Science & Technology Facilities Council

Technical Report

On the complexity of steepest descent, Newton's method and regularized Newton methods for nonconvex unconstrained optimization

C Cartis N I M Gould Ph L Toint

November 9, 2009

RAL-TR-2009-023

Library and Information Services
SFTC Rutherford Appleton Laboratory
Harwell Science and Innovation Campus
Didcot
OX11 0QX
UK
Tel: +44 (0)1235 445384
Fax: +44(0)1235 446403
Email: library@rl.ac.uk

The STFC ePublication archive (epubs), recording the scientific output of the Chilbolton, Daresbury, and Rutherford Appleton Laboratories is available online at: http://epubs.cclrc.ac.uk/

ISSN 1358-6254

Neither the Council nor the Laboratory accept any responsibility for loss or damage arising from the use of information contained in any of their reports or in any communication about their tests or investigation

RAL-TR-2009-023

On the complexity of steepest descent, Newton's method and regularized Newton methods for nonconvex unconstrained optimization

Coralia Cartis[1,2], Nicholas I. M. Gould[2,3] and Philippe L. Toint[4,5]

ABSTRACT

It is shown that the steepest descent and Newton's method for unconstrained nonconvex optimization under standard assumptions may require numbers 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 method is tight, and that Newton's method may be as slow as steepest descent in the worst case. The improved evaluation complexity bound of $O(\epsilon^{-3/2})$ evaluations known for cubically-regularised Newton methods is also shown to be tight.

---

[1] School of Mathematics, The King's Buildings, University of Edinburgh, EH9 3JZ, Scotland, EU. Email: coralia.cartis@ed.ac.uk .

[2] This work was supported by the EPSRC grants EP/E053351/1 and EP/F005369/1.

[3] Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, England, EU. Email: nick.gould@stfc.ac.uk . Current reports available from "http://www.numerical.rl.ac.uk/reports/reports.shtml".

[4] This work was supported by the EPSRC grant EP/G038643/1 and 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)"

[5] Department of Mathematics, Facultés Universitaires ND de la Paix, 61, rue de Bruxelles, B-5000 Namur, Belgium, EU. Email : philippe.toint@fundp.ac.be . Current reports available from "http://www.fundp.ac.be/~phtoint/pht/publications.html".

Computational Science and Engineering Department
Atlas Centre
Rutherford Appleton Laboratory
Oxfordshire OX11 0QX

November 9, 2009

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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, amongst which the steepest-descent and Newton method 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 behaviour of steepest descent 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

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

where $H_k \overset{\text{def}}{=} \nabla_{xx} f(x_k)$ is assumed to be positive definite. This algorithm to known to converge locally and quadratically to strict local minimizers of the objective function f, but in general convergence from arbtrary 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 surpringly 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 may be found in Dennis and Schnabel (1983), of which 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 (2008, 2009) and Ueda (2009), which essentially proves the same result for a variant of Newton's method using Levenberg-Morrison-Marquardt regularization.

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*a*), 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

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

where $\sigma_k \geq 0$ is a suitably chosen regularization parameters (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 (2007) 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 (2009*b*), 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 behaviour 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 favourable 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 method can be shown, while Section 4 further discusses the implications of these examples (and the interpretation of 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. Some conclusions are finally 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

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

In order to achieve this goal, a suitable condition is to require that, for all $k \geq 0$,

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

An arbitrarily close approximation can considered by requiring that, for any $\tau > 0$, Newton's method needs $O(\epsilon^{-2+\tau})$ iterations to achieve (2.1), which leads to the 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

$$f_k = f_1(x_k), \quad \text{and} \quad g_k = \nabla_x f_1(x_k) \tag{2.4}$$

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, for all $k \geq 0$,

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

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

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

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

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

and $\zeta(t) \stackrel{\rm def}{=} \sum_{k=1}^{\infty} k^{-t}$ 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 $\alpha < \beta$,

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

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

We now exhibit function $f_1(x)$ which satisfies AS.0 and (2.4)-(2.9). For this purpose, we use polynomial Hermite interpolation on the interval $[0, x_{k+1} - x_k]$, which we will subsequently translate. We are thus seeking a polynomial of the form

$$p_k(t) \stackrel{\rm def}{=} c_{0,k} + c_{1,k}t + c_{2,k}t^2 + c_{3,k}t^3 + c_{4,k}t^4 + c_{5,k}t^5 \tag{2.11}$$

on the interval $[0, \mu_k]$ (where $\mu_k = s_k$) such that

$$p_k(0) = \alpha_k(1-\tfrac{1}{2}\alpha_k)\left(\frac{1}{k+1}\right)^{1+2\eta}, \quad p_k(\mu_k) = 0, \tag{2.12}$$

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

and we also impose that $p_k''(0) = p_k''(\mu_k) = 1$. These conditions immediately give that

