Math. Program., Ser. A 91: 349–356 (2002)

Digital Object Identifier (DOI) 10.1007/s101070100249

Chong Li · Xinghua Wang

On convergence of the Gauss-Newton method for convex composite optimization*

Received: July 24, 1998 / Accepted: November 29, 2000
Published online September 3, 2001 – © Springer-Verlag 2001

Abstract. The local quadratic convergence of the Gauss-Newton method for convex composite optimization $f = h \circ F$ is established for any convex function h with the minima set C, extending Burke and Ferris' results in the case when C is a set of weak sharp minima for h.

Key words. Gauss-Newton method – weak sharp minima – regular point – quadratic convergence

1. Introduction

The famous Gauss-Newton method, which was proposed for finding the least-squares solutions of nonlinear equations by Gauss in the early nineteenth century, is extended to solve the following convex composite optimization:

$$(P) \qquad \min f(x) := h(F(x)),$$

where $h : R^m \to R$ is convex and $F : R^n \to R^m$ is continuously differentiable. This problem has recently been studied in [2,5] and the cited references there, and justifiable so since many class of problems in optimization theory can be cast within this framework, e.g., convex inclusion, minimax problems, penalization methods and goal programming. Moreover, this model provides a unifying framework for the development and analysis of algorithmic solution techniques.

In 1985, Womersley [16] proved the local quadratic convergence of Gauss-Newton methods under the assumption of strong uniqueness, extending the work of Jittorntrum and Osborne [9] for the case when h is a norm. However, Womersley's assumption just ensures the Gauss-Newton sequence converges to a local minima of (P). Recently, Burke and Ferris [5] have made a great progress in the study of convergence of Gauss-Newton methods. An important distinction from Womersley is that they do not require that the minima set for h be a singleton or even a bounded set. Their research is based on two

C. Li: Department of Applied Mathematics, Southeast University, Nanjing 210096, P.R. China,
e-mail: cli@seu.edu.cn

X. Wang: Department of Mathematics, Zhejiang University, Hangzhou 310028, P.R. China

Mathematics Subject Classification (2000): 90C30, 65K10

* Supported by the National (Grant No 19971013) and Jiangsu Provincial (Grant No BK99001) Natural Science Foundation of China

assumptions: (1) the set of minima for the function h, denoted by C, is a weak sharp minima for h, and (2) there is a regular point for the inclusion

$$F(x) \in C. \tag{1}$$

Under the above assumptions, they established the local quadratic convergence of the Gauss-Newton sequence. In addition, the convergence properties of a globalization strategy based on a backtracking linear- search were also provided.

Based on the work of Burke and Ferris, we continue carrying out investigation in this direction. It is unexpected that we find that the local quadratic convergence of the method is independent of the other properties of the convex function h. The purpose of this paper is to establish the local quadratic convergence of the Gauss-Newton method for an arbitrary convex function h and the quadratic convergence of globalization strategies under a much weaker assumption than Burke and Ferris'. Moreover, we also propose a relaxation version of the Gauss-Newton method and establish the local superlinear convergence.

We need some notations. The polar of $U \subset R^n$ is the set $U^\circ := \{x^* \in R^n :< x^*, x > \leq 1, \forall x \in U\}$ while cone$U$ stands for the cone generated by U. Let $\|\cdot\|$ denote a norm on R^n and B be the closed unit ball so that $x + rB$ is the closed ball with center x radius r. The distance of a point x to a set U is given by $d(x, U) := \inf\{\|x - u\| : u \in U\}$ and the set of all nearest points to x is denoted by $P_U(x)$. Finally, the set ker$A$ represents the kernel of the linear map A.

2. Preliminaries

For $\Delta > 0$ and $x \in R^n$, let $D_\Delta(x)$ represent the set of solutions to the minimization problem

$$\min\{h(F(x) + F'(x)d) : \|d\| \leq \Delta\}. \tag{2}$$

The basic algorithm considered in [5,8,16] is as follows.

