Mathematical Programming manuscript No.
(will be inserted by the editor)

Walter F. Mascarenhas

Newton's iterates can converge to non-stationary points

(Manuscript submitted to Mathematical Programming)

Received: date / Revised version: date

Abstract In this note we discuss the convergence of Newton's method for minimization. We present examples in which the Newton iterates satisfy the Wolfe conditions and the Hessian is positive definite at each step and yet the iterates converge to a non-stationary point. These examples answer a question posed by Fletcher in his 1987 book *Practical methods of optimization*.

1. Introduction

The convergence of Newton's method has always been an important subject in nonlinear programming [2]. Following this tradition, in this note we analyze the convergence of Newton's method to minimize functions $f : \mathbb{R}^n \to \mathbb{R}$ of class C^2, i.e. functions with continuous second order derivatives, without constraints. We discuss the convergence of the Newton iterates x_k given by

$$\nabla^2 f(x_k)s_k = -\alpha_k \nabla f(x_k), \tag{1}$$

$$x_{k+1} = x_k + s_k. \tag{2}$$

At the kth step of this method we solve the linear system $\nabla^2 f(x_k)d_k = -\nabla f(x_k)$ for d_k and then pick a positive "step size" α_k such that $s_k = \alpha_k d_k$, f, x_k and x_{k+1} in (2) satisfy the Wolfe conditions

$$f(x_{k+1}) - f(x_k) \leq \eta \nabla f(x_k)^t s_k, \tag{3}$$

$$\nabla f(x_{k+1})^t s_k \geq \beta \nabla f(x_k)^t s_k, \tag{4}$$

for parameters η and β with $0 < \eta < \beta < 1$. In this work we also require that

$$\nabla^2 f(x_k) \text{ is positive definite for all } k. \tag{5}$$

We are particularly interested in the convergence of the sequence $\{x_k\}$ to non-stationary points, i.e. points z with $\nabla f(z) \neq 0$. Our motivation comes from a remark by Fletcher in [2] regarding the relation between the solution of nonlinear systems of the form $F(x) = 0$ and minimization, which we restate as:

W.F. Mascarenhas: Computer Science Department, University of São Paulo, São Paulo, Brazil. e-mail: walterfm@ime.usp.br

Mathematics Subject Classification (1991): 20E28, 20G40, 20C20

If $F'(x_k)$ loses rank in the limit, then convergence of Newton's method for systems of equations to a non-stationary point can occur [5]. The situation may therefore be more severe than with Newton's method for minimization, for which no example satisfying (1)–(5) in which x_k accumulates at a non-stationary point has been developed to our knowledge.

Byrd, Marazzi and Nocedal [1] worked on this question and proved this theorem:

Theorem 1. *Suppose $f \in C^2$ and the sequence $\{x_k\}$ satisfies (1)–(5). If*

(a) $\nabla^2 f(z)$ has rank $n-1$ and
(b) $\nabla f(z) \notin$ Range($\nabla^2 f(z)$)

then x_k is bounded away from z □

They then asked if theorem 1 could be proved under weaker assumptions. In this work we show that both conditions (a) and (b) are needed in theorem 1 and that Newton's method for minimization is as vulnerable as Newton's method for solving equations regarding convergence to non-stationary points. With this aim, we present two examples in which the conditions (1)–(5) are satisfied but yet the Newton iterates converge to a non-stationary point. In the first example the condition (a) is satisfied but (b) is not. In the second example (a) is violated but (b) holds.

The examples involve only two variables. Their iterates are given by simple expressions and their objective functions are highly oscillatory combinations of polynomials, sines and logs. We also found examples in which the objective functions are simple combinations of low degree polynomials and the line searches are exact. They are similar to the example in [3]. However, they involve functions of three variables and the examples presented here are more concise. Actually, the theory developed in [4] suggests that the existence of examples in which the iterates converge to non-stationary points should not surprise us. However, the existence of the examples presented here, with simple explicit expressions for the iterates and objective functions, did surprise us.