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

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

$$\begin{pmatrix} \mu_k^3 & \mu_k^4 & \mu_k^5 \\ 3\mu_k^2 & 4\mu_k^3 & 5\mu_k^4 \\ 6\mu_k & 12\mu_k^2 & 20\mu_k^3 \end{pmatrix} \begin{pmatrix} c_3 \\ c_4 \\ c_5 \end{pmatrix} = \begin{pmatrix} 0 \\ p_k'(\mu_k) \\ 0 \end{pmatrix},$$

whose solution turns out to be

$$\begin{pmatrix} c_{3,k} \\ c_{4,k} \\ c_{5,k} \end{pmatrix} = \begin{pmatrix} -4\,\dfrac{\phi_k}{\mu_k} \\ 7\,\dfrac{\phi_k}{\mu_k^2} \\ -3\,\dfrac{\phi_k}{\mu_k^3} \end{pmatrix} \tag{2.14}$$

where

$$\phi_k = \frac{1}{\alpha_k}(1-\alpha_k-\psi_k) \;\text{ with }\; \psi_k \overset{\text{def}}{=} \left(\frac{k+1}{k+2}\right)^{\frac{1}{2}+\eta}. \tag{2.15}$$

The definition of ψ_k implies that $|\psi_k| \in (0,1)$ for all $k \geq 0$, The function f_1 is then recursively defined on the nonnegative reals[1] by

$$f_1(x) = p_k(x-x_k) + f_{k+1} \;\text{ for }\; x \in [x_k, x_{k+1}] \;\text{ and }\; k \geq 0. \tag{2.16}$$

The graph of this function and its first three derivatives are given on the first 16 intervals and for $\eta = 10^{-4}$ and $\alpha_k = 1$ by Figure 2.1.

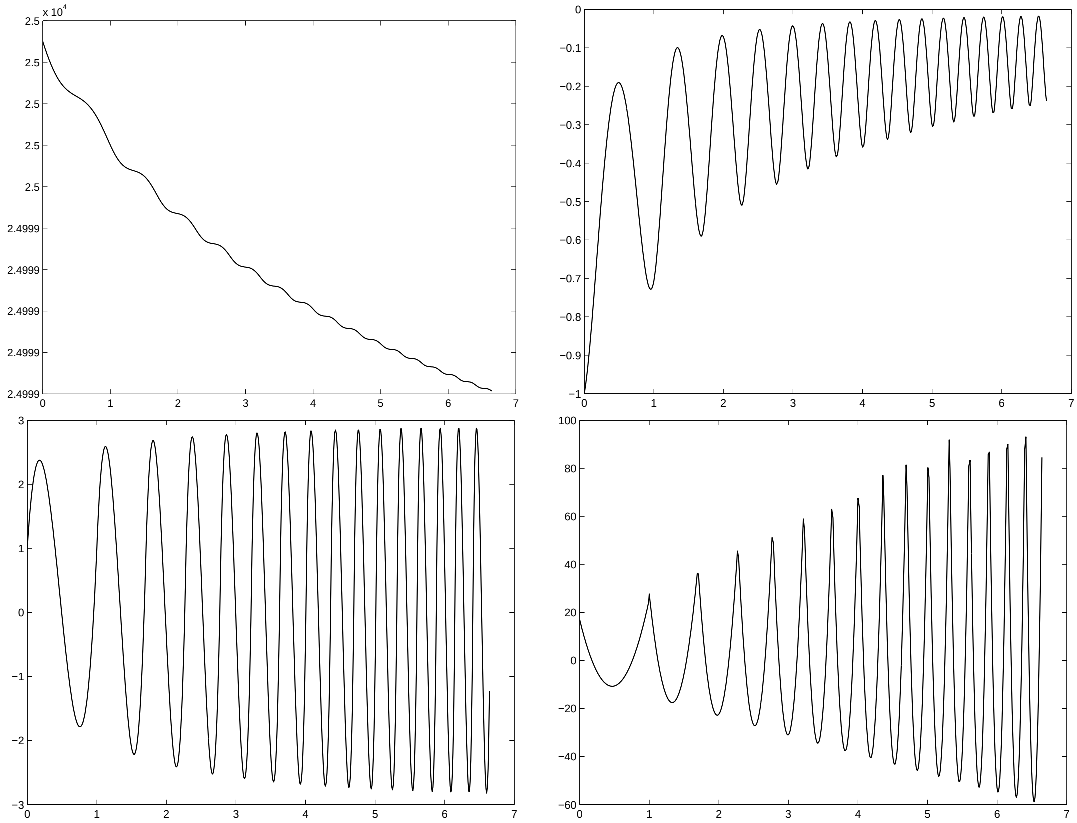

Figure 2.1: The function f_1 and its first three derivatives (from top to bottom and left to right) on the first 16 intervals