Algorithm 1. Let $\eta \geq 1$, $\Delta \in (0, +\infty]$ and $x^0 \in R^n$ be given. For $k = 0, 1, \cdots$, having x^k, determine x^{k+1} as follows.

If $h(F(x^k)) = \min\{h(F(x^k) + F'(x^k)d) : \|d\| \leq \Delta\}$, then stop; otherwise, choose $d^k \in D_\Delta(x^k)$ to satisfy $\|d^k\| \leq \eta d(0, D_\Delta(x^k))$, and set $x^{k+1} = x^k + d^k$.

Definition 1. *The set $C \subset R^m$ is a set of weak sharp minima for h if there is $\lambda > 0$ such that $\forall y \in R^m$*

$$h(y) \geq h_{\min} + \lambda d(y, C),$$

where $h_{\min} = \min_y h(y)$.

There are many examples of convex functions that have a set of weak sharp minima, see for example [4,7]. This notion generalizes the notion of a sharp [12] or strongly unique [6,11,16]. Their applications in the convergence analysis can be found in [4,6,9, 12,16]. In [4], Burke and Ferris also provided some sufficient and necessary conditions

for a minima set for a convex function to be a set of weak sharp minima. However, most of the convex functions on R^m do not have a set of weak sharp minima. Such an important class of convex functions is the class of convex functions which are Gateaux differentiable in every direction on R^m.

Definition 2. *A point $\bar{x} \in R^n$ is a regular point for the inclusion* (1) *if*

$$\ker(F'(\bar{x})^T) \cap \Gamma_C(F(\bar{x})) = \{0\},$$

where the multifunction $\Gamma_C : R^m \to R^m$ is given by

$$\Gamma_C(y) = (\mathrm{cone}(C - y))^0, \quad \forall y \in R^m.$$

The notion of regularity is related to various notions of regularity that can be found in the papers [1,3,13–15], which have played an important role in the study of nonsmooth optimizations. Some equivalent conditions on the regular points for (1) are given in paper [5]. The following proposition is useful to the convergence analysis of the Gauss-Newton method.

Proposition 1 [5]. *If $\bar{x}$ is a regular point of* (1)*, then for $\forall \Delta > d(0, D_{+\infty}(\bar{x}))$, there is some neighborhood $N(\bar{x})$ of $\bar{x}$ and a $\beta > 0$ satisfying*

$$d(0, D_\Delta(x)) \le \beta d(F(x), C), \qquad \forall x \in N(\bar{x})$$

and

$$\{d \in R^n : \|d\| \le \Delta, F(x) + F'(x)d \in C\} \ne \emptyset, \qquad \forall x \in N(\bar{x}).$$

3. Quadratic convergence

In [5], Burke and Ferris proved the quadratic convergence theorem under the assumption that h has a set of weak sharp minima and is Lipschitz continuous [5, Theorem 4.1]. But the comments of last section show that the assumption that h has a set of weak sharp minima is strong. The key difference of our approach to Burke and Ferris' is that we do not require that h have a set of weak sharp minima. The main theorem is stated as follows.

Theorem 1. *Let $\bar{x} \in R^n$ be a regular point of inclusion* (1) *where C is a minima set for h and suppose the conclusions of Proposition 1 are satisfied on the set $\bar{x} + \bar{\delta}B$ for $\bar{\delta} > 0$, with $\bar{\delta} < \Delta$. Assume that F' is Lipschitz continuous on $\bar{x} + \bar{\delta}B$ with Lipschitz constant L. If there is $\delta > 0$ such that*

a) $\delta < \min\{\bar{\delta}/2, 1\}$,
b) $d(F(\bar{x}), C) < \delta/2\eta\beta$ *and*
c) $\eta L \delta \beta < 2$,

then there is a neighborhood $M(\bar{x})$ of $\bar{x}$ such that the sequence $\{x^k\}$ generated by Algorithm 1 with initial point in $M(\bar{x})$ converges at a quadratic rate to some x^* with $F(x^*) \in C$, that is x^* solves (P).

