SIAM J. OPTIM.
Vol. 20, No. 6, pp. 2833–2852

ON THE COMPLEXITY OF STEEPEST DESCENT, NEWTON'S AND REGULARIZED NEWTON'S METHODS FOR NONCONVEX UNCONSTRAINED OPTIMIZATION PROBLEMS*

C. CARTIS†, N. I. M. GOULD‡, AND PH. L. TOINT§

Abstract. It is shown that the steepest-descent and Newton's methods for unconstrained nonconvex optimization under standard assumptions may both require a number of iterations and function evaluations arbitrarily close to $O(\epsilon^{-2})$ to drive the norm of the gradient below ϵ. This shows that the upper bound of $O(\epsilon^{-2})$ evaluations known for the steepest descent is tight and that Newton's method may be as slow as the steepest-descent method in the worst case. The improved evaluation complexity bound of $O(\epsilon^{-3/2})$ evaluations known for cubically regularized Newton's methods is also shown to be tight.

Key words. nonlinear optimization, unconstrained optimization, steepest-descent method, Newton's method, trust-region methods, cubic regularization, global complexity bounds, global rate of convergence

AMS subject classifications. 90C30, 65K05, 49M37, 49M15, 49M05, 58C15, 90C60, 68Q25

DOI. 10.1137/090774100

1. Introduction.

We consider the numerical solution of the unconstrained (possibly nonconvex) optimization problem

$$\min_x f(x), \tag{1.1}$$

where we assume that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and bounded below. All practical methods for the solution of (1.1) are iterative and generate a sequence $\{x_k\}$ of iterates approximating a local minimizer of f. A variety of algorithms of this form exist, among which the steepest-descent and Newton's methods are preeminent.

At iteration k, the steepest-descent method chooses the new iterate x_{k+1} by minimizing (typically inexactly) $f(x_k - tg_k)$, for $t \geq 0$, where $g_k = \nabla_x f(x_k)$. This first-order method has the merit of simplicity and a theoretical guarantee of convergence under weak conditions (see Dennis and Schnabel (1983), for instance). The number of iterations required in the worst case to generate an iterate x_k such that $\|g_k\| \leq \epsilon$ (for $\epsilon > 0$ arbitrarily small) is known to be at most $O(\epsilon^{-2})$ (see Nesterov (2004), page 29), but the question of whether this latter bound is tight has remained open. The practical

*Received by the editors October 19, 2009; accepted for publication (in revised form) June 16, 2010; published electronically September 9, 2010.
http://www.siam.org/journals/siopt/20-6/77410.html

†School of Mathematics, The King's Buildings, University of Edinburgh, Edinburgh, EH9 3JZ, Scotland, UK (coralia.cartis@ed.ac.uk). This author is grateful to the Royal Society for its support through the International Joint Project 14265.

‡Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, England, UK (nick.gould@sftc.ac.uk). The work of this author was supported by EPSRC grant EP/E053351/1.

§Department of Mathematics, FUNDP—University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium (philippe.toint@fundp.ac.be). This author gratefully acknowledges the support of EPSRC grant EP/G038643/1, and of the ADTAO project funded by the Sciences et Technologies pour l'Aéronautique et l'Espace (STAE) Foundation (Toulouse, France) within the Réseau Thématique de Recherche Avancée (RTRA).

behavior of the steepest-descent method may be poor on ill-conditioned problems, and it is not often used for solving general unconstrained optimization problems.

By contrast, Newton's method and its variants are popular and effective. At iteration k, this method (in its simplest and standard form) chooses the next iterate by minimizing the quadratic model

$$m_k(x_k + s) = f(x_k) + g_k^T s + \tfrac{1}{2} s^T H_k s, \tag{1.2}$$

where $H_k \stackrel{\text{def}}{=} \nabla_{xx} f(x_k)$ is assumed to be positive definite. This algorithm is known to converge locally and quadratically to strict local minimizers of the objective function f, but, in general, convergence from arbitrary starting points cannot be guaranteed, in particular, because the Hessian H_k may be singular or indefinite, making the minimization of the quadratic model (1.2) irrelevant. However, Newton's method works surprisingly often without this guarantee and, when it does, is usually remarkably effective. We again refer the reader to classics in optimization like Dennis and Schnabel (1983) and Nocedal and Wright (1999) for a more extensive discussion of this method. To the best of our knowledge, no worst-case analysis is available for this standard algorithm applied on possibly nonconvex problems (a complexity analysis is, however, available for the case where the objective function is convex; see Nesterov (2004), for instance).

Globally convergent variants of Newton's method have been known and used for a long time in the linesearch, trust-region, or filter frameworks, descriptions of which may be found in Dennis and Schnabel (1983), Conn, Gould, and Toint (2000), and Gould, Sainvitu, and Toint (2005), respectively. Although theoretically convergent and effective in practice, the complexity of most of these variants applied on general nonconvex problems has not yet been investigated. The authors are only aware of the analysis by Gratton, Sartenaer, and Toint (2008, Corollary 4.10), where a bound on the complexity of an inexact variant of the trust-region method is shown to be of the same order as that of steepest descent, and of the analysis by Ueda and Yamashita (2009), (2010) and Ueda (2009), which essentially proves the same result for a variant of Newton's method using Levenberg–Morrison–Marquardt regularization. Interestingly, the result by Gratton, Sartenaer, and Toint (2008) does not assume accurate minimization of the model m_k but only relies on the standard but weaker Cauchy condition (see Assumption AA.1, page 131, in Conn, Gould, and Toint (2000)).

Another particular globally convergent variant of Newton's method for the solution of nonconvex unconstrained problems of the form (1.1) is of special interest, because it is covered by a better worst-case complexity analysis. Independently proposed by Griewank (1981), Weiser, Deuflhard, and Erdmann (2007), and Nesterov and Polyak (2006) and subsequently adapted in Cartis, Gould, and Toint (2009), this method uses a cubic regularization of the quadratic model (1.2) in that the new iterate is found at iteration k by globally minimizing the cubic model

$$m_k(x_k + s) = f(x_k) + g_k^T s + \tfrac{1}{2} s^T H_k s + \tfrac{1}{3}\sigma_k \|s\|^3, \tag{1.3}$$

where $\sigma_k \geq 0$ is a suitably chosen regularization parameter (the various cited authors differ in how this choice is made). This method, which we call the adaptive regularization with cubics (ARC) algorithm, has been shown to require at most $O(\epsilon^{-3/2})$ iterations to produce an iterate x_k such that $\|g_k\| \leq \epsilon$, provided the objective function is twice continuously differentiable, bounded below, and provided $\nabla_{xx} f(x)$ is globally Lipschitz continuous on each segment $[x_k, x_{k+1}]$ of the piecewise linear path defined by the iterates. This result, due to Nesterov and Polyak (2006) when the model minimization is global and exact and to Cartis, Gould, and Toint (2010a) for the case where this minimization is only performed locally and approximately, is obviously considerably better than that for the steepest-descent method. We note here that even better complexity results *in the convex case* are discussed for ARC by Nesterov (2008) and Cartis, Gould, and Toint (2010b) and for other regularized Newton's methods by Polyak (2009) and Ueda (2009).

But obvious questions remain. For one, whether the steepest-descent method may actually require $O(\epsilon^{-2})$ functions evaluations on functions with Lipschitz continuous gradients is of interest. The first purpose of this paper is to show that this is so. The lack of complexity analysis for the standard Newton's method also raises the possibility that, despite its considerably better performance on problems met in practice, its worst-case behavior could be as slow as that of steepest descent. A second objective of this paper is to show that this is the case, even if the objective function is assumed to be bounded below and twice continuously differentiable with Lipschitz continuous Hessian on each segment of the piecewise linear path defined by the iterates. This establishes a clear distinction between Newton's method and its ARC variant, for which a substantially more favorable analysis exists. The question then immediately arises to decide whether this better bound for ARC is actually the best that can be achieved. The third aim of the paper is to demonstrate that it is indeed the best.

The paper is organized as follows. Section 2 introduces an example for which the steepest-descent method is as slow as its worst-case analysis suggests. Section 3 then exploits the technique of section 2 for constructing examples for which slow convergence of Newton's method can be shown; while section 4 further discusses the implications of these examples (and the interpretation of worst-case complexity bounds in general). Section 5 then again exploits the same technique for constructing an example where the ARC algorithm is as slow as is implied by the aforementioned complexity analysis. Finally, some conclusions are drawn in section 6.