This figure confirms the properties inherited from the construction of the function f_1, namely that it is twice continuoulsy differentiable with bounded second derivatives. This last observation results from the bound

$$\begin{aligned} |p''(t)| &= 2c_{2,k} + 6c_{3,k}t + 12c_{4,k}t^2 + 20c_{5,k}t^3 \\ &\leq 2|c_{2,k}| + 6|c_{3,k}|\mu_k + 12|c_{4,k}|\mu_k^2 + 20|c_{5,k}|\mu_k^3 \\ &\leq 1 + 150|\phi_k| \\ &\leq 1 + 150\max[1,\overline{\alpha}]/\underline{\alpha} \end{aligned} \tag{2.17}$$

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

for all $k \geq 0$ and all $t \in [0, \mu_k]$, where we used (2.14) 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 it 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

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

and the definition of the Riemann ζ function (note that $\zeta(1.0002) \approx 50000.6$).

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

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

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

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 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

$$f_k = f_2(x_k), \quad g_k = \nabla_x f_2(x_k) \text{ and } H_k = \nabla_{xx} f_2(x_k). \tag{3.1}$$

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 had 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 we define, for all $k \geq 0$,

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

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

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

where, as in (2.10), $\eta = \tau/(4-2\tau) > 0$ and $\zeta(t) \stackrel{\rm def}{=} \sum_{k=1}^{\infty} k^{-t}$ 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

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

and remembering (1.2), this amounts to verifying that

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

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

and that

$$f(x_k+s_k) = m_k(x_k+s_k) \tag{3.8}$$

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. In our case, (3.6) becomes, using (3.5), (3.2) and (3.4),

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

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

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

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 i-th component of the vector x and where the univariate $f_{2,1}$ and $f_{2,2}$ are computed separately. Since our conditions involve, for both functions, fixed values of the function

$$f_{2,1}(0) = \frac{1}{2}\zeta(1+2\eta), \quad f_{2,1}([x_{k+1}]_1) = f_{2,1}([x_k]_1) - \frac{1}{2}\left(\frac{1}{k+1}\right)^{1+2\eta}, \tag{3.9}$$

$$f_{2,2}(0) = 1/2\zeta(2), \quad f_{2,2}([x_{k+1}]_2) = f_{2,2}([x_k]_2) - \frac{1}{2}\left(\frac{1}{k+1}\right)^2, \tag{3.10}$$

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 satsify 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

$$f_{2,1}([x]_1) = f_1([x]_1).$$

Let us now consider $f_{2,2}$. Again, we seek a polynomial

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

on the interval $[0, 1]$ 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

$$f_{2,2}([x]_2) = q_k([x_2 - x_k]_2) + f_{2,2}([x_{k+1}]_2) \text{ for } [x]_2 \in [[x_k]_2, [x_{k+1}]_2] \text{ and } k \geq 0,$$

and that

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

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 $\eta = 10^{-4}$ by 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 continuous differentiable and has uniformly bounded second derivative. Its second derivative is now globally Lipschitz continuous, as it can be seen in Figure 3.1 where one observes that the third derivative is bounded above in norm for all k. The fact that f_2 is bounded below by zero results from (3.10) and the fact that $\zeta(2) = \pi^2/6$.

