Mathematical Programming 32 (1985) 69-89
North-Holland

LOCAL PROPERTIES OF ALGORITHMS FOR MINIMIZING NONSMOOTH COMPOSITE FUNCTIONS

R.S. WOMERSLEY

School of Mathematics, University of New South Wales, P.O. Box 1, Kensington, N.S.W. 2033, Australia

Received 8 May 1984
Revised manuscript received 6 November 1984

This paper considers local convergence and rate of convergence results for algorithms for minimizing the composite function $F(x)=f(x)+h(c(x))$ where f and c are smooth but $h(c)$ may be nonsmooth. Local convergence at a second order rate is established for the generalized Gauss-Newton method when h is convex and globally Lipschitz and the minimizer is strongly unique. Local convergence at a second order rate is established for a generalized Newton method when the minimizer satisfies nondegeneracy, strict complementarity and second order sufficiency conditions. Assuming the minimizer satisfies these conditions, necessary and sufficient conditions for a superlinear rate of convergence for curvature approximating methods are established. Necessary and sufficient conditions for a two-step superlinear rate of convergence are also established when only reduced curvature information is available. All these local convergence and rate of convergence results are directly applicable to nonlinearing programming problems.

Key words: Composite Functions, Nonsmooth Optimization, Structure Functionals, Superlinear Convergence, Second Order Convergence, Strong Uniqueness, Reduced Curvature.

AMS(MOS) Subject Classification: 90 C 30.

1. Introduction

This paper considers local convergence and rate of convergence results for algorithms for minimizing the composite function $F:\mathbb{R}^n \to \mathbb{R}$ defined by

$$F(x)=f(x)+h(c(x)), \tag{1.1}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ and $c:\mathbb{R}^n \to \mathbb{R}^m$ are smooth (at least once continuously differentiable) but $h:\mathbb{R}^m \to \mathbb{R}$ may be nonsmooth. The function h is always locally Lipschitz, moreover it is often a convex or a polyhedral convex function in which case F is regular. Differential properties, optimality conditions and model algorithms for minimizing the composite function (1.1) are discussed by Womersley [21], providing a background to the current work. The problem of minimizing the composite function has also been considered by Fletcher [7, Chapter 14] when h is a polyhedral convex function, and by Osborne [15] and Powell [19] when h is a norm. Numerous special cases have also been considered by various authors.

This work was done while the author was a Research fellow at the Mathematical Sciences Research Centre, Australian National University.

The algorithms considered here are based on modelling the function F at the point $x^{(k)}$ by

$$\Psi(\delta ; x^{(k)})=f^{(k)}+\delta^{\mathrm{T}} g^{(k)}+\tfrac{1}{2}\delta^{\mathrm{T}} B^{(k)}\delta+h(c^{(k)}+A^{(k)\mathrm{T}}\delta), \tag{1.2}$$

where $f^{(k)}=f(x^{(k)})$, $g^{(k)}=\nabla f(x^{(k)})$, $c^{(k)}=c(x^{(k)})$ and $A^{(k)}=[\nabla c_1^{(k)} \cdots \nabla c_m^{(k)}]$. A minimizer $\delta^{(k)}$ of (1.2) is used as a correction term to try and improve the estimate $x^{(k)}$ of a minimizer x^* of F. This general framework includes a large number of published algorithms for minimizing F (see [7], [15] or [19] for reviews). In particular if $B^{(k)}=0$ for all k then minimizing $\Psi(\delta ; x^{(k)})$ corresponds to the *generalized Gauss-Newton method*, which has received a great deal of attention for nonlinear discrete approximation problems. If f and c are twice continuously differentiable and one takes $B^{(k)}=W^{(k)}$, for all k, where

$$W^{(k)}=G^{(k)}+\sum_{i=1}^{m} \lambda_i^{(k)}\nabla^2 c_i^{(k)}, \tag{1.3}$$

$G^{(k)}=\nabla^2 f(x^{(k)})$ and $\lambda^{(k+1)}\in\partial h(c^{(k)}+A^{(k)\mathrm{T}}\delta^{(k)})$ is used to calculate $W^{(k+1)}$, then minimizing (1.2) corresponds to a *generalized Newton method.* If h is a polyhedral convex function this is closely related to the Successive Quadratic Programming (SQP) method for nonlinear programming (see [7] or [21]). Another very popular idea is to use an approximation $B^{(k)}$ to $W^{(k)}$, often based on quasi-Newton methods.

This paper establishes local convergence at a second order rate for the generalized Gauss-Newton method when h is convex and globally Lipschitz and the minimizer is strongly unique, extending the work of Jittorntrum and Osborne [11] for the case when h is a norm. It is also shown that if h is just locally Lipschitz and the generalized Gauss-Newton method converges to a strongly unique minimizer then the rate of convergence is second order.

Local convergence at a second order rate is also established for the generalized Newton method when h is a polyhedral convex function and x^* satisfies a non-degeneracy condition, a strict complementarity condition and a second order sufficiency condition. This corresponds to the well known result for the SQP method [7] for nonlinear programming problems.

Under the above conditions on h and x^* necessary and sufficient conditions on the matrices $B^{(k)}$ for a superlinear rate of convergence for methods based on the model function (1.2) are established. These correspond to the results of Dennis and Moré [6] when h is twice continuously differentiable, of Boggs, Tolle and Wang [1] and Powell [18] for nonlinear programming problems, and Powell and Yuan [20] for l_1 and l_∞ approximation problems. Finally necessary and sufficient conditions for a two-step superlinear rate of convergence when only reduced curvature information is available are given. These results generalize those of Powell [18] for nonlinear programming problems, Han [10] for the minimax problem, and Coleman and Conn [3] for an exact l_1 penalty function method. Throughout only mild conditions are imposed upon the matrices $B^{(k)}$, namely that they are uniformly positive definite *on an appropriate subspace*, which is the natural condition to be satisfied if second order sufficiency conditions are to be satisfied.

In all cases these results deal with the sequence $\{x^{(k)}\}$ generated by taking $x^{(k+1)}=x^{(k)}+\delta^{(k)}$, where $\delta^{(k)}$ is a minimizer of (1.2). However modifications to ensure global convergence of the method (for example adding a line search or a trust region) may prevent this (see [12] and [24] for examples). Consequently the results of this paper represent local properties that are achievable by algorithms. An area for further work is modifications which guarantee global convergence but do not affect these local convergence properties (see Chamberlain et al. [2] and Fletcher [9] for example).

In Section 7 the connection with smooth problems is discussed. The results of Sections 5 and 6 are presented in such a way that they are all applicable to the nonlinear programming problem

$$\begin{aligned} &\underset{x\in\mathbb{R}^n}{\text{minimize}} && f(x) \\ &\text{subject to} && c_i(x)=0, \quad i\in E, \\ & && c_i(x)\geqslant 0, \quad i\in I, \end{aligned} \tag{1.4}$$

where f and c are twice continuously differentiable, and E and I are finite index sets. The subproblem corresponding to that of minimizing (1.2) is

$$\begin{aligned} &\underset{\delta\in\mathbb{R}^n}{\text{minimize}} && f^{(k)}+\delta^{\mathrm{T}}g^{(k)}+\tfrac{1}{2}\delta^{\mathrm{T}}B^{(k)}\delta \\ &\text{subject to} && c_i^{(k)}+\delta^{\mathrm{T}}\nabla c_i^{(k)}=0 \quad \text{for } i\in E, \\ & && c_i^{(k)}+\delta^{\mathrm{T}}\nabla c_i^{(k)}\geqslant 0 \quad \text{for } i\in I. \end{aligned} \tag{1.5}$$

All the local convergence and rate of convergence results established for the composite function (1.1) also hold when the subproblem (1.5) is used to solve (1.4). In this case many of the results are known (see Powell [18] and Boggs, Tolle and Wang [1] for example). However the necessary and sufficient conditions for a two-step superlinear rate of convergence when only reduced curvature information is available are new.