Proof. Let L_0 be the Lipschitz constant for F on $\bar{x} + \delta B$ and define

$$\bar{\beta} = \frac{\delta - 2\eta\beta d(F(\bar{x}), C)}{2\eta\beta L_0}, \qquad r_0 = \min\{\delta, \bar{\beta}\}.$$

Then for $\forall x^0 \in \bar{x} + r_0 B$, we have

$$d(F(x^0), C) \le \|F(x^0) - F(\bar{x})\| + d(F(\bar{x}), C) \le L_0\bar{\beta} + d(F(\bar{x}), C) \le \frac{\delta}{2\eta\beta}.$$

It follows from Proposition 1 that

$$\|d^0\| \le \eta\beta d(F(x^0), C) \le \frac{\delta}{2}$$

and

$$\|x^1 - \bar{x}\| \le \|d^0\| + \|x^0 - \bar{x}\| \le \bar{\delta}.$$

We claim that for $k = 0, 1, 2, \cdots$,

$$\|x^k - \bar{x}\| \le \bar{\delta} \quad \text{and} \quad \|d^k\| \le \frac{1}{2}\eta L\beta\|d^{k-1}\|^2 \le \left(\frac{1}{2}\right)^{2^k}\delta. \tag{3}$$

The proof proceeds by induction on k.

Note that (3) holds for $k = 0$. Assume that (3) holds for $k \le s - 1$. Then, for $k = s$,

$$\|x^s - \bar{x}\| \le \sum_{i=0}^{s-1}\|d^i\| + \|x^0 - \bar{x}\| \le \delta\sum_{i=0}^{s-1}\left(\frac{1}{2}\right)^{2^i} + \delta \le 2\delta \le \bar{\delta}$$

and so, using Proposition 1 again, we have

$$\begin{aligned}
\|d^s\| &\le \eta\beta d(F(x^s), C)\\
&\le \eta\beta\|F(x^s) - F(x^{s-1}) - F'(x^{s-1})d^{s-1}\|\\
&\le \eta\beta\|\int_0^1 (F'(x^{s-1} + td^{s-1}) - F'(x^{s-1}))d^{s-1}\mathrm{d}t\|\\
&\le \eta\beta\int_0^1 Lt\|d^{s-1}\|^2\mathrm{d}t\\
&= \frac{1}{2}\eta\beta L\|d^{s-1}\|^2\\
&\le \delta\left(\frac{1}{2}\right)^{2^s}.
\end{aligned}$$

Hence (3) holds for any $k = 0, 1, \cdots$. This means x^k converges to some x^* at a quadratic rate and proves the theorem.

□

Now let us consider the quadratic convergence of the global algorithm proposed by Burke and Ferris in [5]. The algorithm is simply stated as follows.

Algorithm 2. Let $\eta \geq 1$, $\Delta \in (0, +\infty]$, $c \in (0, 1)$, $\gamma \in (0, 1)$ and $x^0 \in R^n$ be given. For $k = 0, 1, \cdots$, having x^k determine x^{k+1} as follows.

i) If $h(F(x^k)) = \min\{h(F(x^k) + F'(x^k)d) : \|d\| \leq \Delta\}$, then stop; otherwise, choose $d^k \in D_\Delta(x^k)$ to satisfy $\|d^k\| \leq \eta d(0, D_\Delta(x^k))$.

ii) Set $x^{k+1} = x^k + t_k d^k$ where t_k is the maximum value of γ^s, for $s = 1, 2, \cdots$, such that

$$h(F(x^k + \gamma^s d^k)) - h(F(x^k)) \leq c\gamma^s [h(F(x^k) + F'(x^k)d^k) - h(F(x^k))].$$

Thus, because of Theorem 1, the same arguments as in the proof of Theorem 5.2 in [5] give the proof of the following theorem.