Counter examples that solve Fletcher's question are contrived by their very nature. For instance, in a future work we will present a proof of the following lemma:

Lemma 1. *Assume the function f has continuous second order derivatives and the sequence x_k satisfies (1) – (3) and (5). If a subsequence x_{k_i} of x_k converges to z with $\nabla f(z) \neq 0$ then*

$$\limsup_{i\to\infty} \frac{\alpha_{k_i}}{f(x_{k_i}) - f(x_{k_i+1})} \leq \frac{\|\nabla^2 f(z)\|}{\eta \|\nabla f(z)\|^2}$$

and, as a consequence, $\lim_{i\to\infty} \alpha_{k_i} = 0$. □

Since in practice we usually try the step size $\alpha_k = 1$ the reader may be tempted to conclude from lemma 1 that our examples are not "practical". Indeed, they are not practical, but not for this reason: using the interpolation techniques described in [4] we could change f so that $\alpha_k = 1$ does not satisfy the Wolfe conditions and we would be forced to look for smaller $\alpha_k's$. However, by doing that we would obtain more complex examples which are not that practical either, because lemma 1 implies that α_k would underflow quickly and the dynamics of the iterates would be governed by messy rounding errors and not by clean mathematical formulae. Therefore, we believe science is better

served by the simple examples in the next section than by even more contrived and complex examples that would try to cover all gimmicks one may consider in practical implementations of Newton's method.

It is interesting to contrast our examples with Powell's work [6]. At a first glance, it may seem that the choice of the first minimizer along the search lines would resolve the problems posed by our examples. We believe the opposite: criteria like the one in [6] with no hypothesis about the third order derivatives of f only lead to more contrived counter examples, with more variables. This belief is based on [4]. In this reference we present general techniques to produce examples like the ones presented here.

Finally, we would like to mention that although Fletcher's remark may seem to address a specific point about the convergence of Newton's method we believe that a complete answer to the questions posed by Fletcher, Byrd, Marazzi and Nocedal would lead to a better understanding of the convergence of this method in general. We have found improved versions of theorem 1 and we are now working to combine them with the ideas in [7] in order to present an unified discussion of Fletcher's remark, theorem 1 and the choice of the size of the perturbation to the Hessian in regularized versions of Newton's method.

2. The examples

This section presents two examples of C^2 objective functions [1] of two variables and sequences $\{x_k\}$ and α_k that satisfy conditions (1) – (5) and yet $\lim_{k\to\infty} x_k = 0$ and $\nabla f(0) \neq 0$. The iterates x_k are the same in both examples. They are

$$x_k = \begin{pmatrix} 8^{-k} \\ 2^{-k} \end{pmatrix}. \tag{6}$$

The objective functions f have the form

$$f(x) = \phi(\psi(x)), \tag{7}$$

with ϕ and ψ given below. This expression for f is motivated by these equations:

$$\nabla f(x) = \phi'(\psi(x))\nabla\psi(x), \tag{8}$$

$$\nabla^2 f(x) = \phi''(\psi(x))\nabla\psi(x)\nabla\psi(x)^t + \phi'(\psi(x))\nabla^2\psi(x). \tag{9}$$

Notice that if we define

$$\alpha_k = -\frac{\phi''(\psi(x_k))}{\phi'(\psi(x_k))}\nabla\psi(x_k)^t s_k, \tag{10}$$

then equation (1) for the Newton step is satisfied if $\nabla^2\psi(x_k)s_k = 0$. Equations (7)–(10) allow the analysis of the effects of the addition to ψ of highly oscillatory functions like μ given by

$$\mu(0) = 0 \quad \text{and} \quad \mu(b) = \frac{b^3(\ln 2)^2}{8\pi^2}\sin^2\left(\pi\frac{\ln(b^2)}{\ln 2}\right) \quad \text{for } b \neq 0. \tag{11}$$

[1] Actually, the second derivatives of these objective functions are locally Lipschitz continuous in $\mathbb{R}^2$ and analytic outside the x axis.

This function has continuous second order derivatives at all $b \in \mathbb{R}$ and if we consider a C^2 function $h : \mathbb{R} \to \mathbb{R}$ and add $h(b)\mu(b)$ to $\psi(a,b)$ then $\nabla^2\psi(x_k)$ changes by

$$h(2^{-k})2^{-k}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \tag{12}$$

but $\psi(x_k)$ and $\nabla\psi(x_k)$ do not change, because $\mu(0) = \mu'(0) = \mu''(0) = 0$,

$$\mu(2^{-k}) = \mu'(2^{-k}) = 0 \qquad \text{and} \qquad \mu''(2^{-k}) = 2^{-k} \quad \text{for all } k \in \mathbb{N}. \tag{13}$$