2. Preliminaries

A key to the uniform treatment of local convergence and rate of convergence results is the use of a concise representation of the generalized gradient in terms of structure functionals, introduced by Osborne [17] for polyhedral convex functions and extended to piecewise smooth functions in [21]. For any piecewise smooth function h there exists a minimal linearly independent set of $l\equiv l(c)$ smooth functionals $\varphi_j(c):\mathbb{R}^m\to\mathbb{R}$ such that

$$\varphi_j(c)=0, \quad j=1,\ldots,l, \tag{2.1}$$

and

$$\partial h(c)=\{\lambda \in \mathbb{R}^m\colon \lambda=\lambda_0(c)+\Phi(c)u,\ u\in U(c)\}, \tag{2.2}$$

where $\Phi(c)=[\nabla\varphi_1(c)\cdots\nabla\varphi_l(c)]$ and $U(c)$ is a nonempty compact convex set in $\mathbb{R}^l$. The $\varphi_j(c)$ are referred to as *structure functionals*. The function h is smooth at c if and only if $l=0$ so $\partial h(c)=\{\lambda_0(c)\}=\{\nabla h(c)\}$. By definition $\Phi(c)$ has rank l and $0\leqslant l\leqslant m$ so there is a one-to-one correspondence between $\lambda\in\partial h(c)$ and $u\in U(c)$. Note that if h is a piecewise linear function (for example a polyhedral convex function) the structure functionals are linear so $\varphi_j(c)=\varphi_j^{\mathrm{T}}c+\xi_j$. The explicit dependence of λ_0, Φ and U upon c may be omitted for notational convenience when no ambiguity results.

The generalized gradient of the composite function (1.1) is given by

$$\begin{aligned}\partial F(x)&=\{v\in\mathbb{R}^n\colon v=g(x)+A(x)\lambda,\ \lambda\in\partial h(c(x))\}\\&=\{v\in\mathbb{R}^n\colon v=g(x)+A(x)\lambda_0+A(x)\Phi u,\ u\in U(c(x))\}.\end{aligned} \tag{2.3}$$

The problem is said to be *degenerate* at the point x if the vectors $\nabla\varphi_j(c(x))$ are linearly dependent on the space spanned by the rows of $A(x)$. Thus a common nondegeneracy assumption is that the $n\times l$ matrix $A(x)\Phi$ has rank l, which of course cannot hold if $l>n$. It is important that this nondegeneracy assumption is no stronger than the usual one made in nonlinear programming problems, as it corresponds to an assumption that the gradients of the active constraints are linearly independent.

The point x^* is a *stationary point* if $0\in\partial F(x^*)$. From (2.3) equivalent expressions are

$$\exists\lambda^*\in\partial h(c^*)\text{ such that }0=g^*+A^*\lambda^*, \tag{2.4}$$

or

$$\exists u^*\in U^*\text{ such that }0=g^*+A^*\lambda_0^*+A^*\Phi^*u^*. \tag{2.5}$$

If x^* is nondegenerate the multipliers u^* (and λ^*) are uniquely determined by (2.5). *Strict complementarity* holds at a stationary point x^* if u^* lies in the interior of U^* (regarded as a set in $\mathbb{R}^{l^*}$).

The following second order sufficient conditions (see [7] or [21]) are also required. They assume f and c are twice continuously differentiable.

Proposition 2.1. *Let h be a regular locally Lipschitz function (for instance a convex function). If x^* is a stationary point of* (1.1), *so* (2.4) *holds, and if*

$$s^{\mathrm{T}}W^*s>0\quad\forall s\in S^*, \tag{2.6}$$

where

$$W^*=G^*+\sum_{i=1}^{m}\lambda_i^*\nabla^2c_i^*, \tag{2.7}$$

and

$$S^* = \{s \in \mathbb{R}^n : \|s\| = 1,\ F'(x^*; s) = 0\}, \tag{2.8}$$

then x^* is a strict local minimizer of F.

Remarks. (i) If strict complementarity holds at x^* then, from lemma 4.1 of [21],

$$S^* = \{s \in \mathbb{R}^n : \|s\| = 1, s^{\mathrm{T}} A^* \Phi^* = 0\}, \tag{2.9}$$

which (excluding the $\|s\| = 1$ condition and assuming x^* is nondegenerate) is the tangent space to the surface of nondifferentiability at x^*.

(ii) If $l^* = n$ and strict complementarity holds at x^* then $S^* = \phi$, so the first order conditions (2.4) or (2.5) are sufficient. In this case the solution x^* is *strongly unique.* This is equivalent to the usual definition of strong uniqueness (see [11] or [16] for example), namely there exists a $\gamma > 0$ such that

$$F(x) \geq F(x^*) + \gamma \|x - x^*\| \tag{2.10}$$

for all x in a neighbourhood of x^*. Note that if F is convex then the inequality (2.10) holds for all x.

Many of the results in this paper assume that an algorithm generates a sequence of points $\{x^{(k)}\}$ converging to x^* and establish necessary and sufficient conditions for various rates of convergence. The rate of convergence is *superlinear* (in fact Q-superlinear) if and only if

$$\lim_{k \to \infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|} = 0, \tag{2.11}$$

or equivalently

$$x^{(k+1)} = x^* + \mathrm{o}(\|x^{(k)} - x^*\|). \tag{2.12}$$

The rate of convergence is *two-step superlinear* if and only if

$$\lim_{k \to \infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k-1)} - x^*\|} = 0, \tag{2.13}$$

or equivalently

$$x^{(k+1)} = x^* + \mathrm{o}(\|x^{(k-1)} - x^*\|). \tag{2.14}$$

Finally the rate of convergence is *second order* if and only if

$$\lim_{k \to \infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|^2} < \infty, \tag{2.15}$$

or equivalently

$$x^{(k+1)} = x^* + \mathrm{O}(\|x^{(k)} - x^*\|^2). \tag{2.16}$$

3. Strong uniqueness

Strong uniqueness of the solution (that is (2.10) holds) is a powerful condition which ensures local convergence at a second order rate for the generalized Gauss–Newton method applied to nonlinear discrete approximation problems (see [11] or [16] for example). In fact strong uniqueness of the solution produces local convergence at a second order rate for a wide class of methods applied to the composite function $F(x)=f(x)+h(c(x))$ when h is convex and Lipschitz on $\mathbb{R}^m$. This includes all norms and polyhedral convex functions h. When h is just locally Lipschitz strong uniqueness of the solution implies the rate of convergence is second order if the method is convergent.

Consider modelling the composite function F in a neighbourhood of a point y by

$$\begin{aligned}\psi(x;y)=&f(y)+(x-y)^{\mathrm{T}}g(y)+\tfrac{1}{2}(x-y)^{\mathrm{T}}B(y)(x-y)\\&+h(c(y)+A(y)^{\mathrm{T}}(x-y)).\end{aligned}\tag{3.1}$$

A new estimate of a minimizer of $F(x)$ is obtained by solving the subproblem

$$\min_{x\in\mathbb{R}^n}\psi(x;y),\tag{3.2}$$

so the generalized Gauss–Newton method corresponds to $B(y)\equiv 0$.

Lemma 3.1. *Let h be locally Lipschitz and let the matrix $B(x^*)$ be bounded. Then x^* is a strongly unique minimizer of $F(x)$ if and only if it is a strongly unique minimizer of $\psi(x;x^*)$.*

Proof. From the smoothness properties of f and c, the boundedness of $B(x^*)$, and the locally Lipschitz nature of h it follows that

$$\psi(x;x^*)=F(x)+\mathrm{o}(\|x-x^*\|).$$

If F is strongly unique at x^* then, as $\psi(x^*;x^*)=F(x^*)$,

$$\psi(x;x^*)\geq F(x^*)+\gamma\|x-x^*\|+\mathrm{o}(\|x-x^*\|)\geq\psi(x^*;x^*)+\bar{\gamma}\|x-x^*\|$$

for all x in a small enough neighbourhood of x^*. The argument can be reversed to complete the proof. □

The next result generalizes the local convergence results of [5], [11] and [16] for the generalized Gauss–Newton method when x^* is strongly unique. The proof of the theorem is along similar lines to that given in [11]. It is assumed that f and c are smooth and their first derivatives satisfy a Lipschitz condition.

Theorem 3.1. *Let h be a function Lipschitz on $\mathbb{R}^m$, i.e. there exists a positive constant L such that*

$$|h(c)-h(\bar{c})|\leq L\|c-\bar{c}\|\quad\forall c,\bar{c}\in\mathbb{R}^m.\tag{3.3}$$

Let $\psi(x; x^*)$ be a convex function of x and let x^* be a strongly unique minimizer of $\psi(x; x^*)$. Let the sequence $\{x^{(k)}\}$ be generated by taking $x^{(k+1)}$ as a minimizer of $\psi(x; x^{(k)})$ and let the matrices $B^{(k)}$ in $\psi(x; x^{(k)})$ be uniformly bounded. Then if $x^{(k)}$ is sufficiently close to x^* the sequence $\{x^{(k)}\}$ converges to x^* and the rate of convergence is second order.