Theorem 2. *Suppose F' is locally Lipschitz continuous and C is a minima set for h satisfying*

$$\lim_{d(y,C)\to 0} \frac{h(y) - h_{\min}}{d(y, C)^2} = +\infty. \tag{4}$$

Let $\{x^k\}$ be a sequence generated by Algorithm 2 with initial point in R^n and $\bar{x}$ be a cluster point of this sequence. If $\Delta < +\infty$ and $\bar{x}$ is a regular point of (1)*, then $F(\bar{x}) \in C$ and x^k converges to $\bar{x}$ at a quadratic rate.*

Remark. Burke and Ferris [5] proved Theorem 2 under the assumption that C is a set of weak sharp minima for h. Obviously, condition (4) is much weaker than it. One important class of convex functions such that (4) holds is the class of convex functions h that have a set C of weak sharp minima of order s with $1 \leq s < 2$ in the sense that there exist $\epsilon > 0$ and $\alpha_s > 0$ satisfying

$$h(y) \geq h_{\min} + \alpha_s d(y, C)^s, \quad \forall y \in R^m, d(y, C) \leq \epsilon.$$

This notion generalizes the notion of the strong uniqueness of order s in approximation theory [10]. Clearly it is also a generalization of the concept of weak sharp minima. For two convex functions h_1 and h_2 on R^m with the same minima set C, it follows from the definition that the function $h = h_1 + h_2$ has a set of weak sharp minima of order $s \geq 1$ if at least one of h_1 and h_2 does. The following is an example of convex function h that has a set of weak sharp minima of order $1 < s < 2$ but not a set of weak sharp minima.

Example 1. Let $1 < s < 2$ and $1 = s_1 \leq s_2 \leq \cdots \leq s_m = s$. Define $h : R^m \longrightarrow R$ as follows.

$$h(y) = \sum_{i=1}^{m} |y_i|^{s_i}, \quad \forall y = (y_i) \in R^m.$$

Then the minima set $C = \{0\}$ for h is a set of weak sharp minima of order s for h since it is for the function $h_m(y) := |y_m|$ but it is not a set of weak sharp minima for h. In fact, C is not a set of weak sharp minima of order $\bar{s}$ for any $\bar{s} < s$.

4. A relaxation algorithm

Let us return to the Gauss-Newton method, that is, Algorithm 1. In the iterative procedure, we must determine the exact solution of the convex optimization problem (2). However, this is, in practice, very difficult and impossible for the most cases. Thus, the question arises: Does the Gauss-Newton sequence remain the local quadratic convergence if the correction term d^k is replaced by an approximation of the exact solution to the problem (2)? The purpose of this section is to establish such a result. To this end, we propose a relaxation version of the Gauss-Newton algorithm as follows.

Let $D^k_\Delta(x^k)$ represent the set of all $d \in R^n$ satisfying $\|d\| \le \Delta$ and

$$h(F(x^k) + F'(x^k)d) \le \min\{h(F(x^k) + F'(x^k)d) : \|d\| \le \Delta\} + \|d^{k-1}\|^\alpha.$$

Relaxation Algorithm 3. Let $\eta \ge 1, \alpha > 1, \Delta \in (0, +\infty]$, $x^0 \in R^n$ and $d^{-1} \in R^n$, $d^{-1} \ne 0$ be given. For $k = 0, 1, \cdots$, having x^k, then determine x^k as follows.

i) if

$$h(F(x^k)) \le \min\{h(F(x^k) + F'(x^k)d) : \|d\| \le \Delta\} + \|d^{k-1}\|^\alpha, \tag{5}$$

take $d^k = \|d^{k-1}\|^{\alpha-1}d^{k-1}$; otherwise, choose $d^k \in D^k_\Delta(x^k)$ to satisfy $\|d^k\| \le \eta d(0, D^k_\Delta(x^k))$.

ii) set $x^{k+1} = x^k + d^k$.

It should be noted that, in the step **i** of the algorithm, we need compute $\min\{h(F(x^k) + F'(x^k)d) : \|d\| \le \Delta\}$. There are a lots of ways to do this, for example, the subgradient method, the cutting plane method, the bundle method, etc [17]. However, in fact, we need not the exact value of $\min\{h(F(x^k) + F'(x^k)d) : \|d\| \le \Delta\}$ but only to distinguish (5) holds or not.