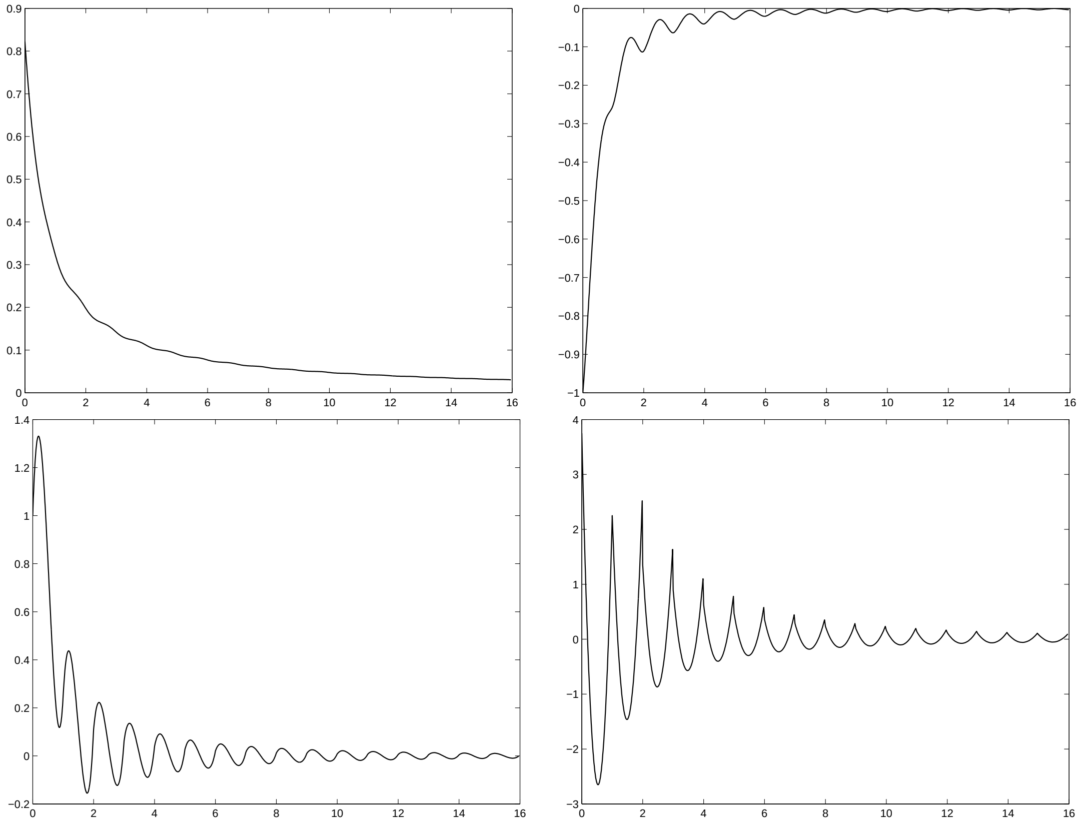

Figure 3.1: The function f_2 and its first three derivatives (from top to bottom and left to right) on the first 16 intervals

One may also compute the third derivative of f_2 along the step, which is given, in the k-th interval, by

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

where we used the inequalities $\|s_k\| > 1$ and $t \le 1$ 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

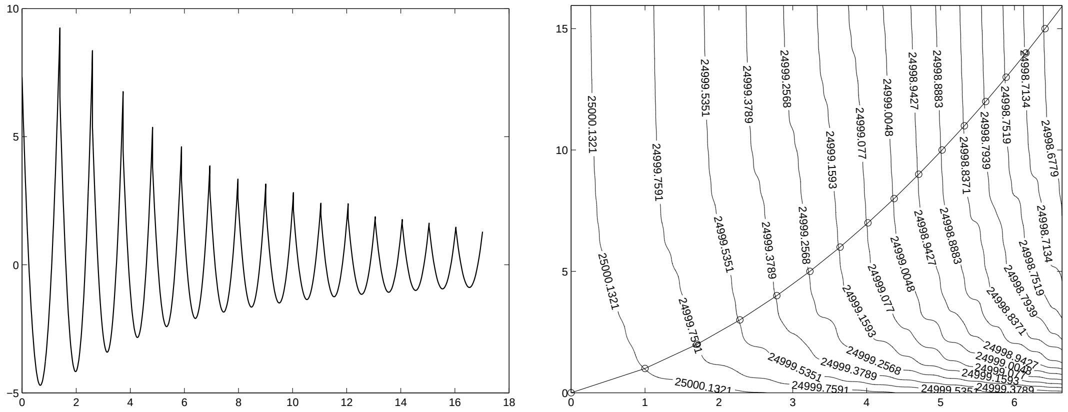

Figure 3.2: The third derivative of the function $f_2(x)$ along the path $[x_o, \ldots, x_{16}]$, and this path on the level curves of f_2.

this path is shown on the left side of Figure 3.2, while the path itself is illustrated on the right side, superposed on the levels curves of f. As a consequence, $f_2(x)$ satisfies AS.1, as desired.

If we are now ready to give up smoothness of the objective function beyond continuous differentiability, it is then possible to construct an example with $\tau = \eta = 0$, thereby guaranteeing that Newton's method takes precisely ϵ^{-2} iterations to generate $\|g_{k-1}\| \leq \epsilon$ when applied to f_3, with a certain x_0 and for any $\epsilon > 0$. Thus we relax our asumptions to

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

This second example is unidimensional and satisfies the conditions

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

for $k \geq 0$, where

$$g_k = -\left(\frac{1}{k+1}\right)^{\frac{1}{2}}, \quad H_k = k+1$$

and

$$f_3(0) = \frac{1}{2}\zeta(2), \quad f_3(x_k + s_k) = m_k(x_k + s_k).$$

One easily checks that

$$f_3(x_k) - m_k(x_k + s_k) = f_3(x_k) - f_3(x_{k+1}) = \frac{1}{2}\left(\frac{1}{k+1}\right)^2.$$

We may now contruct a twice continuously differentiable univariate function from $\mathbb{R}^+$ into $\mathbb{R}$ by constructing, on each interval $[x_k, x_{k+1}]$, a polynomial of the type (2.11) such that