Proof. By the continuity properties of f and c and the uniform boundedness of the $B^{(k)}$ there exists a neighbourhood Ω of x^* such that for all $y \in \Omega$ and for all $x \in \mathbb{R}^n$

(i) $|f(y) - f^* + g(y)^{\mathrm{T}}(x^* - y) + \frac{1}{2}(x^* - y)^{\mathrm{T}} B^{(k)}(x^* - y)| \leq K_1 \|x^* - y\|^2$,

(ii) $|(g(y) - g^*)^{\mathrm{T}}(x - x^*)| \leq K_2 \|x^* - y\| \, \|x - x^*\|$,

(iii) $\|c(y) - c^* + A(y)^{\mathrm{T}}(x^* - y)\| \leq K_3 \|x^* - y\|^2$,

(iv) $\|(A(y) - A^*)^{\mathrm{T}}(x - x^*)\| \leq K_4 \|x^* - y\| \, \|x - x^*\|$,

(v) $(K_2 + LK_4)\|y - x^*\| \leq \gamma/2$,

(vi) $4(K_1 + LK_3)\|y - x^*\|/\gamma \leq \theta < 1$,

where K_1 to K_4 are positive constants. From (i) to (iv) and (3.3) one has that for $x^{(k)} \in \Omega$

$$|\psi(x; x^{(k)}) - \psi(x; x^*)| \leq (K_1 + LK_3)\|x^* - x^{(k)}\|^2 + (K_2 + LK_4)\|x^* - x^{(k)}\| \, \|x - x^*\|, \tag{3.4}$$

for any x. Now strong uniqueness yields

$$\begin{aligned} &\psi(x^*; x^*) + \gamma\|x^{(k+1)} - x^*\| - (K_1 + LK_3)\|x^* - x^{(k)}\|^2 \\ &\quad - (K_2 + LK_4)\|x^* - x^{(k)}\| \, \|x^{(k+1)} - x^*\| \\ &\leq \psi(x^{(k+1)}; x^*) - |\psi(x^{(k+1)}; x^{(k)}) - \psi(x^{(k+1)}; x^*)| \\ &\leq \psi(x^{(k+1)}; x^{(k)}) \leq \psi(x^*; x^{(k)}) \leq \psi(x^*; x^*) + (K_1 + LK_3)\|x^* - x^{(k)}\|^2. \end{aligned}$$

Hence

$$(\gamma - (K_2 + LK_4)\|x^{(k)} - x^*\|)\|x^{(k+1)} - x^*\| \leq 2(K_1 + LK_3)\|x^{(k)} - x^*\|^2.$$

Conditions (v) and (vi) now yield

$$\|x^{(k+1)} - x^*\| \leq 4(K_1 + LK_3)\|x^{(k)} - x^*\|^2/\gamma \leq \theta\|x^{(k)} - x^*\|.$$

Thus $x^{(k+1)} \in \Omega$, the sequence $\{x^{(k)}\}$ converges to x^* and the rate of convergence is second order. □

Remark. The convexity of $\psi(x; x^*)$ is only needed to ensure that

$$\psi(x^{(k+1)}; x^*) \geq \psi(x^*; x^*) + \gamma\|x^{(k+1)} - x^*\|$$

without any requirement that $x^{(k+1)}$ is close to x^*. If h is convex and $B(y) \equiv 0$ then $\psi(x; y)$ is a convex function of x for all y. The Lipschitz condition (3.3) is needed for (iv) to hold for all x (in particular $x^{(k+1)}$). If h is locally Lipschitz then (3.3) can be replaced by an assumption that the sequence $\{x^{(k)}\}$ is uniformly bounded. If one only has that h is locally Lipschitz then the above proof (without any convexity assumption on ψ) shows that if $x^{(k)} \rightarrow x^*$ then the rate of convergence is second order.

4. Basic estimates

Two basic results for algorithms applied to the composite function (1.1) when h is a polyhedral convex function are established in this section. Both results assume that we are in a neighbourhood of a minimizer x^* satisfying

$$A^*\Phi^* \text{ has full rank } l^* \quad \text{(nondegeneracy)} \tag{4.1}$$

and

$$u^* \in \operatorname{int} U^* \quad \text{(strict complementarity)}, \tag{4.2}$$

where U^* is regarded as a set in $\mathbb{R}^{l^*}$. The first result is that the generalized gradient of h at the solution of the subproblem is the same as the generalized gradient of h at x^*. This result directly corresponds to the nonlinear programming result that sufficiently close to a minimizer satisfying a strict complementarity condition a subproblem produced by linearizing the constraints correctly predicts the set of active constraints at the solution. The second result establishes the estimates which are the basis of all the succeeding local convergence and rate of convergence results. The only condition on the matrices $B^{(k)}$ appearing in the model function (1.2) is that $B^{(k)}$ is uniformly positive definite *on the tangent space to the surface of nondifferentiability at* x^*, that is

$$s^{\mathrm{T}}B^{(k)}s \geqslant \beta s^{\mathrm{T}}s > 0 \qquad \forall s\colon \quad s^{\mathrm{T}}A^*\Phi^* = 0, \; s \neq 0. \tag{4.3}$$

If x is sufficiently close to x^* then the nondegeneracy assumption and the smoothness properties of $c(x)$ imply $A(x)\Phi^*$ has full rank l^* and (4.3) can be replaced by

$$s^{\mathrm{T}}B^{(k)}s \geqslant \tilde{\beta} s^{\mathrm{T}}s > 0 \qquad \forall s\colon \quad s^{\mathrm{T}}A(x)\Phi^* = 0, \; s \neq 0. \tag{4.4}$$

If $l^* = n$ then x^* is strongly unique and both (4.3) and (4.4) are trivially satisfied. It is natural to impose (4.3) or (4.4) as they guarantee that any stationary point of (4.5) is a local minimizer of (4.5). Also if (4.1) holds and x is sufficiently close to x^* the second order necessary conditions ([7] or [21]) for (4.5) imply $s^{\mathrm{T}}B^{(k)}s \geqslant 0$ for all s such that $s^{\mathrm{T}}A(x)\Phi^* = 0$. Moreover the second order sufficient conditions for F require that W^* satisfies (4.3), and hence if $x^{(k)}$, $\lambda^{(k)}$ are sufficiently close to x^*, λ^* that $W^{(k)}$ satisfies (4.3) and (4.4). Thus as $B^{(k)}$ is approximating $W^{(k)}$ conditions (4.3) or (4.4) do not impose any unwanted structure upon $B^{(k)}$. Technically (4.3) and (4.4) can be replaced by an assumption that, for all k, $B^{(k)}$ is nonsingular on the appropriate subspace, but one then needs an assumption that the stationary point of (4.5) one has is a minimizer of (4.5).

Lemma 4.1. *Let h be a polyhedral convex function and let x^* be a stationary point of the composite function $F(x)=f(x)+h(c(x))$ satisfying the nondegeneracy condition* (4.1) *and the strict complementarity condition* (4.2). *Consider modelling the function F at the point x by*

$$\Psi(\delta;x)=f(x)+\delta^{\mathrm{T}}g(x)+\tfrac{1}{2}\delta^{\mathrm{T}}B\delta+h(c(x)+A(x)^{\mathrm{T}}\delta), \tag{4.5}$$

where B satisfies (4.3). *If x is sufficiently close to x^* then the unique minimizer $\delta(x)$ of $\Psi(\delta;x)$ satisfies $\partial h(c(x)+A(x)^{\mathrm{T}}\delta(x))=\partial h(c^*)$.*

Proof. Consider the system of linear equations

$$\begin{bmatrix} B & A(x)\Phi^* \\ (A(x)\Phi^*)^{\mathrm{T}} & 0 \end{bmatrix}\begin{bmatrix} \delta \\ u \end{bmatrix}=\begin{bmatrix} -g(x)-A(x)\lambda_0^* \\ \Phi^{*\mathrm{T}}(c^*-c(x)) \end{bmatrix}. \tag{4.6}$$

From (4.1) and (4.3) the matrix

$$\begin{bmatrix} B & A^*\Phi^* \\ (A^*\Phi^*)^{\mathrm{T}} & 0 \end{bmatrix}$$

is nonsingular, and as x^* is a stationary point $\delta=0$ and $u=u^*$ solve the corresponding system (4.6). Thus, by the implicit function theorem, there exists a neighbourhood of x^* such that for any x in this neighbourhood the *unique* solution $\delta(x)$, $u(x)$ to (4.6) changes smoothly with x and $\delta(x^*)=0$, $u(x^*)=u^*$. As $u^*\in\operatorname{int}U^*$ there exists a neighbourhood of x^* such that $u(x)\in\operatorname{int}U^*$, and hence $\lambda(x)=\lambda_0^*+\Phi^*u(x)\in\partial h(c^*)$. Moreover from (4.6)

$$g(x)+B\delta(x)+A(x)\lambda(x)=0. \tag{4.7}$$

Now as h is a polyhedral convex function, we have

$$h(c)=\max_{i=1,\dots,r} h_i^{\mathrm{T}}c+\beta_i.$$

Let

$$\mathscr{A}(c)=\{i\in 1,\dots,r\colon h_i^{\mathrm{T}}c+\beta_i=h(c)\},$$

and

$$\tilde{c}(x)=c(x)+A(x)^{\mathrm{T}}\delta(x).$$

Then there exists a neighbourhood of x^* such that

$$h(\tilde{c}(x))=\max_{i\in\mathscr{A}^*} h_i^{\mathrm{T}}\tilde{c}(x)+\beta_i. \tag{4.8}$$