Theorem 3. *Let $\bar{x} \in R^n$ be a regular point of inclusion* (1) *where C is a weak sharp minima set for h and suppose the conclusions of Proposition 1 are satisfied on the set $\bar{x} + \bar{\delta}B$ for $\bar{\delta} > 0$, with $\bar{\delta} < \Delta$. Assume that F' is Lipschitz continuous on $\bar{x} + \bar{\delta}B$ with Lipschitz constant L. If there is $\delta > 0$ such that*

a) $\delta < \min\{\bar{\delta}/(c+1), 1\}$,
b) $d(F(\bar{x}), C) < \delta/2\eta\beta$ *and*
c) $\eta L\delta\beta/2 + \eta\beta\delta^{\alpha-1}/\lambda < 1$,

where $p = \min\{2, \sqrt{\alpha}\}$ and $c = \sum_{i=0}^{+\infty}(\frac{1}{2})^{p^i}$, then there is a neighborhood $M(\bar{x})$ of $\bar{x}$ such that the sequence $\{x^k\}$ generated by Algorithm 3 with initial point in $M(\bar{x})$, with $\|d^{-1}\| \le \delta/2$, converges at a rate of p degree to some x^* with $F(x^*) \in C$, that is x^* solves (P).

Proof. Let $L_0, \bar{\beta}, r_0$ be as in the proof of Theorem 1. Then for $\forall x^0 \in \bar{x} + r_0 B$, we have

$$d(F(x^0), C) \le \|F(x^0) - F(\bar{x})\| + d(F(\bar{x}), C) \le L_0\bar{\beta} + d(F(\bar{x}), C) \le \frac{\delta}{2\eta\beta}.$$

By Relaxation Algorithm 3, if (5) holds for $k = 0$, then

$$\|d^0\| = \|d^{-1}\|^\alpha \le \frac{\delta}{2};$$

otherwise,

$$\|d^0\| \le \eta d\big(0, D^0_\Delta(x^0)\big).$$

Observe that $D_\Delta(x^0) \subset D^0_\Delta(x^0)$. Thus, in the second case, from Proposition 1, we also have that

$$\|d^0\| \le \eta d\big(0, D_\Delta(x^0)\big) \le \eta\beta d(F(x^0), C) \le \frac{\delta}{2}.$$

Hence,

$$\|x^1 - \bar{x}\| \le \|d^0\| + \|x^0 - \bar{x}\| \le \bar{\delta}.$$

In the following, we will establish by induction that for $k = 0, 1, 2, \cdots,$

$$\|x^k - \bar{x}\| \le \bar{\delta} \quad \text{and} \quad \|d^k\| \le \left(\frac{1}{2}\right)^{p^k} \delta. \tag{6}$$

Note that (6) holds for $k = 0$. Assume that (6) holds for $k \le l - 1$. Then, for $k = l$,

$$\|x^l - \bar{x}\| \le \sum_{i=0}^{l-1} \|d^i\| + \|x^0 - \bar{x}\| \le \delta \sum_{i=0}^{l-1} \left(\frac{1}{2}\right)^{p^i} + \delta \le (c+1)\delta \le \bar{\delta}.$$

Applying Relaxation Algorithm 3, if (5) holds for $k = l$, we have

$$\|d^l = \|d^{l-1}\|^\alpha \le \left(\frac{1}{2}\right)^{p^l} \delta,$$

and otherwise,

$$\|d^l\| \le \eta d\big(0, D^l_\Delta(x^l)\big) \le d\big(0, D_\Delta(x^l)\big)$$

since $D_\Delta(x^l) \subset D^l_\Delta(x^l)$. Then it follows from Proposition 1 that