2. Slow convergence of the steepest-descent method. Consider using the steepest-descent method for solving (1.1). We would like to construct an example on which this algorithm converges at a rate which corresponds to its worst case on general nonconvex objective functions, i.e., such that one has to perform $O(\epsilon^{-2})$ iterations to ensure that

$$\|g_{k+1}\| \leq \epsilon. \tag{2.1}$$

This property is obviously obtained if the sequence of gradients satisfies the (monotonically decreasing) lower bound

$$\|g_k\| \geq \left(\frac{1}{k+1}\right)^{\frac{1}{2}} \tag{2.2}$$

for all $k \geq 0$. An arbitrarily close approximation can also be considered by requiring that, for any $\tau > 0$, the steepest-descent method needs $O(\epsilon^{-2+\tau})$ iterations to achieve (2.1), which leads to the alternative condition that, for all $k \geq 0$,

$$\|g_k\| \geq \left(\frac{1}{k+1}\right)^{\frac{1}{2-\tau}}. \tag{2.3}$$

Our objective is therefore to construct sequences $\{x_k\}$, $\{g_k\}$, $\{H_k\}$, and $\{f_k\}$ such that (2.3) holds and which may be generated by the steepest-descent algorithm together with a twice continuously differentiable function $f^{(1)}(x)$ such that

$$f_k = f^{(1)}(x_k) \text{ and } g_k = \nabla_x f^{(1)}(x_k) \tag{2.4}$$

for all $k \geq 0$. In addition, $f^{(1)}$ must be bounded below and H_k must be positive definite for the algorithm to be well-defined. We also would like $f^{(1)}$ to be as smooth as possible; we are aiming at

AS.0 f is twice continuously differentiable, bounded below, and has bounded Lipschitz continuous gradient,

since these are the standard assumptions under which globalized steepest descent is provably convergent (see Dennis and Schnabel (1983, Theorem 6.3.3)).

Our example is unidimensional, and we define a sequence of iterates $\{x_k\}$ tending to infinity by

$$x_0 = 0, \qquad x_{k+1} = x_k + \alpha_k \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \quad (k > 0) \tag{2.5}$$

for some steplength $\alpha_k > 0$ such that, for constant $\underline{\alpha}$ and $\overline{\alpha}$,

$$0 < \underline{\alpha} \leq \alpha_k \leq \overline{\alpha} < 2, \tag{2.6}$$

giving the step

$$s_k \stackrel{\text{def}}{=} x_{k+1} - x_k = \alpha_k \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta}. \tag{2.7}$$

We also set

$$f_0 = \frac{1}{2}\zeta(1+2\eta), \quad f_{k+1} = f_k - \alpha_k(1 - \tfrac{1}{2}\alpha_k)\left(\frac{1}{k+1}\right)^{1+2\eta}, \tag{2.8}$$

$$g_k = -\left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta}, \text{ and } H_k = 1, \tag{2.9}$$

where

$$\eta = \eta(\tau) \stackrel{\text{def}}{=} \frac{1}{2-\tau} - \frac{1}{2} = \frac{\tau}{4-2\tau} > 0 \tag{2.10}$$

and

$$\zeta(t) \stackrel{\text{def}}{=} \sum_{k=1}^{\infty} \frac{1}{k^t} = \sum_{k=0}^{\infty} \left(\frac{1}{k+1}\right)^t$$

is the Riemann ζ function, which is finite for all $t > 1$ and thus for $t = 1 + 2\eta$. Immediately note that the first part of (2.9) gives (2.3) by construction. In what follow the choice of α_k is arbitrary in the interval $[\underline{\alpha}, \overline{\alpha}]$, but we observe that the selected value of α_k can be seen as resulting from a Goldstein–Armijo linesearch enforcing, for some $\alpha, \beta \in (0,1)$ with $\alpha < \beta$,

$$f(x_k) - f(x_{k+1}) \geq -\alpha s_k^T g_k = \alpha \alpha_k \|g_k\|^2$$

and

$$f(x_k) - f(x_{k+1}) \leq -\beta s_k^T g_k = \beta\alpha_k \|g_k\|^2,$$

since (2.8) ensures that $2(1-\alpha) < \alpha_k < 2(1-\beta)$ and thus that (2.6) holds.

We now exhibit a function $f^{(1)}(x)$ which satisfies AS.0 and (2.4)–(2.9) and whose definition on the nonnegative reals[1] is given by

$$f^{(1)}(x) = p_k(x - x_k) + f_{k+1} \text{ for } x \in [x_k, x_{k+1}] \text{ and } k \geq 0, \tag{2.11}$$

where p_k is a polynomial Hermite interpolant on $[0, s_k]$ of the form

$$p_k(t) \stackrel{\rm def}{=} c_{0,k} + c_{1,k}t + c_{2,k}t^2 + c_{3,k}t^3 + c_{4,k}t^4 + c_{5,k}t^5 \tag{2.12}$$

such that

$$p_k(0) = \alpha_k \left(1 - \frac{1}{2}\alpha_k\right) \left(\frac{1}{k+1}\right)^{1+2\eta}, \quad p_k(s_k) = 0, \tag{2.13}$$

$$p_k'(0) = -\left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta}, \quad p_k'(s_k) = -\left(\frac{1}{k+2}\right)^{\frac{1}{2}+\eta}. \tag{2.14}$$

We also impose that $p_k''(0) = p_k''(s_k) = 1$. These conditions immediately give that

$$c_{0,k} = \alpha_k \left(1 - \frac{1}{2}\alpha_k\right) \left(\frac{1}{k+1}\right)^{1+2\eta}, \quad c_{1,k} = -\left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta}, \text{ and } c_{2,k} = \frac{1}{2}.$$

One then verifies that the remaining interpolation conditions may be written in the form

$$\begin{pmatrix} s_k^3 & s_k^4 & s_k^5 \\ 3s_k^2 & 4s_k^3 & 5s_k^4 \\ 6s_k & 12s_k^2 & 20s_k^3 \end{pmatrix} \begin{pmatrix} c_3 \\ c_4 \\ c_5 \end{pmatrix} = \begin{pmatrix} 0 \\ p_k'(s_k) + (1-\alpha_k)\left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ 0 \end{pmatrix},$$

whose solution turns out to be

$$\begin{pmatrix} c_{3,k} \\ c_{4,k} \\ c_{5,k} \end{pmatrix} = \begin{pmatrix} -4\,\dfrac{\phi_k}{s_k} \\ 7\,\dfrac{\phi_k}{s_k^2} \\ -3\,\dfrac{\phi_k}{s_k^3} \end{pmatrix}, \tag{2.15}$$

where

$$\phi_k = \frac{1}{\alpha_k}(1 - \alpha_k - \psi_k) \text{ with } \psi_k \stackrel{\rm def}{=} \left(\frac{k+1}{k+2}\right)^{\frac{1}{2}+\eta}. \tag{2.16}$$

The definition of ψ_k implies that $|\psi_k| \in (0,1)$ for all $k \geq 0$. The graph of the resulting $f^{(1)}$ and its first three derivatives are given on the first 16 intervals and for $\eta = 10^{-4}$ and $\alpha_k = 1$ in Figure 2.1.

[1] It can be easily smoothly extended to the negative reals while maintaining its boundedness and the bounded nature of its second derivatives.

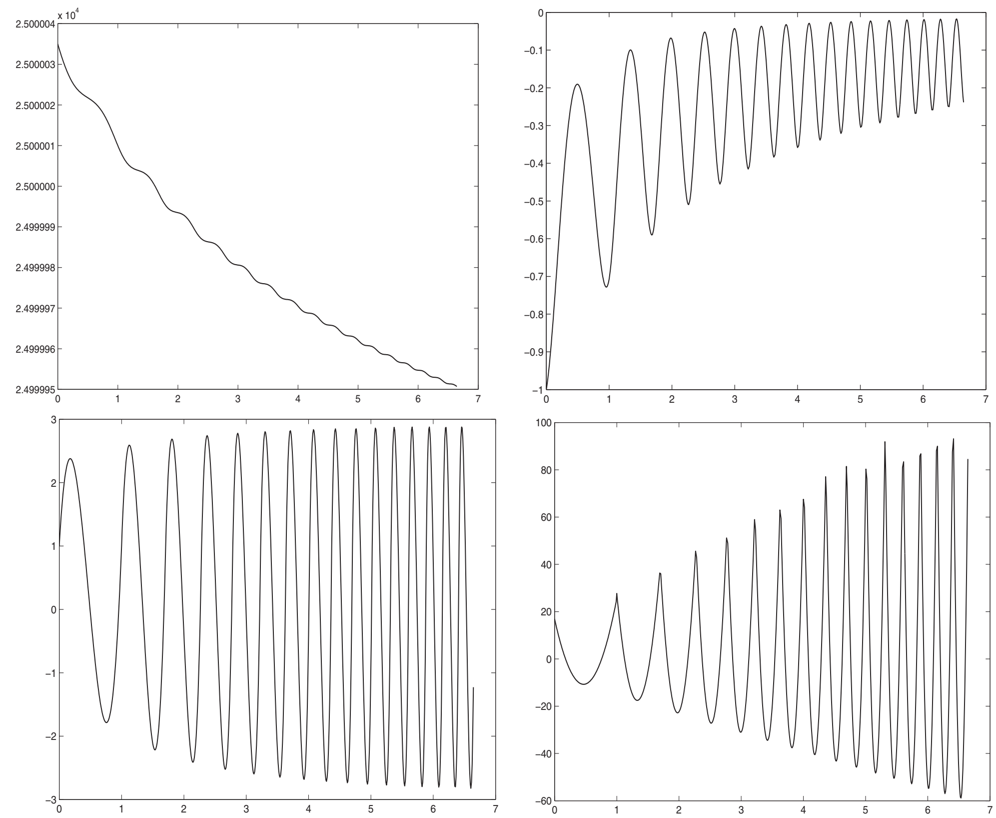

FIG. 2.1. *The function $f^{(1)}$ (top left) and its derivatives of order one (top right), two (bottom left), and three (bottom right) on the first* 16 *intervals.*