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

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

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

$$\begin{pmatrix} s_k^3 & s_k^4 & s_k^5 \\ 3s_k^2 & 4s_k^3 & 5s_k^4 \\ 6s_k & 12s_k^2 & 20s_k^3 \end{pmatrix} \begin{pmatrix} c_3 \\ c_4 \\ c_5 \end{pmatrix} = \begin{pmatrix} 0 \\ p_k'(s_k) \\ 1 \end{pmatrix},$$

whose solution is given by

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

and

$$\begin{pmatrix} c_3 \\ c_4 \\ c_5 \end{pmatrix} = \begin{pmatrix} s_k^{-1}\left(\frac{1}{2} - 4\phi_k\right) \\ s_k^{-2}\left(-1 + 7\phi_k\right) \\ s_k^{-3}\left(\frac{1}{2} - 3\phi_k\right), \end{pmatrix}$$

where now

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

To complete this example, we may then set

$$f_3(x) = p_k(x - x_k) + f_3(x_{x+1}) \;\text{ for }\; x \in [x_k, x_{k+1}].$$

Observe, as above that we may extend $f_3(x)$ to the negative reals by defining $f_3(x) = f_3(0) + xf_3'(0) + 1/2x^2 f_3''(0)$ for $x < 0$, and beyond $x_* = \sum_{k=0}^{\infty} s_k = \zeta(3/2)$ by symmetrizing it with respect to this point, i.e.

$$f_3(x + \zeta(3/2)) = f_3(\zeta(3/2) - x) \;\text{ for }\; x > 0.$$

The resulting function is bounded below (by zero), continuously differentiable on $\mathbb{R}$ (as thus satisfies AS.2) and twice continuously differentiable everywhere except at $\zeta(3/2)$, where both left and right second derivatives are infinite (it is therefore not Lipschitz continuous either). It also has a unique minimizer in $\zeta(3/2)$. The graph of this function and its first three derivatives on the first 16 intervals are shown in Figure 3.3.

It is unclear whether an example with $\tau = \eta = 0$ can be found without weakening the smoothness assumptions made at the start of this section, as we have just done. Interestingly, yet another example of $\Theta(\epsilon^{-2})$ convergence for Newton's method may be constructed along the lines of the one just presented, by defining H_k, the Hessian at x_k, to be $\sqrt{k+1}$ instead of $k+1$. The minimum of the function f is then at infinity, but continuous second derivatives are preserved although they remain unbounded.

4 How slow is slow?

Having shown an example where 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.

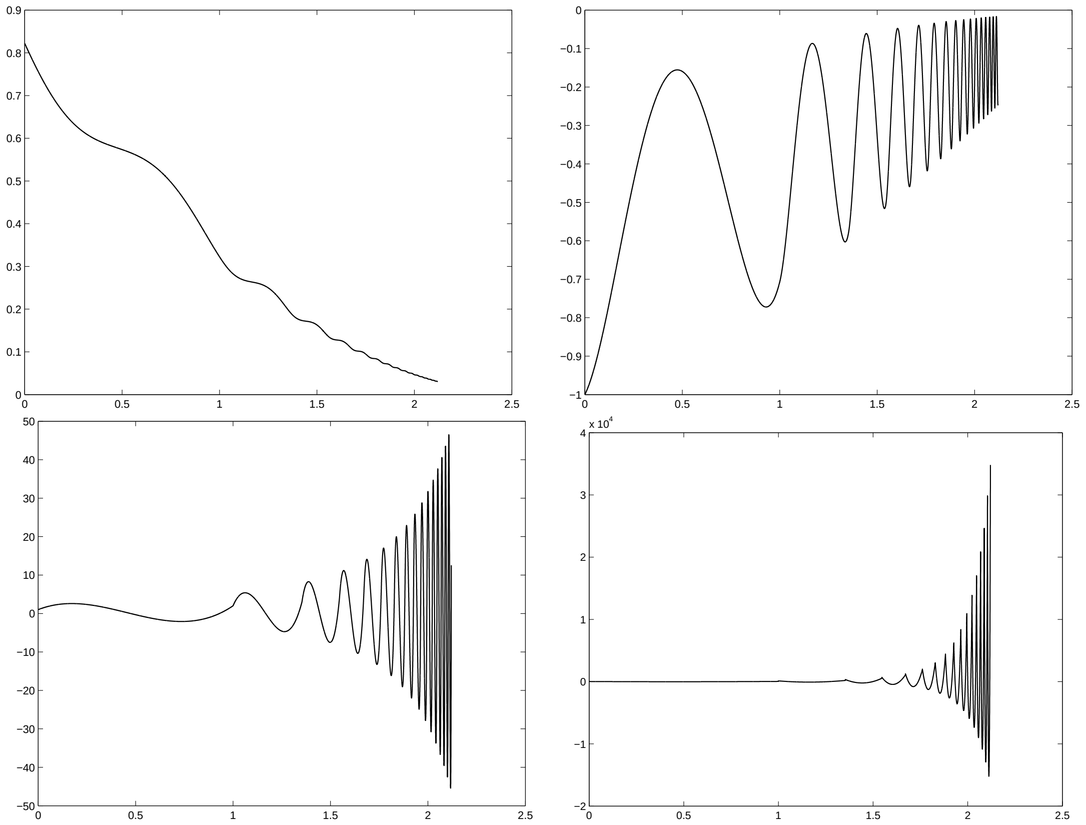

Figure 3.3: The function f_3 and its first three derivatives (from top to bottom and left to right) on the first 16 intervals