Also, for any $i\in\mathscr{A}^*$ one has $h_i\in\partial h(c^*)$, so there exists a $u_i\in U^*$ such that $h_i=\lambda_0^*+\Phi^*u_i$. Thus for any $i,j\in\mathscr{A}^*$

$$h_i^{\mathrm{T}}c+\beta_i-h_j^{\mathrm{T}}c-\beta_j=(u_i-u_j)^{\mathrm{T}}\Phi^*c+\beta_i-\beta_j.$$

Using this equation with $c=\tilde{c}(x)$ and $c=c^*$, the definition of $\mathscr{A}^*$ and (4.6) yield

$$h_i^{\mathrm{T}}\tilde{c}(x)+\beta_i-h_j^{\mathrm{T}}\tilde{c}(x)-\beta_j=(u_i-u_j)^{\mathrm{T}}\Phi^*(\tilde{c}(x)-c^*)=0.$$

Then (4.8) implies $\mathscr{A}(\tilde{c}(x))=\mathscr{A}^*$ and $\partial h(\tilde{c}(x))=\partial h(c^*)$ for all x in a neighbourhood of x^*. It now follows from (4.7) that $\delta(x)$ is a stationary point of $\Psi(\delta;x)$, and uniqueness follows as $\delta(x)$, $u(x)$ is the unique solution of (4.6) with $\delta(x^*)=0$ and $u(x^*)=u^*$. Also as B satisfies (4.3) and hence (4.4), $\delta(x)$ satisfies the sufficient conditions (Proposition 2.1) for a minimizer of $\Psi(\delta;x)$. □

Remark. An immediate consequence of $\partial h(c(x)+A(x)\delta(x))=\partial h(c^*)$ is that

$$\Phi^{*\mathrm{T}}(c(x)+A(x)^{\mathrm{T}}\delta(x))=\Phi^{*\mathrm{T}}c^*, \tag{4.9}$$

which corresponds to (2.1) and is established directly in the proof.

For the rest of the paper it is assumed f and c are twice continuously differentiable, and their second derivatives satisfy a Lipschitz condition.

Lemma 4.2. *Let h be a polyhedral convex function and let x^* be a stationary point of the composite function $F(x)=f(x)+h(c(x))$ satisfying the nondegeneracy condition* (4.1) *and the strict complementarity condition* (4.2). *Let $\delta^{(k)}$ be a stationary point of the model function* (1.2) *where $B^{(k)}$ satisfies* (4.3), *and let $\lambda^{(k+1)}\in\partial h(c^{(k)}+A^{(k)\mathrm{T}}\delta^{(k)})$ be the corresponding multipliers. Define $e^{(k)}=x^{(k)}-x^*$ and $e^{(k+1)}=x^{(k)}+\delta^{(k)}-x^*$. If $x^{(k)}$ is sufficiently close to x^* then $\lambda^{(k+1)}=\lambda_0^*+\Phi^*u^{(k+1)}$ and*

$$\begin{bmatrix} B^{(k)} & A^{(k)}\Phi^* \\ (A^{(k)}\Phi^*)^{\mathrm{T}} & 0\end{bmatrix}\begin{bmatrix} e^{(k+1)} \\ u^{(k+1)}-u^*\end{bmatrix}$$

$$=\begin{bmatrix}(B^{(k)}-W^{(k)})e^{(k)} \\ 0\end{bmatrix}+\begin{bmatrix}\mathrm{O}(\|e^{(k)}\|^2)+\mathrm{O}(\|e^{(k)}\|\,\|\lambda^{(k)}-\lambda^*\|) \\ \mathrm{O}(\|e^{(k)}\|^2)\end{bmatrix} \tag{4.10}$$

where

$$W^{(k)}=G^{(k)}+\sum_{i=1}^{m}\lambda_i^{(k)}\nabla^2c_i^{(k)}. \tag{4.11}$$

Proof. The smoothness properties of f and c imply

$$g^*=g^{(k)}-G^{(k)}e^{(k)}+\mathrm{O}(\|e^{(k)}\|^2) \tag{4.12}$$

and

$$\nabla c_i^*=\nabla c_i^{(k)}-\nabla^2c_i^{(k)}e^{(k)}+\mathrm{O}(\|e^{(k)}\|^2)\quad \text{for } i=1,\ldots,m. \tag{4.13}$$

As $\delta^{(k)}$ is a stationary point of $\psi(\delta;x^{(k)})$ and $\lambda^{(k+1)}\in\partial h(c^{(k)}+A^{(k)\mathrm{T}}\delta^{(k)})$ are the corresponding multipliers

$$g^{(k)}+B^{(k)}\delta^{(k)}+A^{(k)}\lambda^{(k+1)}=0. \tag{4.14}$$

Also as x^* is a stationary point of F there exists a $\lambda^*\in\partial h(c^*)$ such that

$$g^*+A^*\lambda^*=0. \tag{4.15}$$

Now (4.11) to (4.15) yield

$$B^{(k)}e^{(k+1)}+A^{(k)}(\lambda^{(k+1)}-\lambda^*)$$
$$=(B^{(k)}-W^{(k)})e^{(k)}+\mathrm{O}(\|e^{(k)}\|^2)+\mathrm{O}(\|e^{(k)}\|\,\|\lambda^{(k)}-\lambda^*\|). \tag{4.16}$$

For $x^{(k)}$ sufficiently close to x^* Lemma 4.1 implies $\partial h(c^{(k)}+A^{(k)\mathrm{T}}\delta^{(k)})=\partial h(c^*)$, so

$$\lambda^{(k+1)}-\lambda^*=\Phi^*(u^{(k+1)}-u^*) \tag{4.17}$$

and

$$\Phi^{*\mathrm{T}}(c^{(k)}+A^{(k)\mathrm{T}}\delta^{(k)})=\Phi^{*\mathrm{T}}c^*. \tag{4.18}$$

Now $c^*=c^{(k)}-A^{(k)}e^{(k)}+\mathrm{O}(\|e^{(k)}\|^2)$ and $\delta^{(k)}=e^{(k+1)}-e^{(k)}$ so (4.18) gives

$$(A^*\Phi^*)^{\mathrm{T}}e^{(k+1)}=\mathrm{O}(\|e^{(k)}\|^2). \tag{4.19}$$

Combining (4.16), (4.17) and (4.19) gives the desired result. □

Remarks. (i) The minor role played by the multipliers is illustrated by the fact that $\|\lambda^{(k)}-\lambda^*\|$ appears only linearly on the right hand side of (4.10). Also the multipliers used in (4.11) are those generated by the subproblem (1.2).

(ii) Equation (4.19) shows that there is a second order decrease in the error terms in the space orthogonal to the surface of nondifferentiability at x^*. Significantly (4.19) does not depend upon the matrix $B^{(k)}$.

5. Newton like methods

When h is a polyhedral convex function then the problem of minimizing F is equivalent to a nonlinear programming problem. Also the SQP method (see [7] for example) is equivalent to minimizing the model function

$$\Psi(\delta;x^{(k)})=f^{(k)}+\delta^{\mathrm{T}}g^{(k)}+\tfrac{1}{2}\delta^{\mathrm{T}}W^{(k)}\delta+h(c^{(k)}+A^{(k)\mathrm{T}}\delta) \tag{5.1}$$

(see [7] or [21] for more details). As the SQP method is known to converge locally at a second order rate a corresponding result is available when using (5.1) to minimize the composite function (1.1). In fact using Lemma 4.1 the proof of this result for the SQP method (see [7, p. 141] for example) directly extends to (5.1).

Theorem 5.1. *Let h be a polyhedral convex function and x^* a minimizer of the composite function $F(x)=f(x)+h(c(x))$ satisfying* (4.1), (4.2) *and the second order sufficiency condition*

$$s^{\mathrm{T}}W^*s>0 \qquad \forall s \quad s^{\mathrm{T}}A^*\Phi^*=0,\ s\neq 0. \tag{5.2}$$

Let $\delta^{(k)}$ be a minimizer of (5.1) and let $\lambda^{(k+1)}\in\partial h(c^{(k)}+A^{(k)\mathrm{T}}\delta^{(k)})$ be the corresponding multipliers. If $x^{(1)}$ is sufficiently close to x^* and $\lambda^{(1)}$ is chosen so that the matrix

$$\begin{bmatrix} W^{(1)} & A^{(1)}\Phi^* \\ (A^{(1)}\Phi^*)^{\mathrm{T}} & 0 \end{bmatrix} \tag{5.3}$$

is nonsingular and $\Psi(\delta; x^{(1)})$ has a well defined minimizer $\delta^{(1)}$, then the sequence $\{x^{(k)}\}$ generated by $x^{(k+1)} = x^{(k)} + \delta^{(k)}$ converges to x^* and the rate of convergence is second order.