This figure confirms the properties inherited from the construction of the function $f^{(1)}$, namely, that it is twice continuously differentiable with bounded second derivatives. This last observation results from the bound

$$\begin{aligned} |p''(t)| &= 2c_{2,k} + 6c_{3,k}t + 12c_{4,k}t^2 + 20c_{5,k}t^3 \\ &\leq 2|c_{2,k}| + 6|c_{3,k}|s_k + 12|c_{4,k}|s_k^2 + 20|c_{5,k}|s_k^3 \\ &\leq 1 + 168|\phi_k| \\ &\leq 1 + 168\max[1,\overline{\alpha}]/\underline{\alpha} \end{aligned} \tag{2.17}$$

for all $k \geq 0$ and all $t \in [0, s_k]$, where we used (2.15) and the inequality $|\phi_k| \leq 1$. The gradient of $f^{(1)}$ is therefore Lipschitz continuous, but it is not the case for its second derivative, as can be seen in Figure 2.1 where one observes a linear increase in the third-derivative peaks with k. The fact that $f^{(1)}$ is bounded below by zero finally results from the bound

$$f_k - f_{k+1} = \alpha_k \left(1 - \frac{1}{2}\alpha_k\right)\left(\frac{1}{k+1}\right)^{1+2\eta} \leq \frac{1}{2}\left(\frac{1}{k+1}\right)^{1+2\eta},$$

which in turn yields, using (2.8) and the definition of the Riemann ζ function, that

$$f_{k+1} \geq f_0 - \frac{1}{2}\sum_{j=0}^{k}\left(\frac{1}{j+1}\right)^{1+2\eta} \geq \frac{1}{2}\zeta(1+2\eta) - \frac{1}{2}\zeta(1+2\eta) = 0$$

for all $k \geq 0$; note that, with the choice $\eta = 10^{-4}$ used for Figure 2.1, we obtain that $1 + 2\eta = 1.0002$ and $f_0 = (1/2)\zeta(1.0002) \approx 25000.3$.

This example thus implies that, for any $\tau > 0$, the steepest-descent method (with a Goldstein–Armijo linesearch) may require, for any $\epsilon \in (0,1)$, at least

$$\left\lfloor \frac{1}{\epsilon^{2-\tau}} \right\rfloor$$

iterations for producing an iterate x_k such that $\|g_k\| \leq \epsilon$. This bound is arbitrarily close to the upper bound of $O(\epsilon^{-2})$, which proves that this latter bound is essentially sharp.

3. Slow convergence of Newton's method.

3.1. Bounded second derivatives. Now consider using Newton's method for solving (1.1). We now would like to construct an example on which this algorithm converges at a rate which corresponds to the worst case known for the steepest-descent method on general nonconvex objective functions, i.e., such that one has to perform $O(\epsilon^{-2})$ iterations to ensure (2.1). As above, a suitable condition for achieving this goal is to require that (2.2) holds for all $k \geq 0$, and an arbitrarily close approximation can be considered by requiring that, for any $\tau > 0$, Newton's method needs $O(\epsilon^{-2+\tau})$ iterations to achieve (2.1), leading to the requirement that (2.3) holds for all $k \geq 0$. Our current objective is therefore to construct sequences $\{x_k\}$, $\{g_k\}$, $\{H_k\}$, and $\{f_k\}$ such that this latter condition holds and which may now be generated by Newton's algorithm together with a twice continuously differentiable function $f^{(2)}(x)$ such that, for all $k \geq 0$,

$$f_k = f^{(2)}(x_k), \quad g_k = \nabla_x f^{(2)}(x_k), \text{ and } H_k = \nabla_{xx} f^{(2)}(x_k). \tag{3.1}$$

In addition, $f^{(2)}$ must be bounded below and H_k must be positive definite for the algorithm to be well-defined. We also would like $f^{(2)}$ to be as smooth as possible; we are aiming at

AS.1 f is twice continuously differentiable, bounded below, and has bounded and Lipschitz continuous second derivatives along each segment $[x_k, x_{k+1}]$,

since these are the standard assumptions under which globalized Newton's method is provably convergent (see Dennis and Schnabel (1983, Theorem 6.3.3), Fletcher (1987, Theorem 2.5.1), or Nocedal and Wright (1999, Theorem 3.2)).

Our example is bidimensional and its iterates are tending to infinity. They are defined, for all $k \geq 0$, by

$$x_0 = (0,0)^T, \quad x_{k+1} = x_k + \begin{pmatrix} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ 1 \end{pmatrix}, \tag{3.2}$$

$$f_0 = \frac{1}{2}\left[\zeta(1+2\eta) + \zeta(2)\right], \quad f_{k+1} = f_k - \frac{1}{2}\left[\left(\frac{1}{k+1}\right)^{1+2\eta} + \left(\frac{1}{k+1}\right)^{2}\right], \tag{3.3}$$

$$g_k = -\begin{pmatrix} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ \left(\frac{1}{k+1}\right)^{2} \end{pmatrix}, \text{ and } H_k = \begin{pmatrix} 1 & 0 \\ 0 & \left(\frac{1}{k+1}\right)^{2} \end{pmatrix}, \tag{3.4}$$

where, as in (2.10), $\eta = \tau/(4-2\tau) > 0$ and $\zeta(t)$ is the Riemann ζ function. The first part of (3.4) then immediately gives (2.3) by construction, since the norm of that vector is at least equal to the absolute value of its first component.

We now verify that, provided (3.1) holds, the sequences given by (3.2)–(3.4) may be generated by Newton's method. Defining

$$s_k \stackrel{\text{def}}{=} x_{k+1} - x_k = \left(\begin{array}{c} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ 1 \end{array} \right) \stackrel{\text{def}}{=} \left(\begin{array}{c} \mu_k \\ 1 \end{array} \right) \tag{3.5}$$

and remembering (1.2), this amounts to verifying that

$$g_k + H_k s_k = 0, \tag{3.6}$$

$$H_k \text{ is positive definite,} \tag{3.7}$$

and

$$f_{k+1} = m_k(x_k + s_k) \tag{3.8}$$

for all $k \geq 1$. Note that, by definition, $\mu_k \in (0, 1]$. The first two of these conditions say that the quadratic model (1.2) is globally minimized exactly, while the third ensures perfect equality between the predicted and achieved decrease in the model and objective function, respectively. In our case, (3.6) becomes, using (3.5), (3.2), and (3.4),

$$g_k + H_k s_k = -\left(\begin{array}{c} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ \left(\frac{1}{k+1}\right)^2 \end{array} \right) + \left(\begin{array}{cc} 1 & 0 \\ 0 & \left(\frac{1}{k+1}\right)^2 \end{array} \right) \left(\begin{array}{c} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ 1 \end{array} \right) = 0,$$

as desired, while (3.7) also follows from (3.4). Using (3.3) and (3.4), we also obtain that

$$\begin{aligned} m_k(x_k + s_k) &= f(x_k) + g_k^T s_k + \frac{1}{2} s_k^T H_k s_k \\ &= f(x_k) - \frac{1}{2} \left(\frac{1}{k+1}\right)^{1+2\eta} - \frac{1}{2} \left(\frac{1}{k+1}\right)^2 \\ &= f(x_{k+1}), \end{aligned}$$

which in turn yields (3.8).

