrease

$$f(x_{\pi(k+1)}) - f(x_{k+1}) \geq -\alpha\, t_k \langle g_k, d_k \rangle \geq \alpha \frac{2(1-\beta)\kappa_1}{L_g \kappa_2^2} \kappa_1 \|g_k\|^2, \tag{3.23}$$

where we used (2.3), (2.4), the Cauchy-Schwarz inequality, (2.2) and Lemma 3.1 successively.

- If the linesearch is not terminated in Step 4, Step 6 must be entered, with a bracket $[t_{\text{low}}, t_{\text{up}}]$ where

$$t_{\text{low}} = 2^{-j_3} t_{\text{ini}} = \frac{1}{2} t_{\text{up}}.$$

Thus

$$t_{\text{up}} - t_{\text{low}} = \frac{1}{2} 2^{-j_3+1} t_{\text{ini}} = 2^{-j_3} t_{\text{ini}}.$$

We know from Lemma 3.3 that the length of the admissible interval is at least equal to $\kappa_{\text{int}} > 0$, where this constant only depends on α, β and L_g. Thus, the number $j_4 \geq 1$ of bisection (and function evaluations) within Step 6 is bounded above by the smallest integer such that

$$2^{-j_4}(t_{\text{up}} - t_{\text{low}}) = 2^{-j_4} 2^{-j_3} t_{\text{ini}} \geq \kappa_{\text{int}},$$

which then yields that the total number of function evaluations in Step 4 and 6 is bounded by

$$j_3 + j_4 \leq \left| \log_2 \left(\frac{\kappa_{\text{int}}}{t_{\text{ini}}} \right) \right| \overset{\text{def}}{=} n_2.$$

If we know compute the decrease obtained, we deduce, again from (2.3), (2.4) and Lemma 3.1, that (3.23) also holds in this case.

- Assume now that (2.4) fails in Step 3.3, and thus that Step 5 is entered. Assume furthermore that $j_2 \geq 1$ doubling of t_k (and j_2 function evaluations) occur in Step 5.1 (we know that j_2 is finite because of (3.4)). If the lineasearch is terminated in Step 5.2 (i.e. branching to Step 7 occurs), we obtain that the function decrease obtained is bounded below by

$$f(x_{\pi(k+1)}) - f(x_{k+1}) \geq -\alpha t_k \langle g_k, d_k \rangle \geq \alpha 2^{j_2} t_{\text{ini}} \kappa_1 \|g_k\|^2.$$

- The final case is when Step 6 is entered after Step 5, in which case the initial bracket for Step 6 is given by $[t_{\text{low}}, t_{\text{up}}]$ where

$$t_{\text{low}} = 2^{j_2-1} t_{\text{ini}} = \frac{1}{2} t_{\text{up}}.$$

Thus

$$t_{\text{up}} - t_{\text{low}} = \frac{1}{2} 2^{j_2} t_{\text{ini}} = 2^{j_2 - 1} t_{\text{ini}}.$$

Just as in the case where Step 6 is entered after Step 4, we now deduce that the number j_4 of bisections and function evaluations needed to reduce this bracket to the minimum possible value κ_{int} is limited by the inequality

$$2^{-j_4} 2^{j_2 - 1} t_{\text{ini}} = 2^{-j_4}(t_{\text{up}} - t_{\text{low}}) \geq \kappa_{\text{int}},$$

yielding a maximum number of bisection (and function evaluation) in Step 6 bounded by

$$j_4 \leq j_2 - 1 + \left| \log_2 \left(\frac{\kappa_{\text{int}}}{t_{\text{ini}}} \right) \right| = j_2 - 1 + n_2 \leq j_2(n_2 + 1).$$

In this final case, since $t_k \geq t_{\text{low}} = 2^{j_2 - 1} t_{\text{ini}}$ and (2.3) holds at t_k, the function decrease is bounded below by

$$f(x_{\pi(k+1)}) - f(x_{k+1}) \geq -\alpha t_k \langle g_k, d_k \rangle \geq \alpha 2^{j_2 - 1} t_{\text{ini}} \kappa_1 \|g_k\|^2.$$