Returning to 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 and has uniformly bounded Hessian. Assume also that H_k is positive definite for all k and that the unit step is taken at every iteration of this process. 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 et al., 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

$$f(x_k) - m_k(x_k + s_k) \geq \frac{\kappa \|g_k\|^4}{2 g_k^T H_k g_k} \geq \frac{\kappa \|g_k\|^2}{2\|H_k\|} \geq \frac{\kappa}{2\kappa_H}\|g_k\|^2, \tag{4.1}$$

where we used the Cauchy-Schwartz inequality to deduce the penultimate inequality and where κ_H is an upper bound on the Hessian norms. Because unit steps are taken, we obtain from (2.2) and (4.1) that

$$f(x_0) - f_{\rm low} \geq \kappa \sum_{k=0}^{\infty} f(x_k) - m_k(x_k + s_k) \geq \frac{\kappa}{2\kappa_H} \sum_{k=0}^{\infty} \frac{1}{k+1}, \tag{4.2}$$

where $f_{\rm low}$ is a lower bound on $f(x)$. But this last inequality is impossible because the harmonic series diverges. 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 (4.2) implies that

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

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

$$\lim_{\ell \to \infty} \frac{|\,\mathcal{K} \cap \mathcal{N}_\ell\,|}{10^\ell - |\,\mathcal{K} \cap \mathcal{N}_\ell\,|} = 0, \tag{4.3}$$

where $\mathcal{N}_\ell \stackrel{\rm def}{=} \{p \in {\rm I\!N} \mid 0 \leq p \leq 10^\ell\}$. 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 (4.3),

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

Thus, if $\ell(k)$ is defined k 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

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

where ${\rm Prob}_k[\cdot]$ is the probability with uniform density on $\{10^{\ell(k)-1}+1, \ldots, 10^{\ell(k)}\}$. 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 doe not contradicts the worst-case bound of $O(\epsilon^{-2})$ mentioned above, which is

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

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 developped for Newton's method also turn out to apply for the steepest descent method, as it can also be shown for this case that

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

for some $\kappa_{\rm SD} > 0$ depending on the maximal curvature of the objective function (see, for instance, Conn et al., 2000, Theorem 6.3.3 with Δ_k sufficiently large, or Nesterov, 2004, relation (1.2.13) page 27). This inequality then replaces (4.1) in the above reasoning.

5 Less slow convergence for ARC

Now 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 behaviour for general nonconvex objective functions, i.e. such that one has to perform $O(\epsilon^{-\frac{3}{2}})$ iterations to ensure (2.1). In order to achieve this goal, a suitable condition is now to require 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 (3.1) holds, 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 now unidimensional and we define, for all $k \geq 0$,

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

$$f_{4,0} = \frac{2}{3}\zeta(1+3\eta), \quad f_{4,k+1} = f_{4,k} - \frac{2}{3}\left(\frac{1}{k+1}\right)^{1+3\eta}, \tag{5.3}$$

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

where now

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

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. Using (1.3), this amounts to verifying that

$$g_k^T s_k + s_k^T H_k s_k + \sigma_k \|s_k\|^3 = 0, \tag{5.5}$$

$$s_k^T H_k s_k + \sigma_k \|s_k\|^3 \geq 0, \tag{5.6}$$

$$\sigma_k > 0, \quad \sigma_{k+1} \leq \sigma_k \tag{5.7}$$

and

$$f_4(x_k + s_k) = m_k(x_k + s_k) \tag{5.8}$$

for all $k \geq 1$. Because the model is unidimensional, the first two of these conditions says that the cubic model is globally minimized exactly. Observe first that (5.7) immediately results from (5.4). In our case, (5.5) becomes, using (5.2) and (5.4),

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

as desired, while inequality (5.6) also follows from (5.2) and (5.4). Using (5.3) and (5.4), we also obtain that

$$\begin{aligned} m_k(x_k + s_k) &= f_4(x_k) + g_k^T s_k + \tfrac{1}{2} s_k^T H_k s_k + \tfrac{1}{3}\sigma_k \|s_k\|^3 \\ &= f_4(x_k) - \tfrac{2}{3}\left(\tfrac{1}{k+1}\right)^{1+3\eta} \\ &= f_4(x_{k+1}), \end{aligned}$$

which in turn yields (5.8).