We now have to exhibit a function $f^{(2)}(x)$ which satisfies AS.1 and (3.1)–(3.4). The above equations suggest a function of the form

$$f^{(2)}(x) = f^{(2,1)}([x]_1) + f^{(2,2)}([x]_2),$$

where $[x]_i$ is the ith component of the vector x and where the univariates $f^{(2,1)}$ and $f^{(2,2)}$ are computed separately. Since our conditions involve, for both functions, fixed values of the function

$$f_{1,0} = \frac{1}{2}\zeta(1+2\eta), \quad f_{1,k+1} = f_{1,k} - \frac{1}{2}\left(\frac{1}{k+1}\right)^{1+2\eta}, \tag{3.9}$$

$$f_{2,0} = \frac{1}{2}\zeta(2), \quad f_{2,k+1} = f_{2,k} - \frac{1}{2}\left(\frac{1}{k+1}\right)^2, \tag{3.10}$$

and of its first and second derivatives at the endpoints of the interval $[x_k, x_{k+1}]$, we again consider applying polynomial Hermite interpolation on the interval $[0, x_{k+1} - x_k]$, which we will subsequently translate. Considering $f^{(2,1)}$ first, we note that it has to satisfy conditions that are identical to those stated for $f^{(1)}$ in section 2 for the case where $\alpha_k = 1$ for all k. We may then choose

$$f^{(2,1)}([x]_1) = f^{(1)}([x]_1)$$

for all $[x]_1$. Let us now consider $f^{(2,2)}$. Again, we consider a function defined by

$$f^{(2,2)}([x]_2) = q_k([x - x_k]_2) + f_{2,k+1} \ \text{ for } \ [x]_2 \in \big[[x_k]_2, [x_{k+1}]_2\big] \ \text{ and } \ k \geq 0,$$

where q_k is a polynomial Hermite interpolant on the interval $[0,1]$ of the form

$$q_k(t) \stackrel{\text{def}}{=} d_{0,k} + d_{1,k}t + d_{2,k}t^2 + d_{3,k}t^3 + d_{4,k}t^4 + d_{5,k}t^5$$

such that

$$q_k(0) = \frac{1}{2}\left(\frac{1}{k+1}\right)^2, \quad q_k(1) = 0,$$

$$q_k'(0) = -\left(\frac{1}{k+1}\right)^2, \quad q_k'(1) = -\left(\frac{1}{k+2}\right)^2,$$

$$q_k''(0) = \left(\frac{1}{k+1}\right)^2, \text{ and } q_k''(1) = \left(\frac{1}{k+2}\right)^2.$$

These conditions immediately give that

$$d_{0,k} = \frac{1}{2}\left(\frac{1}{k+1}\right)^2, \quad d_{1,k} = -\left(\frac{1}{k+1}\right)^2, \text{ and } d_{2,k} = \frac{1}{2}\left(\frac{1}{k+1}\right)^2.$$

Applying the same interpolation technique as above, one verifies that

$$\begin{pmatrix} d_{3,k} \\ d_{4,k} \\ d_{5,k} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 9\left(\frac{1}{k+2}\right)^2 - \left(\frac{1}{k+1}\right)^2 \\ -16\left(\frac{1}{k+2}\right)^2 + 2\left(\frac{1}{k+1}\right)^2 \\ 7\left(\frac{1}{k+2}\right)^2 - \left(\frac{1}{k+1}\right)^2 \end{pmatrix},$$

yielding in turn that

$$\begin{array}{rcl} |q''(t)| & = & 2d_{2,k} + 6d_{3,k}t + 12d_{4,k}t^2 + 20d_{5,k}t^3 \\ & \leq & 2|d_{2,k}| + 6|d_{3,k}| + 12|d_{4,k}| + 20|d_{5,k}| \\ & \leq & 1 + 6\times 5 + 12\times 9 + 20\times 4 \\ & = & 219 \end{array}$$

for all $k \geq 0$ and all $t \in [0,1]$.

The graph of this function and its first three derivatives are given on the first 16 intervals and for $\eta = 10^{-4}$ in Figure 3.1. As for $f^{(2,1)} = f^{(1)}$, this figure confirms the properties inherited from the construction of the function $f(x)$, namely, that it is twice continuously differentiable and has uniformly bounded second derivatives. Its second derivative is now globally Lipschitz continuous, as can be seen in Figure 3.1 where one observes that the third derivative is bounded above in norm for all k. As in section 2, the fact that $f^{(2)}$ is bounded below by zero results from (3.10) and the fact that $\zeta(2) = \pi^2/6$.

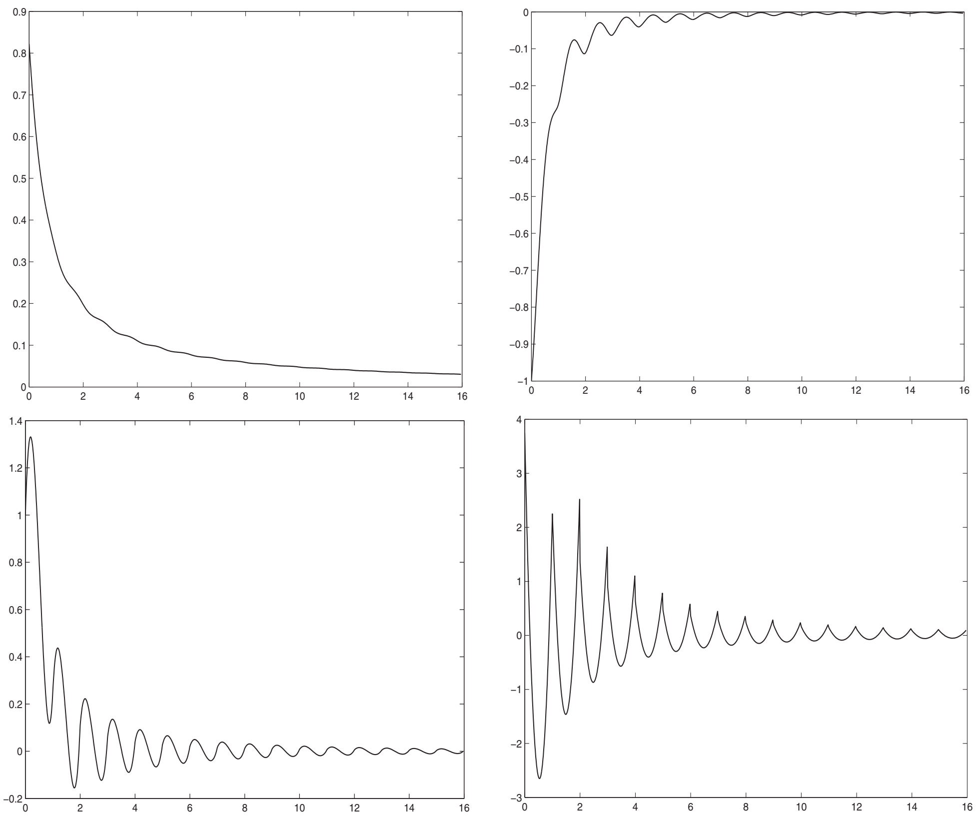

FIG. 3.1. *The function $f^{(2,2)}$ (top left) and its derivatives of order one (top right), two (bottom left), and three (bottom right) on the first* 16 *intervals.*

One may also compute and bound the third derivative of $f^{(2)}$ along the step, which is given in the kth interval by

$$\begin{array}{rcl}\frac{1}{\|s_k\|^3}[p_k'''(t)(s_k)_1^3 + q_k'''(t)] & < & p_k'''(t)(s_k)_1^3 + q_k'''(t)\\ & \leq & (6c_{3,k} + 24c_{4,k}t + 60c_{5,k}t^2)\mu_k^3\\ & & \quad + 6d_{3,k} + 24d_{4,k}t + 60d_{5,k}t^2\\ & < & 6|c_{3,k}|\mu_k + 24|c_{4,k}|\mu_k^2 + 60|c_{5,k}|\mu_k^3\\ & & \quad + 6|d_{3,k}| + 24|d_{4,k}| + 60|d_{5,k}|\\ & \leq & 6\times 4 + 24\times 7 + 60\times 3 + 6\times 5 + 24\times 9 + 60\times 4\\ & = & 858,\end{array}$$