$$\begin{aligned}
\|d^l\| &\le \eta\beta d(F(x^l), C) \\
&\le \eta\beta d(F(x^l), C_{l-1}) + \eta\beta \sup_{x \in C_{l-1}} d(x, C) \\
&\le \eta\beta \|F(x^l) - F(x^{l-1}) - F'(x^{l-1})d^{l-1}\| + \eta\beta \sup_{x \in C_{l-1}} d(x, C) \\
&= \frac{1}{2}\eta\beta L\|d^{l-1}\|^2 + \frac{\eta\beta}{\lambda}\|d^{l-2}\|^\alpha \\
&\le \frac{1}{2}\eta\beta L\left(\left(\frac{1}{2}\right)^{p^{l-1}} \delta\right)^2 + \frac{\eta\beta}{\lambda}\left(\left(\frac{1}{2}\right)^{p^{l-2}} \delta\right)^\alpha \\
&\le \left(\frac{1}{2}\right)^{p^l} \delta,
\end{aligned}$$

where

$$C_k = \{y \in R^m : h(y) \leq h_{\min} + \|d^{k-1}\|^\alpha\}.$$

Hence (6) holds for any $k = 0, 1, \cdots$. This means that x^k converges to some x^* at a rate of p degree and proves the theorem.

□

Remark. Comparing Theorem 3 with Theorem 1, we note that Theorem 3 requires that C be a set of weak sharp minima for h. It is not surprising since, in Relaxation Algorithm 3, the correction term d^k is an approximating solution to the problem (2) such that the approximation error of the function value is controlled within a suitable bound. Thus it is closely related to the extent of the sharpness of h. Of course, the weak sharpness assumption on the minima set C for h can be replaced by the assumption that C a set of weak sharp minima of order $s \geq 1$ for h. In this case, if $\alpha > s$, the conclusion of Theorem 3 remains true for $p = \min\{2, \frac{\alpha}{s}\}$.

Acknowledgements. The authors thank the referees for their valuable comments and suggestion.

References

1. Borwein, J.M. (1986): Stability and regular points of inequality systems. J. Optim. Theory Appl. **48**, 9–52
2. Burke, J.V. (1987): Second order necessary and sufficient conditions for convex composite NDO. Math. Program. **38**, 287–302
3. Burke, J.V. (1991): An exact penalization viewpoint of constrained optimization. SIAM J. Control Optim. **29**, 968–998
4. Burke, J.V., Ferris, M.C. (1993): Weak sharp minima in mathematical programming. SIAM J. Control Optim. **31**, 1340–1359
5. Burke, J.V., Ferris, M.C. (1995): A Gauss-Newton method for convex composite optimization. Math. Program. **71**, 179–194
6. Crommer, L. (1978): Strong uniqueness. a far-reaching criterion for the convergence analysis of iterative procedures. Numer. Math. **29**, 179–193
7. Ferris, M.C. (1988): Weak sharp minima and penalty functions in mathematical programming. Ph.D. Thesis, University of Cambridge (Cambridge)
8. Garcia-Palomares, U.M., Restuccia, A. (1981): A global quadratic algorithm for solving a system of mixed equalities and inequalities. Math. Program. **21**, 290–300
9. Jittorntrum, K., Osborne, M.R. (1980): Strong uniqueness and second order convergence in nonlinear discrete approximation. Numer. Math. **34**, 439–455
10. Lin, P.-K. (1989): Strongly unique best approximation in uniformly convex Banach spaces. J. Approx. Theory **56**, 101–107
11. Osborne, M.R., Womersley, R.S. (1990): Strong uniqueness in sequential linear programming. J. Austral. Math. Soc. Ser. B **31**, 379–384
12. Polyak, B.T. (1987): Introduction to Optimization. (Optimization Software, New York)
13. Robinson, S. (1975): Stability theory for systems of inequalities, part I: linear systems. SIAM J. Numer. Anal. **12**, 754–769
14. Robinson, S. (1976): Stability theory for systems of inequalities, part II: differentiable nonlinear systems. SIAM J. Numer. Anal. **13**, 479–513
15. Rockafellar, R.T. (1988): First and second-order epi-differentiability in nonlinear programming. Trans. Am. Math. Soc. **307**, 75–108
16. Womersley, R.S. (1985): Local properties of algorithms for minimizing nonsmooth composite functions. Math. Program. **32**, 69–89
17. Yuan, Y., Sun, W. (1997): Optimization Theory and Methods. (Science Press, Beijing)