As was the case in the previous sections, the only remaining question is to exhibit 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 may once more consider applying polynomial Hermite interpolation on the interval $[0, x_{k+1}-x_k]$. Thus we are seeking a polynomial of the form (2.11) on the interval $[0, s_k]$ such that

$$p_k(0) = \frac{2}{3}\left(\frac{1}{k+1}\right)^{1+3\eta}, \quad p_k(s_k) = 0 \tag{5.9}$$

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

These conditions immediately give that

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

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

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

whose solution is now given by

$$\begin{pmatrix} c_{3,k} \\ c_{4,k} \\ c_{5,k} \end{pmatrix} = \begin{pmatrix} \frac{10}{3} - 4\phi_k \\ \frac{1}{s_k}[-5 + 7\phi_k] \\ \frac{1}{s_k^2}[2 - 3\phi_k] \end{pmatrix} \tag{5.11}$$

with

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

The definition of ϕ_k implies that $\phi_k \in (0,1)$ for all $k \geq 0$, and hence, using (5.11), that

$$\begin{array}{rcl} |p'''(t)| & = & 6c_{3,k} + 24c_{4,k}t + 60c_{5,k}t^2 \\ & \leq & 6c_{3,k} + 24c_{4,k}s_k + 60c_{5,k}s_k^2 \\ & \leq & 6 \times \frac{10}{3} + 24 \times 13 + 60 \times 2 \\ & = & 452 \end{array} \tag{5.12}$$

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 recursively defined on the nonnegative reals[2] by

$$f_4(x) = p_k(x - x_k) + f_4(x_{k+1}) \;\; \text{for} \;\; x \in [x_k, x_{k+1}] \;\; \text{and} \;\; k \geq 0,$$

and clearly satisfies AS.1. The graph of this function and its first three derivatives are given on the first 16 intervals and for $\eta = 10^{-4}$ by Figure 5.1. This figure confirms the properties

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

of the function $f_4(x)$, namely that it is twice continuous 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 (5.11) and (5.12) 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 $f(x)$. 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 $\zeta(1.0003) \approx 33333.9$).

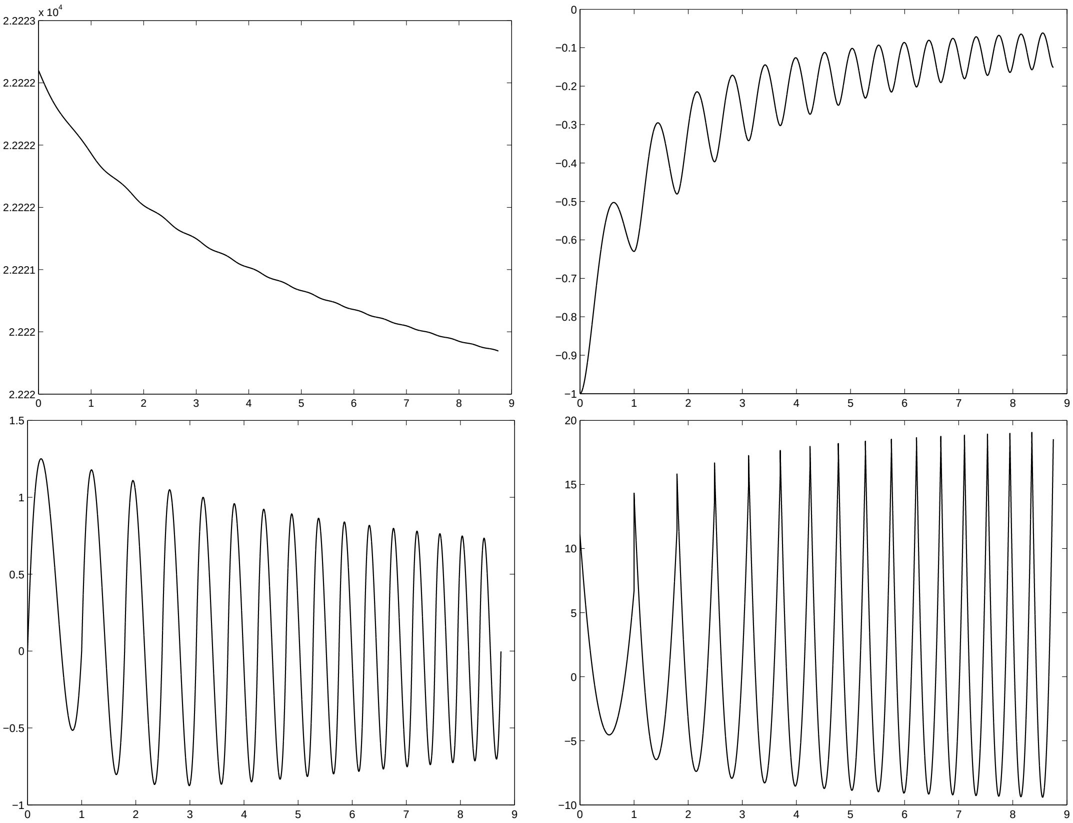

Figure 5.1: The function f_4 and its first three derivatives (from top to bottom and left to right) on the first 16 intervals

6 Conclusions

We now summarize the result obtained in this paper. Considering the steepest method first and assuming Lipzschitz 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 (5.8) hold ensures that our conclusions are also valid if the standard Newton's method is embedded in a trust-region globalization framework (see Conn et al., 2000 for an extensive coverage of such methods), since it guarantees that every iteration is very successful in that case, 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, 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 $\|g_{k-1}\| \leq \epsilon$, but had to give up boundedness of second derivatives to obtain this sharper bound. In addition, we have provided some analysis in an attempt to quantitfy how pessimistic the obtained worst-case bounds can be.