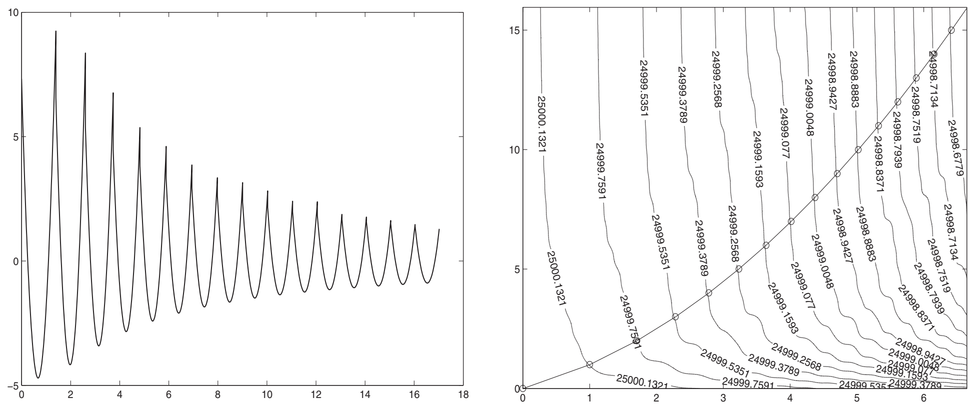

FIG. 3.2. *The third derivative of the function $f^{(2)}(x)$ along the path $[x_o, \ldots, x_{16}]$ (left) and this path on the level curves of $f^{(2)}$ (right).*

where we used the inequalities $\|s_k\| > 1$ and $t \leq 1$, and hence, because of the mean-value theorem, $f^{(2)}(x)$ has Lipschitz continuous second derivatives in each segment of the piecewise linear path $\cup_{k=0}^{\infty}[x_k, x_{k+1}]$. The actual value of the third derivative on the first segments of this path is shown on the left side of Figure 3.2, and the path itself is illustrated on the right side, superposed on the level curves of $f^{(2)}$. As a consequence, $f^{(2)}(x)$ satisfies AS.1, as desired.

This last example shows that Newton's method takes *at least* $O(\epsilon^{-2+\tau})$ iterations in the worst case. It cannot, however, take more than $O(\epsilon^{-2})$ iterations, since it can be included in globalization schemes such as trust regions which on this example would allow Newton steps to be taken and for which an upper bound of $O(\epsilon^{-2})$ iterations is known in the worst case for the same problem class (see Gratton, Sartenaer, and Toint (2008)). Note this latter bound requires bounded second derivatives, as is also the case for our last example.

3.2. Unbounded second derivatives. Remarkably, if we are now ready to allow unbounded Hessians and give up smoothness of the objective function beyond continuous differentiability, then a very different picture emerges. It is possible in this case to construct an example where Newton's method is arbitrarily slow in the sense that, *for any* $p > 0$, it takes precisely ϵ^{-p} iterations to generate $\|g_{k+1}\| \leq \epsilon$ when applied to a well-chosen $f^{(3)}$, with a certain x_0 and for any $\epsilon > 0$. Thus we relax our assumptions to

AS.2 f is twice continuously differentiable and bounded below.

This second (family) of examples is unidimensional and satisfies the conditions

$$x_0 = 0, \quad x_{k+1} = x_k - \frac{g_k}{H_k} \stackrel{\text{def}}{=} x_k + s_k,$$

for $k \geq 0$, where

$$g_k = -\left(\frac{1}{k+1}\right)^{\frac{1}{p}}, \quad H_k = (k+1)^q,$$

and

$$f_0 = \frac{1}{2}\zeta\left(\frac{2}{p}+q\right), \quad f_{k+1} = f_k - \frac{1}{2}\left(\frac{1}{k+1}\right)^{\frac{2}{p}+q},$$

where $p > 0$ and q is chosen to satisfy

$$1 - \frac{1}{p} \geq q > 1 - \frac{2}{p}. \tag{3.11}$$

One easily checks that $f_k - m_k(x_k + s_k) = f_k - f_{k+1}$ and that the iterates x_k tend to infinity.

We may now construct a twice continuously differentiable univariate function $f^{(3)}$ from $\mathbb{R}^n_+$ to $\mathbb{R}$ such that

$$f_k = f^{(3)}(x_k), \quad g_k = \nabla_x f^{(3)}(x_k), \quad \text{and} \quad H_k = \nabla_{xx} f^{(3)}(x_k)$$

for all $k \geq 0$ by construction on each interval $[x_k, x_{k+1}]$,

$$f^{(3)}(x) = p_k(x - x_k) + f_{k+1} \quad \text{for} \quad x \in [x_k, x_{k+1}],$$

where p_k is a polynomial of the type (2.12) such that

$$p_k(0) = \frac{1}{2}\left(\frac{1}{k+1}\right)^{\frac{2}{p}+q}, \quad p_k(s_k) = 0,$$

$$p_k'(0) = -\left(\frac{1}{k+1}\right)^{\frac{1}{p}}, \quad p_k'(s_k) = -\left(\frac{1}{k+2}\right)^{\frac{1}{p}},$$

as well as $p_k''(0) = (k+1)^q$ and $p_k''(s_k) = (k+2)^q$. Writing the interpolation conditions, one finds that

$$c_{0,k} = \frac{1}{2}\left(\frac{1}{k+1}\right)^{\frac{2}{p}+q}, \quad c_{1,k} = -\left(\frac{1}{k+1}\right)^{\frac{1}{p}}, \quad c_{2,k} = \frac{1}{2}(k+1)^q,$$

and that the remaining coefficients satisfy

$$\begin{pmatrix} s_k^3 & s_k^4 & s_k^5 \\ 3s_k^2 & 4s_k^3 & 5s_k^4 \\ 6s_k & 12s_k^2 & 20s_k^3 \end{pmatrix} \begin{pmatrix} c_{3,k} \\ c_{4,k} \\ c_{5,k} \end{pmatrix} = \begin{pmatrix} 0 \\ p_k'(s_k) \\ \Delta h_k \end{pmatrix},$$

where

$$\Delta h_k \stackrel{\text{def}}{=} (k+2)^q - (k+1)^q.$$

This system gives

$$\begin{pmatrix} c_{3,k} \\ c_{4,k} \\ c_{5,k} \end{pmatrix} = \begin{pmatrix} s_k^{-1}(\frac{1}{2}\Delta h_k - 4\phi_k) \\ s_k^{-2}(-\Delta h_k + 7\phi_k) \\ s_k^{-3}(\frac{1}{2}\Delta h_k - 3\phi_k) \end{pmatrix},$$

where now

$$\phi_k = \frac{p_k'(s_k)}{s_k} = -(k+1)^q \cdot \left(\frac{k+1}{k+2}\right)^{\frac{1}{p}}.$$

The resulting function $f^{(3)}$ is bounded below by zero and twice continuously differentiable (and thus satisfies AS.2), with minimum at infinity. Note, however, that its second derivatives are neither unbounded above nor Lipschitz continuous. The graph of this function and its first three derivatives on the first 16 intervals are shown in Figure 3.3.

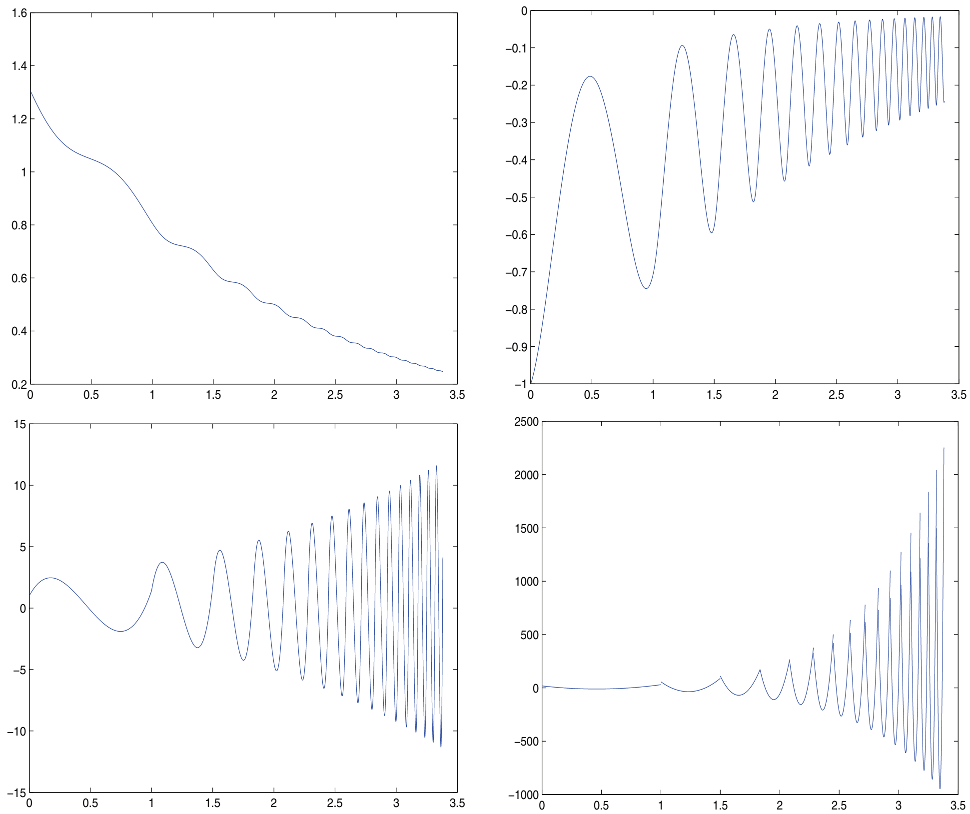

FIG. 3.3. *The function $f^{(3)}$ (top left) and its derivatives of order one (top right), two (bottom left), and three (bottom right) on the first* 16 *intervals.*

4. How slow is slow? Having shown in section 3.1 an example where, under standard assumptions, the performance of Newton's method is arbitrarily close to the worst case known for steepest descent, we now wish to comment on the degree of pessimism of this bound.