Our study of such functions $h(b)\mu(b)$ lead to these ψ's for the examples 1 and 2:

$$\psi_1(a,b) = a + b^3 - 28\,b^3\,a + 24\,a^2 + 3\left(105\,b^3 - 2\right)\mu(b), \tag{14}$$
$$\psi_2(a,b) = a + b^3 - 11\,b^4\,a + 8\,b\,a^2 + \left(181\,b^4 - 6\right)\mu(b). \tag{15}$$

Finally, we chose these functions ϕ:

$$\phi_1(a) = a + a^2 \qquad \text{and} \qquad \phi_2(a) = a + a^3. \tag{16}$$

The definition of the examples is now complete and it is summarized in table 1. From here to the end of this section we validate them. We compute the terms in condi-

Table 1. The two examples

Example	x_k	f	ψ	$\phi(a)$	α_k	$\nabla f(0)$	$\nabla^2 f(0)$
1	$\begin{pmatrix} 8^{-k} \\ 2^{-k} \end{pmatrix}$	$\phi(\psi(x))$	see (11)–(14)	$a + a^2$	see (10)	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 50 & 0 \\ 0 & 0 \end{pmatrix}$
2	$\begin{pmatrix} 8^{-k} \\ 2^{-k} \end{pmatrix}$	$\phi(\psi(x))$	see (11)–(15)	$a + a^3$	see (10)	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$

tions (1)–(5) explicitly and show that they are satisfied. We assume that the indexes k start at $k = 1000$. As a consequence, $\varepsilon = 2^k \leq 2^{-1000}$ for all k of interest. This ε is tiny and we suggest that, in order to get the gist of our examples, you look only at leading order terms in ε in the expression below in a first reading. After you have a global view of why these expressions make sense it will be easier to come back and fill in the details regarding the higher order terms. We begin the verification of the examples by using (6)

and (12)–(15) to deduce that, for $x_k = (a,b) = (\varepsilon^3, \varepsilon)^t = (2^{-3k}, 2^{-k})^t$,

$$s_k = x_{k+1} - x_k = -\frac{\varepsilon}{8}\begin{pmatrix} 7\varepsilon^2 \\ 4 \end{pmatrix}, \tag{17}$$

$$\psi_1(x_k) = 2(1-2\varepsilon^3)\varepsilon^3, \tag{18}$$

$$\nabla\psi_1(x_k) = \begin{pmatrix} 1+20\varepsilon^3 \\ 3(1-28\varepsilon^3)\varepsilon^2 \end{pmatrix}, \tag{19}$$

$$\nabla^2\psi_1(x_k) = \begin{pmatrix} 48 & -84\varepsilon^2 \\ -84\varepsilon^2 & 147\varepsilon^4 \end{pmatrix}, \tag{20}$$

$$\psi_2(x_k) = (2-3\varepsilon^4)\varepsilon^3, \tag{21}$$

$$\nabla\psi_2(x_k) = \begin{pmatrix} 1+5\varepsilon^4 \\ 3(1-12\varepsilon^4)\varepsilon^2 \end{pmatrix}, \tag{22}$$

$$\nabla^2\psi_2(x_k) = \varepsilon\begin{pmatrix} 16 & -28\varepsilon^2 \\ -28\varepsilon^2 & 49\varepsilon^4 \end{pmatrix}. \tag{23}$$

The evaluation of (20) $\times$ (17) and (23) $\times$ (17) show that $\nabla^2\psi(x_k)s_k = 0$ in both examples. Thus, (9) and (10) imply that the iterates follow the Newton steps in (1) (✓).

To check the first Wolfe condition (3) and verify that α_k is positive, we use the bound $\varepsilon \le 2^{-1000}$ and equations (18) and (21) to obtain

$$\psi_1(x_{k+1}) - \psi_1(x_k) = 7\frac{9\varepsilon^3 - 4}{16}\varepsilon^3 < -\frac{7}{4}\varepsilon^3(1-\varepsilon) < 0, \tag{24}$$

$$\psi_2(x_{k+1}) - \psi_2(x_k) = \frac{381\varepsilon^4 - 224}{128}\varepsilon^3 < -\frac{7}{4}\varepsilon^3(1-\varepsilon) < 0. \tag{25}$$