We have then extended the methodology to cover the Adaptive Regularization with Cubics (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.

In our examples for the Newton's and ARC methods, exact global model minimization is carried out, covering the "exact" variants of these algorithms. But the conditions used ((3.6)-(3.7) and (5.5)-(5.6)) only require this exact minimization to occur along the step s_k, which makes the conclusions presented in this paper applicable if one prefers using approximate minimization where the global model minimum is only sought in subspaces, as in the case for truncated conjugate-gradients (see Steihaug, 1983, and Toint, 1981), GLTR (Gould, Lucidi, Roma and Toint, 1999, LSTR and LSRT (Cartis, Gould and Toint, 2009*c*), or for other subspace methods (Ni and Yuan, 1997, Hager, 2001, Erway, Gill and Griffin, 2009). This is however less surprising, as one could expect approximate minimization to deteriorate the global effiency of the minimization algorithm.

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 exluding the (unlikely) possibility that steepest descent 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 mentoned 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.

References

H. Behforooz. Thinning out the harmonic series. *Mathematics Magazine*, **68**(4), 289–293, 1995. As cited on http://www.cut-the-knot.org/arithmetic/algebra/HarmonicSeries.shtml.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case iteration complexity. Technical Report 07/05, Department of Mathematics, FUNDP - University of Namur, Namur, Belgium, 2007.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming, Series A*, 2009*a*. DOI: 10.1007/s10107-009-0286-5, 51 pages.

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, 2009*b*.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Trust-region and other regularisation of linear least-squares problems. *BIT*, **49**(1), 21–53, 2009*c*.

A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. Number 01 *in* 'MPS-SIAM Series on Optimization'. SIAM, Philadelphia, USA, 2000.

J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1983. Reprinted as *Classics in Applied Mathematics 16*, SIAM, Philadelphia, USA, 1996.

J. B. Erway, P. E. Gill, and J. D. Griffin. Iterative methods for finding a trust-region step. *SIAM Journal on Optimization*, **20**(2), 1110–1131, 2009.

R. Fletcher. *Practical Methods of Optimization*. J. Wiley and Sons, Chichester, England, second edn, 1987.

N. I. M. Gould, S. Lucidi, M. Roma, and Ph. L. Toint. Solving the trust-region subproblem using the Lanczos method. *SIAM Journal on Optimization*, **9**(2), 504–525, 1999.

N. I. M. Gould, C. Sainvitu, and Ph. L. Toint. A filter-trust-region method for unconstrained optimization. *SIAM Journal on Optimization*, **16**(2), 341–357, 2005.

S. Gratton, A. Sartenaer, and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, **19**(1), 414–444, 2008.

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.

W. W. Hager. Minimizing a quadratic over a sphere. *SIAM Journal on Optimization*, **12**(1), 188–208, 2001.

Yu. Nesterov. *Introductory Lectures on Convex Optimization*. Applied Optimization. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

Yu. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming, Series A*, **112**(1), 159–181, 2008.

Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, **108**(1), 177–205, 2006.

Q. Ni and Y. Yuan. A subspace limited memory quasi-Newton algorithm for large-scale nonlinear bound constrained optimization. *Mathematics of Computation*, **66**(220), 1509–1520, 1997.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Series in Operations Research. Springer Verlag, Heidelberg, Berlin, New York, 1999.

R. Polyak. Regularized Newton method for unconstrained convex optimization. *Mathematical Programming, Series A*, **120**(1), 125–145, 2009.

T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, **20**(3), 626–637, 1983.

Ph. L. Toint. Towards an efficient sparsity exploiting Newton method for minimization. *in* I. S. Duff, ed., 'Sparse Matrices and Their Uses', pp. 57–88, London, 1981. Academic Press.

K. Ueda. *Regularized Newton Method without Line Search for Unconstrained Optimization*. PhD thesis, 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. Technical Report 2008-015, Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto, Japan, 2008.

K. Ueda and N. Yamashita. 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.

M. Weiser, P. Deuflhard, and B. Erdmann. Affine conjugate adaptive Newton methods for nonlinear elastomechanics. *Optimization Methods and Software*, **22**(3), 413–431, 2007.