Returning to the multidimensional case, let us assume that (2.2) holds for some sequence of iterates $\{x_k\} \subset \mathbb{R}^n$ generated by Newton's method on a twice continuously differentiable objective function from $\mathbb{R}^n$ into $\mathbb{R}$ which is also bounded below by a constant f_{low} and has a uniformly bounded Hessian. Assume also that H_k is positive definite for all k and that every iteration of this process is successful in the sense that

$$f(x_k) - f(x_k + s_k) \geq \eta_1[m_k(x_k) - m_k(x_k + s_k)] \tag{4.1}$$

for some constant $\eta_1 \in (0, 1)$. Assume finally that the quadratic model (1.2) is minimized accurately enough to guarantee a model reduction at least as large as a fraction

κ of that obtained at the Cauchy point, which is defined as the solution of the (strictly convex) problem

$$\min_{t\geq 0} m_k(x_k - t g_k).$$

It is known (see Conn, Gould, and Toint (2000, section 6.3.2), for instance) that the solution t_k^C of this last problem and the associated model reduction satisfy

$$f(x_k) - m_k(x_k - t_k^C g_k) \geq \frac{\|g_k\|^4}{2 g_k^T H_k g_k}.$$

Thus our assumption yields that

$$f(x_k) - m_k(x_k + s_k) \geq \frac{\kappa\|g_k\|^4}{2 g_k^T H_k g_k} \geq \frac{\kappa\|g_k\|^2}{2\|H_k\|} \geq \frac{\kappa}{2\kappa_H}\|g_k\|^2, \tag{4.2}$$

where we used the Cauchy–Schwarz inequality to deduce the penultimate inequality and where κ_H is an upper bound on the Hessian norms. Because of (4.1), we obtain from (2.2) and (4.2) that

$$f(x_0) - f_{\rm low} \geq \eta_1 \sum_{k=0}^{\infty}[f(x_k) - m_k(x_k + s_k)] \geq \frac{\kappa\eta_1}{2\kappa_H}\sum_{k=0}^{\infty}\frac{1}{k+1}. \tag{4.3}$$

But this last inequality contradicts the divergence of the harmonic series. Hence we conclude that (2.2) cannot hold for our sequence of iterates. Thus a gradient sequence satisfying (2.3) is essentially as close to (2.2) as possible if the example is to be valid for all ϵ sufficiently small.

We may even pursue the analysis a little further. Let $\mathcal{K}$ denote the subset of the integers such that (2.2) holds. Then (4.3) implies that

$$\sum_{k\in\mathcal{K}} \frac{1}{k+1} < +\infty.$$

We then know from Behforooz (1995) that, in this case,

$$\lim_{\ell\to\infty} \frac{|\,\mathcal{K}\cap\mathcal{N}_\ell\,|}{10^\ell - |\,\mathcal{K}\cap\mathcal{N}_\ell\,|} = 0, \tag{4.4}$$

where $\mathcal{N}_\ell \stackrel{\rm def}{=} \{p\in\mathbb{N} \mid 0\leq p\leq 10^\ell\}$. But

$$\frac{|\,\mathcal{K}\cap(\mathcal{N}_\ell\setminus\mathcal{N}_{\ell-1})\,|}{10^\ell - 10^{\ell-1}} \leq \frac{10|\,\mathcal{K}\cap\mathcal{N}_\ell\,|}{9\times 10^\ell} \leq \frac{10}{9}\,\frac{|\,\mathcal{K}\cap\mathcal{N}_\ell\,|}{10^\ell - |\,\mathcal{K}\cap\mathcal{N}_\ell\,|},$$

and therefore, using (4.4),

$$\lim_{\ell\to\infty}\frac{|\,\mathcal{K}\cap(\mathcal{N}_\ell\setminus\mathcal{N}_{\ell-1})\,|}{|\,\mathcal{N}_\ell\setminus\mathcal{N}_{\ell-1}\,|} = \lim_{\ell\to\infty}\frac{|\,\mathcal{K}\cap(\mathcal{N}_\ell\setminus\mathcal{N}_{\ell-1})\,|}{10^\ell - 10^{\ell-1}} = 0.$$

Thus, if $\ell(k)$ is defined such that $k\in\mathcal{N}_{\ell(k)}\setminus\mathcal{N}_{\ell(k)-1}$, we have that $\lim_{k\to\infty}\ell(k) = \infty$ and therefore that

$$\begin{aligned}
\lim_{k\to\infty}{\rm Prob}_k\big[\,\|g_k\|\geq (k+1)^{-2}\,\big] &= \lim_{k\to\infty}{\rm Prob}_k[\,k+1\in\mathcal{K}\,]\\
&= \lim_{k\to\infty}{\rm Prob}_k[\,k+1\in\mathcal{K}\cap(\mathcal{N}_{\ell(k)}\setminus\mathcal{N}_{\ell(k)-1})\,]\\
&= 0,
\end{aligned}$$

where $\text{Prob}_k[\cdot]$ is the probability with uniform density on $\{10^{\ell(k)-1}+1,\dots,10^{\ell(k)}\}$. As a consequence, *the probability that the termination test* (2.1) *is satisfied for an arbitrary k in the range* $[\,10^{\ell(\lfloor\epsilon^{-1/2}\rfloor)-1}+1, 10^{\ell(\lfloor\epsilon^{-1/2}\rfloor)}\,]$ *tends to one when ϵ tends to zero.*

How do we interpret these results? What we have shown is that, under the conditions stated before, the statement

$$\text{there exists } \theta>0 \text{ such that, for all } k \text{ arbitrarily large, } ||g_k|| \geq \theta\left(\frac{1}{k+1}\right)^2$$

is false. This is to say that

$$\text{for all } \theta>0 \text{ there exists } k \text{ arbitrarily large such that } ||g_k|| < \theta\left(\frac{1}{k+1}\right)^2.$$

In fact, we have proved that the proportion of "good" k's for which this last inequality holds (for a given θ) grows asymptotically. But it is important to notice that this last statement does not contradicts the worst-case bound of $O(\epsilon^{-2})$ mentioned above, which is

$$\text{there exists } \theta>0 \text{ such that, for all } \epsilon>0 \text{ and } k\geq\frac{\theta}{\epsilon^2},\ \|g_k\|\leq\epsilon.$$

Indeed, if ϵ is given, there is no guarantee that the particular k such that $k=\theta(k+1)^{-2}$ belongs to the set of "good" k's. As a consequence, we see that the worst-case analysis is increasingly pessimistic for ϵ tending to zero.

We conclude this section by noting that the arguments developed for Newton's method also turn out to apply for the steepest-descent method, as it can also be shown for this case that

$$f(x_k) - m_k(x_k - t_k^C g_k) \geq \kappa_{\text{SD}}\|g_k\|^2$$

for some $\kappa_{\text{SD}}>0$ depending on the maximal curvature of the objective function (see, for instance, Conn, Gould, and Toint (2000, Theorem 6.3.3), with Δ_k sufficiently large, or Nesterov (2004, relation (1.2.13), page 27)). This inequality then replaces (4.2) in the above reasoning.

5. Less slow convergence for ARC. Now we consider using the ARC algorithm for solving (1.1), using exact second-order information. As above, we would like to construct an example on which ARC converges at a rate which corresponds to its worst-case behavior for general nonconvex objective functions, i.e., such that one has to perform $O(\epsilon^{-\frac{3}{2}})$ iterations to ensure (2.1). This goal is clearly achieved by a gradient sequence such that

$$\|g_k\| \geq \left(\frac{1}{k+1}\right)^{\frac{2}{3}}.$$

An arbitrarily close approximation is again considered by requiring that, for any $\tau>0$, the ARC method needs $O(\epsilon^{-\frac{3}{2}+\tau})$ iterations to achieve (2.1), which leads to the condition that, for all $k\geq 0$,

$$\|g_k\| = \left(\frac{1}{k+1}\right)^{\frac{2}{3-2\tau}}. \tag{5.1}$$

Our new objective is therefore to construct sequences $\{x_k\}$, $\{g_k\}$, $\{H_k\}$, $\{\sigma_k\}$, and $\{f_k\}$ such that (5.1) holds and which may be generated by the ARC algorithm together with a function $f^{(4)}(x)$ satisfying AS.1 such that

$$f_k = f^{(4)}(x_k), \quad g_k = \nabla_x f^{(4)}(x_k), \text{ and } H_k = \nabla_{xx} f^{(4)}(x_k)$$

for all $k \geq 0$, which is bounded below and whose Hessian $\nabla_{xx} f^{(4)}(x)$ is Lipschitz continuous with global Lipschitz constant $L \geq 0$.