Proof. From (5.2) it follows that for $x^{(k)}$, $\lambda^{(k)}$ in some neighbourhood of x^*, λ^*

$$s^{\mathrm{T}} W^{(k)} s \geq \beta s^{\mathrm{T}} s > 0 \qquad \forall s\colon\ s^{\mathrm{T}} A^{(k)} \Phi^* = 0,\ s \neq 0.$$

Hence from (4.1) the matrix

$$\begin{bmatrix} W^{(k)} & A^{(k)}\Phi^* \\ (A^{(k)}\Phi^*)^{\mathrm{T}} & 0 \end{bmatrix} \tag{5.4}$$

is nonsingular in a neighbourhood of x^*, λ^*. Lemma 4.2 yields

$$\begin{bmatrix} e^{(k+1)} \\ u^{(k+1)} - u^* \end{bmatrix} = \mathrm{O}(\|e^{(k)}\|^2) + \mathrm{O}(\|e^{(k)}\|\ \|\lambda^{(k)} - \lambda^*\|). \tag{5.5}$$

If $x^{(k)}$ is sufficiently close to x^* lemma 4.1 shows that $\lambda^{(k+1)} \in \partial h(c^*)$ so

$$\lambda^{(k+1)} - \lambda^* = \Phi^*(u^{(k+1)} - u^*).$$

From (5.5) it follows that there exists a constant $K_1 > 0$ such that

$$\max(\|e^{(k+1)}\|, \|\lambda^{(k+1)} - \lambda^*\|) \leq K_1 \|e^{(k)}\| \max(\|e^{(k)}\|, \|\lambda^{(k)} - \lambda^*\|). \tag{5.6}$$

Now there exists a neighbourhood of x^*, λ^* such that $\theta = K_1 \max(\|e^{(k)}\|, \|\lambda^{(k)} - \lambda^*\|) < 1$ so

$$\max(\|e^{(k+1)}\|, \|\lambda^{(k+1)} - \lambda^*\|) \leq \theta \|e^{(k)}\| \leq \theta \max(\|e^{(k)}\|, \|\lambda^{(k)} - \lambda^*\|).$$

Thus the iteration converges and the rate is second order from (5.6). Now suppose only $x^{(1)}$ is in a neighbourhood of x^* and $\lambda^{(1)}$ is chosen so (5.3) is nonsingular. Then $\|\lambda^{(1)} - \lambda^*\| \geq \|e^{(1)}\|$ and as above there exists a constant $K_2 > 0$ such that

$$\max(\|e^{(2)}\|, \|\lambda^{(2)} - \lambda^*\|) \leq K_2 \|e^{(1)}\|\ \|\lambda^{(1)} - \lambda^*\|.$$

Let $x^{(1)}$ be close enough to x^* so that $\|e^{(1)}\| < 1/(K_1 K_2 \|\lambda^{(1)} - \lambda^*\|)$ then $\max(\|e^{(2)}\|, \|\lambda^{(2)} - \lambda^*\|) < 1/K_1$. Thus $x^{(2)}$, $\lambda^{(2)}$ is in the neighbourhood for which convergence occurs. □

6. Curvature approximating methods

In this section necessary and sufficient conditions on the matrices $B^{(k)}$ in (1.2) for a quadratic, superlinear, and two-step superlinear rate of convergence are established. The results generalize those of Dennis and Moré [6] for twice continuously differentiable functions, of Powell [18] and Boggs, Tolle and Wang [1] for the nonlinear programming problem, and Powell and Yuan [20] for l_1 and l_∞ approximation problems.

Throughout this section the following assumptions are used.

Assumptions A

(i) f and c are twice continuously differentiable, the second derivatives of f and c satisfy a Lipschitz condition, and h is a polyhedral convex function.

(ii) x^* is a stationary point of the composite function (1.1) satisfying the nondegeneracy condition (4.1), the strict complementarity condition (4.2) and the second order sufficiency condition (5.2).

(iii) $\delta^{(k)}$ is a stationary point of the model function (1.2) where the matrices $B^{(k)}$ satisfy the second order sufficiency condition (4.3), so $\delta^{(k)}$ is a minimizer of (1.2), and $\lambda^{(k+1)} \in \partial h(c^{(k)} + A^{(k)\mathrm{T}}\delta^{(k)})$ are the corresponding multipliers.

(iv) the sequence $\{x^{(k)}\}$, generated by $x^{(k+1)} = x^{(k)} + \delta^{(k)}$, converges to x^*.

The results of this section are in terms of the orthogonal projection matrices $P^{(k)}$ and $Q^{(k)} = I - P^{(k)}$ where

$$P^{(k)} : \mathbb{R}^n \to S_0^{(k)} = \{s \in \mathbb{R}^n : s^\mathrm{T} A^{(k)} \Phi^* = 0\}. \tag{6.1}$$

Note that for $x^{(k)}$ sufficiently close to x^* the nondegeneracy condition (4.1) ensures that the $n \times l^*$ matrix $A^{(k)}\Phi^*$ has rank l^*. When $0 < l^* < n$ this implies $S_0^{(k)}$ is the tangent space at the point $x^{(k)}$ to the surface of nondifferentiability of F which passes through x^*. If $l^* = n$ then the surface of nondifferentiability reduces to the point x^* so $P^{(k)} = I$ and $Q^{(k)} = 0$. Alternatively if $l^* = 0$ then F is smooth in a neighbourhood of x^* so $P^{(k)} = 0$ and $Q^{(k)} = I$. Also as $P^{(k)}$ and $Q^{(k)}$ are orthogonal projection operators, when $0 < l^* < n$ there exist an $n \times (n - l^*)$ orthogonal matrix $Z^{(k)}$ and an $n \times l^*$ orthogonal matrix $Y^{(k)}$ such that $Z^{(k)\mathrm{T}} Y^{(k)} = 0$,

$$P^{(k)} = Z^{(k)} Z^{(k)\mathrm{T}}, \qquad Q^{(k)} = Y^{(k)} Y^{(k)\mathrm{T}}, \tag{6.2}$$

$$Z^{(k)\mathrm{T}} A^{(k)} \Phi^* = 0 \quad \text{and} \quad A^{(k)}\Phi^* = Y^{(k)} R^{(k)}, \tag{6.3}$$

where $R^{(k)}$ is nonsingular. In a neighbourhood of x^* the matrix $A^{(k)}\Phi^*$ depends continuously upon $x^{(k)}$ so $P^{(k)}$, $Q^{(k)}$, $Y^{(k)}$ and $R^{(k)}$ depend continuously upon $x^{(k)}$. In particular $P^{(k)} \to P^*$, $Q^{(k)} \to Q^*$, $Y^{(k)} \to Y^*$ and $R^{(k)} \to R^*$ as $x^{(k)} \to x^*$. However $Z^{(k)}$ may be chosen in a number of ways so a particular method may not produce a $Z^{(k)}$ which depends continuously upon $x^{(k)}$. The continuity properties of $Z^{(k)}$ are discussed in [4].