The definition of ϕ_1 and ϕ_2 in (16) show that

$$\phi_1''(\xi) = 2 > 0 \qquad \text{for} \qquad \xi \in \mathbb{R}, \tag{26}$$

$$\phi_2''(\xi) = 6\xi > 0 \qquad \text{for} \qquad \xi > 0. \tag{27}$$

Equations (18) and (21) and $\varepsilon \le 2^{-1000}$ lead to $\psi_1(x_k) > 0$ and $\psi_2(x_k) > 0$. The bounds (24)–(25) show that $\psi_1(x_k)$ and $\psi_2(x_k)$ decrease with k and, since the second derivatives in (26)–(27) are positive, (16) yields

$$1 \le \phi_1'(\psi_1(x_k)) \le \phi_1'(\psi_1(x_1)) = 1 + 2\psi_1(x_1) = \frac{11}{8}, \tag{28}$$

$$1 \le \phi_2'(\psi_2(x_k)) \le \phi_2'(\psi_2(x_1)) = 1 + 3\psi_2(x_1)^2 = \frac{18907}{16384} < \frac{6}{5}. \tag{29}$$

Equations (17), (19), (22) and $\varepsilon \le 2^{-1000}$ imply that

$$-\frac{19}{8}\varepsilon^3(1+\varepsilon) < \nabla\psi_1(x_k)^t s_k = \frac{196\varepsilon^3 - 19}{8}\varepsilon^3 < -\frac{19}{8}\varepsilon^3(1-\varepsilon) < 0, \tag{30}$$

$$-\frac{19}{8}\varepsilon^3(1+\varepsilon) < \nabla\psi_2(x_k)^t s_k = \frac{109\,\varepsilon^4 - 19}{8}\varepsilon^3 < -\frac{19}{8}\varepsilon^3(1-\varepsilon) < 0. \tag{31}$$

Thus, (26)–(31) show that α_k defined in (10) is positive in both examples (✓).

Combining (8) and (28)–(31) we obtain

$$-\frac{11\times 19}{64}(1+\varepsilon)\varepsilon^3 < \nabla f_1(x_k)^t s_k < -\frac{19}{8}\varepsilon^3(1-\varepsilon), \tag{32}$$

$$-\frac{3\times 19}{20}(1+\varepsilon)\varepsilon^3 < \nabla f_2(x_k)^t s_k < -\frac{19}{8}\varepsilon^3(1-\varepsilon). \tag{33}$$

The last two equations, $\varepsilon \leq 2^{-1000}$ and (24) – (25) imply that

$$\psi_1(x_{k+1}) - \psi_1(x_k) < \frac{7\times 64\times(1-\varepsilon)}{4\times 11\times 19\times(1+\varepsilon)}\nabla f_1(x_k)^t s_k < \frac{1}{2}\nabla f_1(x_k)^t s_k,$$

$$\psi_2(x_{k+1}) - \psi_2(x_k) < \frac{7\times 20\times(1-\varepsilon)}{4\times 3\times 19\times(1+\varepsilon)}\nabla f_2(x_k)^t s_k < \frac{1}{2}\nabla f_2(x_k)^t s_k,$$

because

$$\frac{7\times 64}{4\times 11\times 19} \approx 1.13 \qquad \text{and} \qquad \frac{7\times 20}{4\times 3\times 19} \approx 0.61$$

are both bigger than $1/2$ and $\nabla f_1(x_k)^t s_k$ and $\nabla f_2(x_k)^t s_k$ are negative. Therefore, the examples satisfy the first Wolfe condition (3) with $\eta = 1/2$ (✓). To verify the second Wolfe condition (4), we replace ε by $\varepsilon/2$ in (19) and (22) to compute $\nabla\psi(x_{k+1})$:

$$\nabla\psi_1(x_{k+1}) = \begin{pmatrix} 1+\frac{5}{2}\varepsilon^3 \\ \frac{3}{8}(2-7\varepsilon^3)\varepsilon^2 \end{pmatrix} \qquad \text{and} \qquad \nabla\psi_2(x_{k+1}) = \begin{pmatrix} 1+\frac{5}{16}\varepsilon^4 \\ \frac{3}{16}(4-3\varepsilon^4)\varepsilon^2 \end{pmatrix}.$$

We then recall that $\varepsilon \leq 2^{-1000}$ and use (8), (17), (28) and (29) to obtain

$$\nabla f_1(x_{k+1})^t s_k = \phi_1'(x_{k+1})\nabla\psi_1(x_{k+1})^t s_k = -\phi_1'(x_{k+1})\frac{10+7\varepsilon^3}{8}\varepsilon^3 > -\frac{110}{64}\varepsilon^3(1+\varepsilon), \tag{34}$$

$$\nabla f_2(x_{k+1})^t s_k = \phi_2'(x_{k+1})\nabla\psi_2(x_{k+1})^t s_k = \phi_2'(x_{k+1})\frac{\varepsilon^4-160}{128}\varepsilon^3 > -\frac{3}{2}\varepsilon^3. \tag{35}$$

Therefore, (32) and (33) imply $\nabla f(x_{k+1})^t s_k > 3/4\nabla f(x_k)^t s_k$ in both examples. Since $\beta = 3/4 > \eta = 1/2$, we have verified the second Wolfe condition (4) (✓).