Our example is again unidimensional with iterates tending to infinity. These iterates are defined for all $k \geq 0$ by

$$x_0 = 0, \quad x_{k+1} = x_k + \left(\frac{1}{k+1}\right)^{\frac{1}{3}+\eta}, \tag{5.2}$$

$$f_0 = \frac{2}{3}\zeta(1+3\eta), \quad f_{k+1} = f_k - \frac{2}{3}\left(\frac{1}{k+1}\right)^{1+3\eta}, \tag{5.3}$$

$$g_k = -\left(\frac{1}{k+1}\right)^{\frac{2}{3}+2\eta}, \quad H_k = 0, \text{ and } \sigma_k = 1, \tag{5.4}$$

where now

$$\eta = \eta(\tau) \stackrel{\text{def}}{=} \frac{1}{2}\left(\frac{2}{3-2\tau} - \frac{2}{3}\right) = \frac{2\tau}{9-6\tau} > 0.$$

Observe that (5.4) gives (5.1) by construction.

Let us verify that, provided (3.1) holds, the sequences given by (5.2)–(5.4) may be generated by the ARC algorithm, whose every iteration is very successful[2] in the sense that

$$f^{(4)}(x_k) - f^{(4)}(x_k + s_k) \geq \eta_2[m_k(x_k) - m_k(x_k + s_k)] \tag{5.5}$$

for some constant $\eta_2 \in (0,1)$. Using (1.3), this amounts to verifying that

$$g_k + \big(H_k + \sigma_k\|s_k\|\big)s_k = 0, \tag{5.6}$$

$$H_k + \sigma_k\|s_k\| \text{ is positive semidefinite,} \tag{5.7}$$

$$\sigma_k > 0, \quad \sigma_{k+1} \leq \sigma_k, \tag{5.8}$$

and

$$f_{k+1} = m_k(x_k + s_k) \tag{5.9}$$

[2]The trust-region literature (see Conn, Gould, and Toint (2000), for instance) usually distinguishes between *successful* and *very successful* iterations in that the constant η_1 for the former (in (4.1)) is typically chosen quite small (e.g., 10^{-3}), whereas the constant η_2 for the latter (in (5.5)) is much closer to one (e.g., 0.9). The trust-region radius is not decreased in the latter case. The same distinction is made for the ARC method, in which case the regularization parameter σ_k is not increased on very successful iterations (see Cartis, Gould, and Toint (2009)).

for all $k \geq 1$. (Note that (5.9) ensures that every iteration is very successful, as requested.) Because the model is unidimensional, the first two of these conditions say that the cubic model is globally minimized exactly. Observe first that (5.8) immediately results from (5.4). In our case, (5.6) becomes, using (5.2) and (5.4),

$$g_k + H_k s_k + \sigma_k \|s_k\| s_k = -\left(\frac{1}{k+1}\right)^{\frac{2}{3}+2\eta} + 0 + \left(\frac{1}{k+1}\right)^{\frac{2}{3}+2\eta} = 0,$$

as desired, while (5.7) also follows from (5.2) and (5.4). Using (5.3) and (5.4), we also obtain that

$$\begin{aligned} m_k(x_k + s_k) &= f_k + g_k^T s_k + \tfrac{1}{2} s_k^T H_k s_k + \tfrac{1}{3}\sigma_k \|s_k\|^3 \\ &= f_k - \tfrac{2}{3}\left(\tfrac{1}{k+1}\right)^{1+3\eta} \\ &= f_{k+1}, \end{aligned}$$

which in turn yields (5.9).

As was the case in the previous sections, the only remaining question is to exhibit a bounded below and twice continuously differentiable function $f^{(4)}(x)$ with a Lipschitz continuous Hessian (in each segment $[x_k, x_{k+1}]$) satisfying conditions (5.2)–(5.4), and we define it on the nonnegative reals[3] by the recursion

$$f^{(4)}(x) = p_k(x - x_k) + f_{k+1} \quad \text{for} \quad x \in [x_k, x_{k+1}] \quad \text{and} \quad k \geq 0,$$

where p_k is an polynomial Hermite interpolant of the form (2.12) on the interval $[0, s_k]$ such that,

$$p_k(0) = \frac{2}{3}\left(\frac{1}{k+1}\right)^{1+3\eta}, \quad p_k(s_k) = 0, \tag{5.10}$$

(5.11)

$$p_k'(0) = -\left(\frac{1}{k+1}\right)^{\frac{2}{3}+2\eta}, \quad p_k'(s_k) = -\left(\frac{1}{k+2}\right)^{\frac{2}{3}+2\eta}, \quad p_k''(0) = p_k''(s_k) = 0.$$

These conditions immediately give that

$$c_{0,k} = \frac{2}{3}\left(\frac{1}{k+1}\right)^{1+3\eta}, \quad c_{1,k} = -\left(\frac{1}{k+1}\right)^{\frac{2}{3}+2\eta}, \text{ and } c_{2,k} = 0.$$

In this case, the remaining interpolation conditions may be written in the form

$$\begin{pmatrix} s_k^3 & s_k^4 & s_k^5 \\ 3s_k^2 & 4s_k^3 & 5s_k^4 \\ 6s_k & 12s_k^2 & 20s_k^3 \end{pmatrix} \begin{pmatrix} c_{3,k} \\ c_{4,k} \\ c_{5,k} \end{pmatrix} = \begin{pmatrix} -p_k(0) - p_k'(0)s_k \\ p_k'(s_k) - p_k'(0) \\ 0 \end{pmatrix},$$

whose solution is now given by

$$\begin{pmatrix} c_{3,k} \\ c_{4,k} \\ c_{5,k} \end{pmatrix} = \begin{pmatrix} \frac{10}{3} - 4\phi_k \\ \frac{1}{s_k}[-5 + 7\phi_k] \\ \frac{1}{s_k^2}[2 - 3\phi_k] \end{pmatrix}, \tag{5.12}$$

[3] Again, it can be easily smoothly extended to the negative reals while maintaining its boundedness and the Lipschitz continuity of its second derivatives.

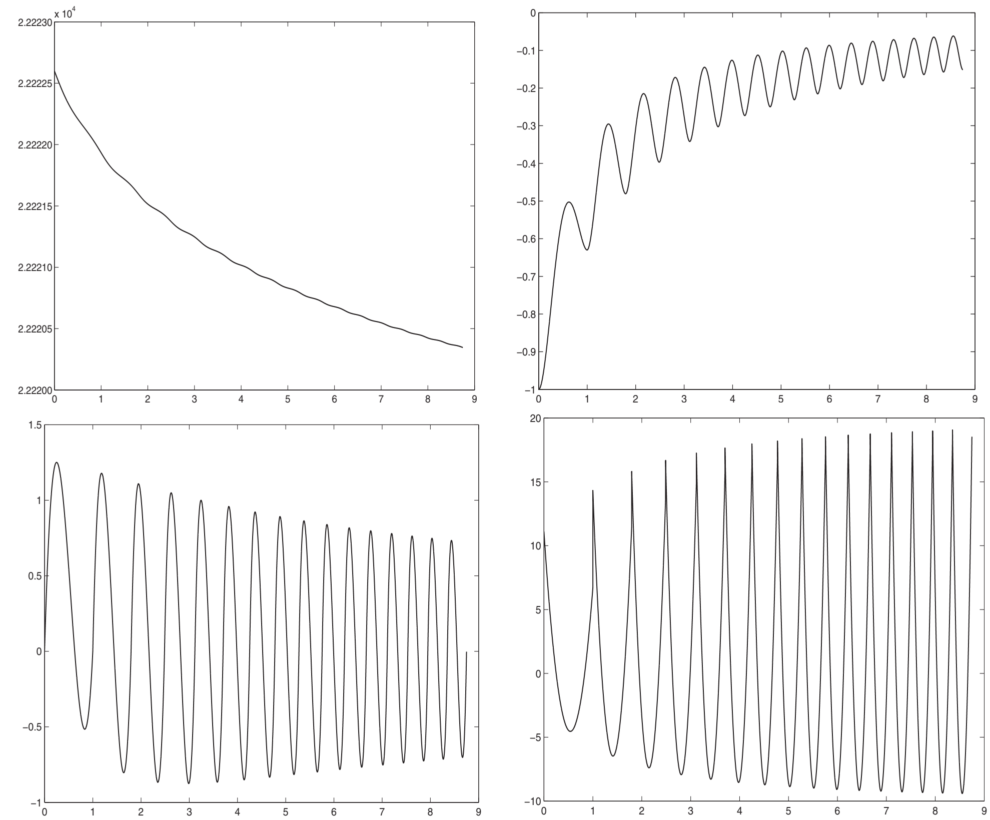

FIG. 5.1. *The function $f^{(4)}$ (top left) and its derivatives of order one (top right), two (bottom left), and three (bottom right) on the first* 16 *intervals.*

with

$$\phi_k \stackrel{\text{def}}{=} \frac{1}{s_k^2}\left[\left(\frac{1}{k+1}\right)^{\mu} - \left(\frac{1}{k+2}\right)^{\mu}\right], \quad \text{where } \mu \stackrel{\text{def}}{=} \frac{2}{3} + 2\eta.$$