Theorem 6.1. *Let the assumptions A be satisfied. Then the rate of convergence is*

(a) *superlinear if and only if*

$$\lim_{k \to \infty} \frac{\|P^{(k)}(W^{(k)} - B^{(k)})(x^{(k)} - x^*)\|}{\|x^{(k)} - x^*\|} = 0, \tag{6.4}$$

(b) *quadratic if and only if*

$$\lim_{k\to\infty} \frac{\|P^{(k)}(W^{(k)}-B^{(k)})(x^{(k)}-x^*)\|}{\|x^{(k)}-x^*\|^2} < \infty. \tag{6.5}$$

Proof. The nondegeneracy condition (4.1) ensures that $\lambda^* = \lambda_0^* + \Phi^* u^*$ is the unique solution of (4.15), and for $x^{(k)}$ sufficiently close to x^* that $\lambda^{(k+1)} = \lambda_0^* + \Phi^* u^{(k+1)}$ is the unique solution to (4.14). Thus $u^{(k)} \to u^*$ and $\lambda^{(k)} \to \lambda^*$ as $x^{(k)} \to x^*$. Moreover from (4.10)

$$A^{(k)}\Phi^*(u^{(k+1)}-u^*) = -B^{(k)}e^{(k+1)} + (B^{(k)}-W^{(k)})e^{(k)} + \mathrm{o}(\|e^{(k)}\|).$$

Hence

$$u^{(k+1)} - u^* = \mathrm{O}(\|e^{(k)}\|), \tag{6.6}$$

as (4.1) implies $A^{(k)}\Phi^*$ has full rank for $x^{(k)}$ sufficiently close to x^* and convergence of the sequence $\{x^{(k)}\}$ implies $e^{(k+1)} = \mathrm{O}(\|e^{(k)}\|)$. Consequently (4.10) reduces to

$$Q^{(k)}e^{(k+1)} = \mathrm{O}(\|e^{(k)}\|^2), \tag{6.7}$$

and

$$P^{(k)}B^{(k)}e^{(k+1)} = P^{(k)}(B^{(k)}-W^{(k)})e^{(k)} + \mathrm{O}(\|e^{(k)}\|^2). \tag{6.8}$$

Now for $x^{(k)}$ sufficiently close to x^* (4.3) implies

$$B_R^{(k)} = Z^{(k)\mathrm{T}}B^{(k)}Z^{(k)} \tag{6.9}$$

is uniformly positive definite. Thus (6.2), (6.7), (6.8) and $P^{(k)} + Q^{(k)} = I$ yield

$$P^{(k)}e^{(k+1)} = Z^{(k)}B_R^{(k)^{-1}}Z^{(k)\mathrm{T}}(B^{(k)}-W^{(k)})e^{(k)} + \mathrm{O}(\|e^{(k)}\|^2). \tag{6.10}$$

As $B_R^{(k)}$ is uniformly positive definite and $e^{(k)} = x^{(k)} - x^*$ equations (6.7) and (6.10) show that (6.4) is equivalent to the superlinear rate of convergence condition (2.11), and that (6.5) is equivalent to the second order rate of convergence condition (2.15).

Remarks. (i) As $\delta^{(k)} = x^{(k+1)} - x^{(k)}$ it follows from (2.12) that when the sequence converges superlinearly $\delta^{(k)} = -(x^{(k)} - x^*) + \mathrm{o}(\|x^{(k)} - x^*\|)$. Thus (6.4) is equivalent to

$$\lim_{k\to\infty} \frac{\|P^{(k)}(W^{(k)}-B^{(k)})\delta^{(k)}\|}{\|\delta^{(k)}\|} = 0. \tag{6.11}$$

Also $P^{(k)} \to P^*$, $\lambda^{(k)} \to \lambda^*$ so $W^{(k)} \to W^*$ as $x^{(k)} \to x^*$. Thus (6.11) is equivalent to

$$\lim_{k\to\infty} \frac{\|P^*(W^*-B^{(k)})\delta^{(k)}\|}{\|\delta^{(k)}\|} = 0, \tag{6.12}$$

which corresponds to the results of Powell and Yuan [20] when $h(c) = \|c\|_1$ and $h(c) = \|c\|_\infty$.

(ii) When the sequence is converging at a second order rate it follows from (2.16) that $\delta^{(k)} = -(x^{(k)} - x^*) + O(\|x^{(k)} - x^*\|^2)$. Thus (6.5) is equivalent to

$$\lim_{k\to\infty} \frac{\|P^{(k)}(W^{(k)} - B^{(k)})\delta^{(k)}\|}{\|\delta^{(k)}\|^2} < \infty.$$

As above $P^{(k)}$ and $W^{(k)}$ may be replaced by P^* and W^* respectively.

Condition (6.4) has the disadvantage that x^* is not known when $B^{(k)}$ is calculated, whilst (6.11) has the similar disadvantage that $\delta^{(k)}$ is not known. However sufficient conditions which contain only information available at the point $x^{(k)}$ are easily obtained from theorem 6.1.

Corollary 6.1. *Let the assumptions A be satisfied. Then the rate of convergence is*

(a) *superlinear if*

$$\lim_{k\to\infty} P^{(k)}(W^{(k)} - B^{(k)}) = 0, \tag{6.13}$$

(b) *quadratic if for all k sufficiently large*

$$P^{(k)}(W^{(k)} - B^{(k)}) = 0. \tag{6.14}$$

Proof. The result comes directly from Theorem 6.1. □

Remarks. (i) As $P^{(k)} + Q^{(k)} = I$, equation (6.13) can be resolved in two components, namely

$$\lim_{k\to\infty} P^{(k)}(W^{(k)} - B^{(k)})P^{(k)} = 0 \quad \text{and} \quad \lim_{k\to\infty} P^{(k)}(W^{(k)} - B^{(k)})Q^{(k)} = 0 \tag{6.15}$$

and similarly for (6.14).

(ii) As before the matrices $P^{(k)}$, $Q^{(k)}$ and $W^{(k)}$ can be replaced by P^*, Q^* and W^* respectively to yield equivalent results.

These conditions can easily be expressed in terms of the matrices $Z^{(k)}$ and $Y^{(k)}$ using (6.2) and their orthogonality. The disadvantage of this form is the possible lack of continuity in the matrices $Z^{(k)}$. If it is assumed that $Z^{(k)} \to Z^*$ as $x^{(k)} \to x^*$ then (6.13) and (6.14) are equivalent to

$$\lim_{k\to\infty} Z^{(k)\mathrm{T}} B^{(k)} Z^{(k)} = Z^{*\mathrm{T}} W^* Z^*$$

and

$$\lim_{k\to\infty} Z^{(k)\mathrm{T}} B^{(k)} Y^{(k)} = Z^{*\mathrm{T}} W^* Y^*.$$

Thus if the reduced curvature term $Z^{(k)\mathrm{T}} B^{(k)} Z^{(k)}$ and the cross curvature term $Z^{(k)\mathrm{T}} B^{(k)} Y^{(k)}$ are correct in the limit one obtains superlinear convergence. Similarly (6.15) shows that if this reduced curvature term and this cross curvature term are exact one obtains a second order rate of convergence. An example of an algorithm that exploits this result is that of Fletcher [8] for an exact l_1 penalty function.

If one uses only the reduced curvature information $P^{(k)}B^{(k)}P^{(k)}$ then significant reductions in storage and gains in efficiency can be obtained for problems with large n and $n-l^* \ll n$. Murray and Overton for minimax [13] and l_1 [14] calculations, Coleman and Conn [3] for an exact l_1 penalty function, and Womersley and Fletcher [22] have proposed methods of this type. When $h(c)=\|c\|_\infty$, Han [10] gives a sufficient condition for a two-step superlinear rate of convergence, namely

$$\lim_{k\to\infty} \frac{\|P^*(W^*-B^{(k)})P^*\delta^{(k)}\|}{\|\delta^{(k)}\|}=0. \tag{6.16}$$

This condition also comes directly from Powell's work [18]. The following result generalizes this result to provide necessary and sufficient conditions for a two-step superlinear rate of convergence using only reduced curvature information.

Theorem 6.2. *Let the assumptions A be satisfied. The rate of convergence is two-step superlinear if and only if*

$$\lim_{k\to\infty} \frac{\|P^{(k)}(W^{(k)}-B^{(k)})P^{(k)}(x^{(k)}-x^*)\|}{\|x^{(k-1)}-x^*\|}=0. \tag{6.17}$$

Proof. From (6.8)

$$P^{(k)}B^{(k)}e^{(k+1)}=P^{(k)}(B^{(k)}-W^{(k)})(P^{(k)}+Q^{(k)})e^{(k)}+\mathrm{O}(\|e^{(k)}\|^2),$$

and, from the continuity properties of the $Q^{(k)}$ and (6.7),

$$\begin{aligned}Q^{(k)}e^{(k)}&=Q^{(k-1)}e^{(k)}+(Q^{(k)}-Q^{(k-1)})e^{(k)}\\&=\mathrm{O}(\|e^{(k-1)}\|^2)+\mathrm{o}(\|e^{(k)}\|)=\mathrm{o}(\|e^{(k-1)}\|),\end{aligned}$$

as convergence implies $e^{(k)}=\mathrm{O}(\|e^{(k-1)}\|)$. Hence

$$P^{(k)}e^{(k+1)}=Z^{(k)}B_R^{(k)^{-1}}Z^{(k)\mathrm{T}}(B^{(k)}-W^{(k)})P^{(k)}e^{(k)}+\mathrm{o}(\|e^{(k-1)}\|). \tag{6.18}$$

Also, from (6.7),

$$Q^{(k)}e^{(k+1)}=\mathrm{O}(\|e^{(k)}\|^2)=\mathrm{O}(\|e^{(k-1)}\|^2). \tag{6.19}$$

As the matrices $B_R^{(k)}$ are uniformly positive definite equations (6.18) and (6.19) show that (6.17) is equivalent to the two-step superlinear rate of convergence condition (2.13). □

Remarks. (i) From (2.14) a two-step superlinear rate of convergence is equivalent to $\delta^{(k)}=-e^{(k)}+\mathrm{o}(\|e^{(k-1)}\|)$. Also $e^{(k)}=e^{(k-1)}+\delta^{(k-1)}$ so (6.17) is equivalent to

$$\lim_{k\to\infty} \frac{\|P^{(k)}(W^{(k)}-B^{(k)})P^{(k)}\delta^{(k)}\|}{\|\delta^{(k-1)}+\delta^{(k)}\|}=0. \tag{6.20}$$

Han's condition (6.16) is obviously sufficient for (6.20) to hold.