Gathering all cases together, we see that function decrease per function evaluation is given, in the worst case, by

$$\min \left[\frac{t_{\text{ini}}}{1}, \frac{2(1-\beta)\kappa_1}{L_g \kappa_2^2 n_1}, \frac{2(1-\beta)\kappa_1}{L_g \kappa_2^2 n_2}, \frac{2^{j_2} t_{\text{ini}}}{j_2}, \frac{2^{j_2 - 1} t_{\text{ini}}}{2 j_2 (n_2 + 1)} \right] \alpha \kappa_1 \|g_k\|^2, \tag{3.24}$$

where, by construction, n_1 and n_2 only depend on α, β, κ_1, κ_2, L_g and t_{ini}. Noting that, for $j_2 \geq 1$,

$$\frac{2^{j_2}}{j_2} \geq 2 \quad \text{and} \quad \frac{2^{j_2 - 1}}{2 j_2} \geq \frac{1}{2}$$

and defining

$$\kappa_{\text{decr}} \stackrel{\text{def}}{=} \alpha \kappa_1 \min \left[\frac{2(1-\beta)\kappa_1}{L_g \kappa_2^2 \max[n_1, n_2]}, \frac{t_{\text{ini}}}{2(n_2 + 1)} \right],$$

we therefore deduce from (3.24) that, as long as the algorithm does not terminate (i.e. as long as $\|g_k\| \geq \epsilon$)

$$f(x_{\pi(k+1)}) - f(x_{k+1}) \geq \kappa_{\text{decr}} \|g_k\|^2 \geq \kappa_{\text{decr}} \epsilon^2.$$

Tracing back the predecessors of iterate x_{k+1} up to x_0 and denoting the composition of j instances of the predecessor operator $\pi(\cdot)$ by $\pi^j(\cdot)$, we also deduce that

$$f(x_{\pi^{j+1}(k+1)}) - f(x_{\pi^j(k+1)}) \geq \kappa_{\text{decr}} \epsilon^2 \qquad (j = 0, \ldots, p_k), \tag{3.25}$$

for p_k such that $x_{\pi^{p_k}(k+1)} = x_0$ and where, by convention, $\pi^0(k+1) \stackrel{\text{def}}{=} k + 1$. Now the definition of $\pi(\cdot)$ in (3.20) implies that, for all ℓ,

$$0 \leq \ell + 1 - \pi(\ell + 1) \leq M + 1, \tag{3.26}$$

and we have, using AF0, that

$$f(x_0) - \kappa_{\text{lbf}} \geq f(x_0) - f(x_{k+1}) = \sum_{j=0}^{p_k} [f(x_{\pi^{j+1}(k+1)}) - f(x_{\pi^j(k+1)})]. \tag{3.27}$$

Using (3.26), we obtain that the sum in the right-side of this expression contains at least

$$\left\lfloor \frac{k+1}{M+1} \right\rfloor$$

terms. Substituting then (3.25) for each term, (3.27) gives that

$$f(x_0) - \kappa_{\text{lbf}} \geq \left\lfloor \frac{k+1}{M+1} \right\rfloor \kappa_{\text{decr}} \epsilon^2 \geq \left(\frac{k+1}{M+1} - 1 \right) \kappa_{\text{decr}} \epsilon^2 = \frac{k-M}{M+1} \kappa_{\text{decr}} \epsilon^2.$$

As a consequence, we obtain that the total number of function evaluations in Algorithm 2.1 is bounded above by

$$\left\lfloor \frac{(M+1)(f(x_0) - \kappa_{\text{lbf}})}{\kappa_{\text{decr}} \epsilon^2} + M \right\rfloor$$

yielding the desired conclusion with $\kappa_{\text{GNL}} = (M+1)/\kappa_{\text{decr}}$. □

We conclude this analysis by an interesting consequence of the proof of this theorem. Assume now that the algorithm is not terminated in Step 1 when the condition $\|g_k\| \leq \epsilon$ occurs, but that an infinite sequence of iterates is computed. Consider iteration ℓ, for some $\ell \geq k_{\max} + M + 1$. Using (3.27), we see that

$$f(x_0) - \kappa_{\text{lbf}} \geq \sum_{j=0}^{p_\ell} [f(x_{\pi^{j+1}(\ell+1)}) - f(x_{\pi^j(\ell+1)})]$$

where p_ℓ is now defined by $x_{\pi^{p_\ell}(\ell+1)} = x_0$, and the sum on the right-hand side now contains at least

$$\left\lfloor \frac{\ell+1}{M+1} \right\rfloor$$

terms, of which the first $\left\lfloor \frac{k_{\max}+1}{M+1} \right\rfloor$ are such that $\|g_{\pi^j(\ell+1)}\| \geq \epsilon$ because $\pi^j(\ell+1) \leq k_{\max}$ for $p_\ell - j + 1 < \left\lfloor \frac{k_{\max}+1}{M+1} \right\rfloor$. If we now assume that $\|g_\ell\| \geq \epsilon$, (3.25) implies that $f(x_{\pi(\ell+1)}) - f(x_{\ell+1}) \geq \kappa_{\text{decr}} \epsilon^2$, and thus, from (3.19),