The definition of ϕ_k implies that $\phi_k \in (0,1)$ for all $k \geq 0$ and, hence, using (5.12), that

$$\begin{aligned} |p'''(t)| &= 6c_{3,k} + 24c_{4,k}t + 60c_{5,k}t^2 \\ &\leq 6c_{3,k} + 24c_{4,k}s_k + 60c_{5,k}s_k^2 \\ &\leq 6 \times \tfrac{10}{3} + 24 \times 13 + 60 \times 2 \\ &= 452 \end{aligned} \tag{5.13}$$

for all $k \geq 0$ and all $t \in [0, s_k]$, and f has Lipschitz continuous second derivatives along the path of iterates, which is $\mathbb{R}^+$. The desired objective function for our final counterexample is then well-defined and clearly satisfies AS.1. The graph of this function and its first three derivatives are given on the first 16 intervals and for $\eta = 10^{-4}$ in Figure 5.1. This figure confirms the properties of the function $f^{(4)}(x)$, namely, that it is twice continuously differentiable and has uniformly bounded third derivative (in Figure 5.1, the maximum is achieved on each interval by the first point in the interval, where (5.12) and (5.13) imply that $|p'''(0)| \leq 20$). Thus its second derivative is globally Lipschitz continuous with constant $L \leq 452$ ($L = 20$ for the function plotted).

As in our first example, the figure reveals the nonconvexity and monotonically decreasing nature of $f(x)$. As in section 2, the fact that $f(x)$ is bounded below by zero finally results from (5.3) and the definition of the Riemann ζ function; note that, with the choice of η, $1 + 3\eta = 1.0003$ and $f_0 = (2/3)\zeta(1.0003) \approx 22222.6$.

6. Conclusions. We now summarize the result obtained in this paper. Considering the steepest method first and assuming Lipschitz continuity of the objective function's gradient along the path of iterates, we have, for any $\tau > 0$, exhibited valid examples for which this algorithm produces a sequence of slowly converging gradients. This in turn implies that, for any $\epsilon \in (0,1)$, at least,

$$\left\lfloor \frac{1}{\epsilon^{2-\tau}} \right\rfloor$$

iterations and function evaluations are necessary for this algorithm to produce an iterate x_k such that $\|g_k\| \leq \epsilon$. This lower bound is arbitrarily close to the upper bound of $O(\epsilon^{-2})$ known for this algorithm. Other examples have also been constructed showing that the same complexity can be achieved by Newton's method for twice continuously differentiable functions whose Hessian is Lipschitz continuous on the path defined by the iterates, thereby proving that Newton's method may be as (in)efficient as the steepest-descent method (in its worst case). The fact that (3.8) and (5.9) hold ensures that our conclusions are also valid if the standard Newton's method is embedded in a trust-region globalization framework (see Conn, Gould, and Toint (2000) for an extensive coverage of such methods), since it guarantees that every iteration is very successful in that case (in the sense that (5.5) holds), and that the initial trust-region may then be chosen large enough to be irrelevant. The conclusions also apply if a linesearch globalization is used (see Dennis and Schnabel (1983) or Nocedal and Wright (1999)), because the unit step is then acceptable at every iteration (in our examples), or in the filter context, because the gradient is monotonically converging to zero. We have also provided an example where Newton's method requires exactly $1/\epsilon^2$ iterations to produce an iterate x_k such that $\|g_k\| \leq \epsilon$, but we had to give up boundedness of second derivatives to obtain this sharper bound. In addition, we have provided some analysis in an attempt to quantify how pessimistic the obtained worst-case bounds can be.

We have then extended the methodology to cover the ARC algorithm, which can be viewed as a regularized version of Newton's method. For any $\tau > 0$, we have exhibited a valid example for which the ARC algorithm produces a sequence of gradients satisfying (5.1). This equality yields that, for any $\epsilon \in (0,1)$, at least

$$\left\lfloor \frac{1}{\epsilon^{\frac{3}{2}-\tau}} \right\rfloor$$

iterations and function evaluations are necessary for this algorithm to produce an iterate x_k such that $\|g_k\| \leq \epsilon$. This lower bound is arbitrarily close to the upper bound of $O(\epsilon^{-3/2})$, thereby proving that this last bound is sharp.

We have not been able to show that the steepest-descent method may take at least $O(\epsilon^{-2})$ evaluations to achieve a gradient accuracy of ϵ on functions with Lipschitz continuous second derivatives, thereby not excluding the (unlikely) possibility that the steepest-descent method could be better than Newton's method on sufficiently smooth functions.

Our result that the ARC method is the best second-order algorithm available so far (from the worst-case complexity point of view) suggests further research directions beyond that of settling the open question mentioned in the previous paragraph. Is the associated complexity bound in $O(\epsilon^{-3/2})$ the best that can be achieved by *any* second-order method for general nonconvex objective functions? And how best to characterize the complexity of an unconstrained minimization problem? These interesting issues remain challenging.

Acknowledgments. Thanks are due to two anonymous referees for their constructive comments.

REFERENCES

H. Behforooz, *Thinning out the harmonic series*, Math. Mag., 68 (1995), pp. 289–293.

C. Cartis, N. I. M. Gould, and Ph. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part* I*: Motivation, convergence and numerical results*, Math. Program., Ser. A, (2009), DOI: 10.1007/s10107-009-0286-5.

C. Cartis, N. I. M. Gould, and Ph. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part* II*: Worst-case function- and derivative-evaluation complexity*, Math. Program., Ser. A, (2010a), DOI: 10.1007/s10107-009-0337-y.

C. Cartis, N. I. M. Gould, and Ph. L. Toint, *Efficiency of Adaptive Cubic Regularisation Methods on Convex Problems*, Technical Report (in preparation), Department of Mathematics, FUNDP—University of Namur, Namur, Belgium, 2010b.

A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, PA, 2000.

J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

R. Fletcher, *Practical Methods of Optimization*, 2nd ed., J. Wiley & Sons, Chichester, England, 1987.

N. I. M. Gould, C. Sainvitu, and Ph. L. Toint, *A filter-trust-region method for unconstrained optimization*, SIAM J. Optim., 16 (2005), pp. 341–357.

S. Gratton, A. Sartenaer, and Ph. L. Toint, *Recursive trust-region methods for multiscale nonlinear optimization*, SIAM J. Optim., 19 (2008), pp. 414–444.

A. Griewank, *The modification of Newton's method for unconstrained optimization by bounding cubic terms*, Technical report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom, 1981.

Yu. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

Yu. Nesterov, *Accelerating the cubic regularization of Newton's method on convex problems*, Math. Program., Ser. B, 112 (2008), pp. 159–181.

Yu. Nesterov and B. T. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program., Ser. A, 108 (2006), pp. 177–205.

J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer Verlag, Heidelberg, Berlin, New York, 1999.

R. Polyak, *Regularized Newton method for unconstrained convex optimization*, Math. Program., Ser. A, 120 (2009), pp. 125–145.

K. Ueda, *Regularized Newton Method without Line Search for Unconstrained Optimization*, Ph.D. thesis, Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto, Japan, 2009.

K. Ueda and N. Yamashita, *A regularized Newton method without line search for unconstrained optimization*, Technical report 2009-007, Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto, Japan, 2009.

K. Ueda and N. Yamashita, *Convergence properties of the regularized Newton method for the unconstrained nonconvex optimization*, Appl. Math. Optim., 62 (2010), pp. 27–46.

M. Weiser, P. Deuflhard, and B. Erdmann, *Affine conjugate adaptive Newton methods for nonlinear elastomechanics*, Optim. Methods Softw., 22 (2007), pp. 413–431.