(ii) A sufficient condition for a two-step superlinear rate of convergence is that

$$\lim_{k\to\infty} P^{(k)}(W^{(k)}-B^{(k)})P^{(k)}=0, \tag{6.21}$$

so that in the limit the reduced curvature information is correct.

(iii) As before the matrices $P^{(k)}$ and $W^{(k)}$ can be replaced by P^* and W^* to obtain expressions equivalent to (6.17), (6.20) and (6.21).

When $l^*>0$ the surface

$$\pi^*=\{x\in\mathbb{R}^n\colon\ \Phi^{*\mathrm{T}}c(x)=\Phi^{*\mathrm{T}}c^*\}, \tag{6.22}$$

is the surface of nondifferentiability of F passing through x^*. From (4.9) the solution of the subproblem, which linearizes these equations, satisfies

$$\Phi^{*\mathrm{T}}(c^{(k)}+A^{(k)}\delta^{(k)})=\Phi^{*\mathrm{T}}c^*. \tag{6.23}$$

This produces the second order rate of convergence (6.7) in directions orthogonal to the surface π^*. When only reduced curvature information is used this second order rate of convergence in directions orthogonal to π^* on the second step corrects any errors caused by the lack of the cross curvature information $P^{(k)}B^{(k)}Q^{(k)}$, producing a two-step superlinear rate of convergence.

In the proofs of the preceding results the step $\delta^{(k)}$ has been split into two components, one orthogonal to the surface π^* at $x^{(k)}$ and one tangential to the surface π^* at $x^{(k)}$, giving

$$\delta^{(k)}=P^{(k)}\delta^{(k)}+Q^{(k)}\delta^{(k)}=v^{(k)}+w^{(k)}. \tag{6.24}$$

Some algorithms (for example [3] and [22]) generate $\delta^{(k)}$ by generating the components $v^{(k)}$ and $w^{(k)}$ separately. As long as $w^{(k)}$ is generated so that (6.7) is satisfied the results of this section still hold. One possibility is the Newton-like step

$$w^{(k)}=(A^{(k)}\Phi^*)^{+\mathrm{T}}\Phi^{*\mathrm{T}}(c^*-c^{(k)})=Y^{(k)}R^{(k)-\mathrm{T}}\Phi^{*\mathrm{T}}(c^*-c^{(k)}), \tag{6.25}$$

where $V^{+\mathrm{T}}$ denotes the generalized inverse of V transposed, produced by (6.23). An alternative, used by Coleman and Conn [3] because of global considerations, is to replace $c^{(k)}$ by $c(x^{(k)}+v^{(k)})$ in (6.23) and (6.25). Using a Taylor series expansion, (6.24) and the fact that $\delta^{(k)}=e^{(k+1)}-e^{(k)}=\mathrm{O}(\|e^{(k)}\|)$ one has

$$c(x^{(k)}+v^{(k)})+A^{(k)\mathrm{T}}\delta^{(k)}-c^*=A^{(k)\mathrm{T}}e^{(k+1)}+A^{(k)\mathrm{T}}P^{(k)}\delta^{(k)}+\mathrm{O}(\|e^{(k)}\|^2).$$

Thus (6.23) with $c(x^{(k)}+v^{(k)})$ replacing $c^{(k)}$ gives

$$(A^{(k)}\Phi^*)^{\mathrm{T}}e^{(k+1)}=-(A^{(k)}\Phi^*)^{\mathrm{T}}P^{(k)}\delta^{(k)}+\mathrm{O}(\|e^{(k)}\|^2),$$

which, using (6.2) and (6.3), implies (6.7).

If the surface π^* is linear then one can obtain a superlinear (rather than two-step superlinear) rate of convergence using only reduced curvature information. Note that the surface π^* being linear does not necessarily imply the functions c are linear, but that $\Phi^{*\mathrm{T}}c(x)$ is linear. For example if $h(c)=\|c\|_\infty$ this condition is satisfied if the functions $c_i(x)$, $i\in\mathcal{A}^*$ are quadratic with identical Hessians.

Corollary 6.2. *Let the assumptions A be satisfied. If the surface π^* is linear then the rate of convergence is*

(a) *superlinear if and only if*

$$\lim_{k\to\infty} \frac{\|P^{(k)}(W^{(k)}-B^{(k)})P^{(k)}(x^{(k)}-x^*)\|}{\|x^{(k)}-x^*\|}=0, \tag{6.26}$$

(b) *quadratic if and only if*

$$\lim_{k\to\infty} \frac{\|P^{(k)}(W^{(k)}-B^{(k)})P^{(k)}(x^{(k)}-x^*)\|}{\|x^{(k)}-x^*\|^2}<\infty. \tag{6.27}$$

Proof. As π^* is linear it follows that for K sufficiently large a step of $\delta^{(K-1)}$ results in $x^{(K)}$, and hence all successive iterates, lying on π^*. Also $P^{(k)}=P^*$ and $Q^{(k)}=Q^*$ for all $k\geqslant K$, hence

$$Q^{(k)}e^{(k+1)}=Q^{(k)}e^{(k)}=0 \quad \forall k\geqslant K.$$

Equation (6.10) then reduces to

$$P^{(k)}e^{(k+1)}=Z^{(k)}B_R^{(k)^{-1}}Z^{(k)\mathrm{T}}(W^{(k)}-B^{(k)})P^{(k)}e^{(k)}+\mathrm{O}(\|e^{(k)}\|^2),$$

which gives the desired results, as $B_R^{(k)}$ is uniformly positive definite. □

Remark. As before the vectors $x^{(k)}-x^*$ in (6.26) and (6.27) can be replaced by $\delta^{(k)}$, whilst $P^{(k)}$ and $W^{(k)}$ can be replaced by P^* and W^* respectively. Also as before there are obvious sufficient conditions for (6.26) and (6.27) which do not involve the vector $x^{(k)}-x^*$.

7. Smooth problems

If h is smooth then $\partial h(c)=\{\nabla h(c)\}$, so the multipliers $\lambda^{(k+1)}$ generated by the subproblem (1.2) are given by $\lambda^{(k+1)}=\nabla h(c^{(k)}+A^{(k)}\delta^{(k)})$. If h is twice continuously differentiable and its second derivatives satisfy a Lipschitz condition then Taylor series expansions can be used to show that when $B^{(k)}=W^{(k)}$

$$\Psi(\delta;x^{(k)})=F^{(k)}+\delta^{\mathrm{T}}\nabla F^{(k)}+\tfrac{1}{2}\delta^{\mathrm{T}}\nabla^2F^{(k)}\delta+\mathrm{O}(\|A^{(k)\mathrm{T}}\delta\|^3), \tag{7.1}$$

where $\Psi(\delta;x^{(k)})$ is given by (1.2). Thus methods based on minimizing $\Psi(\delta;x^{(k)})$ have the same local properties as Newton's method.

Again Taylor series expansions produce a result corresponding a (4.10), namely

$$\Gamma^{(k)}e^{(k+1)}=(B^{(k)}-W^{(k)})e^{(k)}+\mathrm{O}(\|e^{(k)}\|^2)+\mathrm{O}(\|A^{(k)\mathrm{T}}\delta^{(k)}\|^2), \tag{7.2}$$

where $\Gamma^{(k)}=B^{(k)}+A^{(k)}\nabla^2h(c^{(k)})A^{(k)\mathrm{T}}$. Let $\mu^{(k)}=\nabla h(c^{(k)})$ then

$$\nabla^2F^{(k)}=G^{(k)}+\sum_{i=1}^{m}\mu_i^{(k)}\nabla^2c_i^{(k)}+A^{(k)}\nabla^2h(c^{(k)})A^{(k)\mathrm{T}}. \tag{7.3}$$

Second order necessary conditions ensure $\nabla^2 F^*$ is positive semidefinite, whilst second order sufficient conditions require that $\nabla^2 F^*$ is positive definite. Thus for $x^{(k)}$ sufficiently close to x^* a reasonable condition to impose upon a $B^{(k)}$ which is to approximate $W^{(k)}$ is that

$$s^{\mathrm{T}}(B^{(k)}+A^{(k)}\nabla^2 h(c^{(k)})A^{(k)\mathrm{T}})s \geqslant \beta s^{\mathrm{T}} s > 0 \quad \forall s \in \mathbb{R}^n,\ s \neq 0. \tag{7.4}$$

This corresponds to condition (4.3) or (4.4) in the nonsmooth case, and guarantees a stationary point of (1.2) is a minimizer of (1.2).