We now prove that the hessians $\nabla^2 f(x_k)$ are positive definite. In order to simplify the algebra, notice that (9) yields

$$\nabla^2 f(x_k) = \phi'(\psi(x_k))\left(\nabla^2\psi(x_k) + \frac{\phi''(\psi(x_k))}{\phi'(\psi(x_k))}\nabla\psi(x_k)\nabla\psi(x_k)^t\right).$$

Since (26)–(29) show that $\phi_i'(\psi(x_k)) > 0$, $\frac{\phi_1''(\psi(x_k))}{\phi_1'(\psi(x_k))} > 1$ and $\frac{\phi_2''(\psi(x_k))}{\phi_2'(\psi(x_k))} > 8\varepsilon^3$, the positivity of $\nabla^2 f_1(x_k)$ and $\nabla^2 f_2(x_k)$ follow from the positivity of these matrices:

$$A_{1k} = \nabla^2\psi_1(x_k) + \nabla\psi_1(x_k)\nabla\psi_1(x_k)^t \quad \text{and} \quad A_{2k} = \nabla^2\psi_2(x_k) + 8\varepsilon^3\nabla\psi_2(x_k)\nabla\psi_2(x_k)^t,$$

which we take the liberty to write as

$$A_{1k} = \begin{pmatrix} 49+O(\varepsilon^3) & -81\varepsilon^2+O(\varepsilon^3) \\ -81\varepsilon^2+O(\varepsilon^3) & 156\varepsilon^4+O(\varepsilon^5) \end{pmatrix},$$

$$A_{2k} = \varepsilon\begin{pmatrix} 16+8\varepsilon^2+O(\varepsilon^3) & -28\varepsilon^2+24\varepsilon^4+O(\varepsilon^5) \\ -28\varepsilon^2+24\varepsilon^4+O(\varepsilon^5) & 49\varepsilon^4+72\varepsilon^6+O(\varepsilon^7) \end{pmatrix}.$$

A bit of algebra shows that

$$\begin{aligned} \text{Trace}(A_{1k}) &= 49 + O(\varepsilon^3), & \text{Det}(A_{1k}) &= 1083\varepsilon^4 + O(\varepsilon^5), \\ \text{Trace}(A_{2k}) &= 16\varepsilon + O(\varepsilon^3), & \text{Det}(A_{2k}) &= 2888\varepsilon^8 + O(\varepsilon^9). \end{aligned}$$

Thus, the Hessians are positive and the condition (5) is satisfied (✓). The four check marks above indicate that we have verified that α_k is positive and the examples satisfy the conditions (1)–(5) for $\eta = 1/2$ and $\beta = 3/4$. Therefore, the examples are valid.

References

1. R. Byrd, M. Marazzi and J. Nocedal, On the convergence of Newton Iterations to non-stationary points, Math. Prog. **99**, 127-148, 2004.
2. R. Fletcher, *Practical methods of optimization*, J. Wiley and Sons, Chichester, England, second edition, 1987.
3. W.F.Mascarenhas, The BFGS method with exact line searches fails for non-convex objective functions. Math. Prog. **99**, 49-61 , 2004.
4. W.F.Mascarenhas, On the divergence of line search methods, Manuscript submited to publication at the journal Computational and Applied Mathematics. Preliminary version available online at www.ime.usp.br/~walterfm/pub/divergence.pdf
5. Powell, M.J.D.: A hybrid method for nonlinear equations. In: P. Rabinowitz, editor, Numerical Methods for Nonlinear Algebraic Equations, pages 87114, London, 1970. Gordon and Breach
6. Powell, M.J.D.: On the convergence of the DFP algorithm for unconstrained optimization when there are only two variables, Math. Program., Ser. B, 87, 281-301, 2000.
7. Y. Nesterov, B. Polyak, A cubic regularization of a Newton scheme and its global performance. Math. Prog. **108**, 177-205, 2006.