$$\begin{aligned} f(x_0) - \kappa_{\text{lbf}} &\geq \left(\left\lfloor \frac{k_{\max}+1}{M+1} \right\rfloor + 1 \right) \kappa_{\text{decr}} \epsilon^2 \\ &= \left\lfloor \frac{f(x_0) - \kappa_{\text{lbf}}}{\kappa_{\text{decr}} \epsilon^2} \right\rfloor \kappa_{\text{decr}} \epsilon^2 + 2\kappa_{\text{decr}} \epsilon^2 > f(x_0) - \kappa_{\text{lbf}} + \kappa_{\text{decr}} \epsilon^2 \end{aligned}$$

which is impossible. Hence we obtain that, for any $\epsilon \in (0, 1)$ there exists a $k_{\max}$ such that for all $\ell \geq k_{\max} + M + 1$, one has that $\|g_\ell\| < \epsilon$. In other words,

$$\lim_{k \to \infty} \|g_k\| = 0.$$

This confirms, from another perspective, the global convergence result obtained by Grippo et al. [10,11].

4. Conclusions and perspectives

We have shown that gradient-related methods using a non-monotone (and monotone) linesearch will find an ϵ-approximate first-order critical point of a smooth function with Lipschitz gradient in $O(\epsilon^{-2})$ function and gradient evaluations at most. Their worst-case behaviour is therefore, up to a factor, equivalent to that of a simple monotone pure steepest-descent algorithm, albeit their practical performance is often superior [12]. Moreover, it results from Cartis et al. [14], that this bound is sharp.

In the same line of investigation, Cartis et al. [15] show that the same complexity order is obtained for the steepest-descent method with exact linesearch and that it is sharp. One may expect that this result can be extended to the gradient-related algorithms analysed in the present note, although the construction of an example illustrating the sharpness of the complexity bound is likely to be challenging without monotonicity.

Note

1. ‘lbf’ stands for ‘lower bound on the objective function’.

References

[1] Nesterov Yu. Introductory lectures on convex optimization. Applied Optimization. Dordrecht: Kluwer Academic Publishers; 2004.

[2] Gratton S, Sartenaer A, Toint PhL. Recursive trust-region methods for multiscale nonlinear optimization. SIAM J. Optim. 2008;19:414–444.

[3] Nesterov Yu, Polyak BT. Cubic regularization of Newton method and its global performance. Math. Prog. Ser. A. 2006;108:177–205.

[4] Cartis C, Gould NIM, Toint PhL. Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case function-evaluation complexity. Math. Prog. Ser. A. 2011;130:295–319.

[5] Cartis C, Gould NIM, Toint PhL. An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity. IMA J. Numer. Anal. 2012;32:1662–1645.

[6] Cartis C, Gould NIM, Toint PhL. On the complexity of finding first-order critical points in constrained nonlinear optimization. Math. Prog. Ser. A. 2012 (online). doi: 10.1007/s10107-012-0617-9.

[7] Cartis C, Gould NIM, Toint PhL. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. SIAM J. Optim. 2011;21:1721–1739.

[8] Ortega JM, Rheinboldt WC. Iterative solution of nonlinear equations in several variables. London: Academic Press; 1970.

[9] Nocedal J, Wright SJ. Numerical optimization. Series in operations research. Heidelberg: Springer Verlag; 1999.

[10] Grippo L, Lampariello F, Lucidi S. A nonmonotone line search technique for Newton’s method. SIAM J. Numer. Anal. 1986;23:707–716.

[11] Grippo L, Lampariello F, Lucidi S. A truncated Newton method with nonmonotone line search for unconstrained optimization. J. Optim. Theory Appl. 1989;60:401–419.

[12] Toint PhL. An assessment of non-monotone linesearch techniques for unconstrained optimization. SIAM J. Sci. Stat. Comp. 1996;17:725–739.

[13] Bertsekas DP. Nonlinear Programming. Belmont (MA): Athena Scientific; 2008.

[14] Cartis C, Gould NIM, Toint PhL. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization. SIAM J. Optim. 2010;20:2833–2852.

[15] Cartis C, Gould NIM, Toint PhL. On the complexity of the steepest-descent with exact linesearches. Technical Report naXys-16-2012, Namur Centre for Complex Systems (naXys). Belgium: FUNDP-University of Namur, Namur; 2012.