If (7.4) holds so $\Gamma^{(k)}$ is uniformly positive definite, equation (7.2) provides necessary and sufficient conditions for a superlinear rate of convergence corresponding to that of Powell [19] and Dennis and Moré [6].

Theorem 7.1. *Let f, c and h be twice continuously differentiable, and let their second derivatives satisfy a Lipschitz condition. Let* $\delta^{(k)}$ *be a stationary point of* (1.2) *where the matrices* $B^{(k)}$ *satisfy* (7.4) *so* $\delta^{(k)}$ *is a minimizer of* (1.2). *Let the sequence* $\{x^{(k)}\}$ *generated by* $x^{(k+1)} = x^{(k)} + \delta^{(k)}$ *converge to a stationary point* x^* *of F where* $\nabla^2 F^*$ *is positive definite. Then the rate of convergence is*

(a) *superlinear if and only if*

$$\lim_{k\to\infty} \frac{\|(B^{(k)} - W^{(k)})(x^{(k)} - x^*)\|}{\|x^{(k)} - x^*\|} = 0, \tag{7.5}$$

(b) *second order if and only if*

$$\lim_{k\to\infty} \frac{\|(B^{(k)} - W^{(k)})(x^{(k)} - x^*)\|}{\|x^{(k)} - x^*\|^2} < \infty. \tag{7.6}$$

Proof. As the sequence converges $\delta^{(k)} = e^{(k+1)} - e^{(k)} = \mathrm{O}(\|e^{(k)}\|)$, and as the matrices $\Gamma^{(k)}$ are uniformly positive definite equation (7.2) shows that (7.5) is equivalent to the superlinear rate of convergence condition (2.11) and that (7.6) is equivalent to the second order rate of convergence condition (2.15). □

Remarks. (i) The matrix

$$W^{(k)} = G^{(k)} + \sum_{i=1}^{m} \lambda_i^{(k)} \nabla^2 c_i^{(k)}$$

can be generated using either $\lambda^{(k)} = \nabla h(c^{(k)})$ in $W^{(k)}$ or $\lambda^{(k+1)} = \nabla h(c^{(k)} + A^{(k)\mathrm{T}}\delta^{(k)})$ in $W^{(k+1)}$.

(ii) As before $\delta^{(k)}$ can replace $e^{(k)}$ and W^* replace $W^{(k)}$ in (7.5) and (7.6) to obtain equivalent results. Also there are obvious sufficient conditions for (7.5) and (7.6) which do not involve $e^{(k)}$ or $\delta^{(k)}$.

If h is only once continuously differentiable the subproblem (1.2) is well-defined although Newton's method is not. Equation (4.16) with $\lambda^{(k+1)} = \nabla h(c^{(k)} + A^{(k)\mathrm{T}}\delta^{(k)})$ and $\lambda^* = \nabla h(c^*)$ still holds, but the possibilities of discontinuities in the second derivatives of h makes it difficult to estimate the rate at which $\lambda^{(k+1)} \to \lambda^*$ as $x^{(k)} \to x^*$.

Minimizing the subproblem (1.2) may also face difficulties caused by discontinuities in the second derivatives of h.

Finally it is important to note that all the results of Sections 5 and 6 apply directly to the nonlinear programming problem (1.4), when the subproblem (1.5) is used to generate $\delta^{(k)}$. In the context of nonlinear programming problems many of these results are known (see Powell [18] and Boggs, Tolle and Wang [1]). The matrix Φ^* of gradients of the structure functionals at x^* simply corresponds to a matrix which picks out the active constraints at x^*. Let $m = |E| + |I|$ and

$$\mathcal{A}(x) = \{i \in E \cup I\colon c_i(x) = 0\}.$$

Then with $l^* = |\mathcal{A}^*|$ and e_i the ith unit vector in $\mathbb{R}^m$ one has

$$\Phi^* = [e_i, i \in \mathcal{A}^*].$$

As $A(x) = [\nabla c_i(x), i = 1, \ldots, m]$ the nondegeneracy assumption (4.1) that the $n \times l^*$ matrix $A^* \Phi^*$ has rank l^* simply corresponds to an assumption that the gradients of the active constraints at the solution are linearly independent. The set U^* becomes

$$U^* = \{u \in \mathbb{R}^{l^*}\colon u_i \geqslant 0 \text{ for } i \in \mathcal{A}^* \cap I\},$$

and $\lambda_0^* = 0$. The set U^* is nonempty, closed and convex but is not bounded as in the nonsmooth problem. The strict complementarity condition (4.2) corresponds to the condition that $\lambda_i^* > 0$ $\forall i \in \mathcal{A}^* \cap I$. The condition (4.3) on the matrices $B^{(k)}$ requires that they are uniformly positive definite on the tangent space to the active constraints at the solution. The surface of nondifferentiability of F which passes through x^* (namely π^* in (6.22) corresponds to the constraint surface

$$\pi^* = \{x \in \mathbb{R}^n\colon c_i(x) = c_i^* \; \forall i \in \mathcal{A}^*\}.$$

Lemma 4.1 corresponds to the well known result that under nondegeneracy and strict complementarity conditions the set of active constraints at the solution of the subproblem (1.5) is $\mathcal{A}^*$ when $x^{(k)}$ is sufficiently close to x^*, or equivalently

$$\Phi^{*\mathrm{T}}(c^{(k)} + A^{(k)\mathrm{T}}\delta^{(k)}) = \Phi^{*\mathrm{T}}c^*.$$

This is precisely (4.9), which in turn gives (4.19). The Lagrange multipliers at the solution $\delta^{(k)}$ of the subproblem (1.5) can be used as the multiplier estimates $\lambda^{(k+1)}$.

With the above correspondences Lemma 4.2 and all the results of Sections 5 and 6 apply to the nonlinear programming problem (1.4) when it is solved using the subproblem (1.5). For example Theorem 6.2 gives necessary and sufficient conditions for a two-step superlinear rate of convergence using only reduced curvature information when successive quadratic programming methods are used to solve nonlinear programming problems.

Acknowledgement

The detailed comments of a referee are gratefully acknowledged.

References

[1] P.T. Boggs, J.W. Tolle and P. Wang, "On the local convergence of quasi-Newton methods for constrained optimization", *SIAM Journal on Control and Optimization* **20** (1982) 161-171.

[2] R.M. Chamberlain, M.J.D. Powell, C. Lemaréchal and H.C. Pederson, "The watchdog technique in forcing convergence in algorithms for constrained optimization", *Mathematical Programming Study* **16** (1982) 1-17.

[3] T.F. Coleman and A.R. Conn, "Nonlinear programming via an exact penalty function: Asymptotic analysis", *Mathematical Programming* **24** (1982) 123-136.

[4] T.F. Coleman and D.C. Sorensen, "A note on the computation of an orthonormal basis for the null space of a matrix", *Mathematical Programming* **29** (1984) 234-242.

[5] L. Cromme, "Strong uniqueness. A far reaching criterion for the convergence analysis of iterative procedures", *Numerische Mathematik* **29** (1978) 179-194.

[6] J.E. Dennis and J.J. Moré, "A characterization of superlinear convergence and its application to quasi-Newton methods", *Mathematics of Computation* **28** (1974) 549-560.

[7] R. Fletcher, *Practical methods of optimization, volume 2: Constrained optimization* (Wiley, Chichester and New York, 1981).

[8] R. Fletcher, "Numerical experiments with an exact l_1 penalty function method", in: O.L. Mangasarian, R.R. Meyer and S.M. Robinson, eds., *Nonlinear programming 4* (Academic Press, New York and London, 1981).

[9] R. Fletcher, "Second order corrections for nondifferentiable optimization", in: G.A. Watson, ed., *Numerical analysis proceedings, Dundee 1981*, Lecture Notes in Mathematics 912 (Springer-Verlag, 1982).

[10] S.P. Han, "Superlinear convergence of a minimax method", Report TR-78-336, Department of Computer Science, Cornell University (1978).

[11] K. Jittorntrum and M.R. Osborne, "Strong uniqueness and second order convergence in nonlinear discrete approximation", *Numerische Mathematik* **34** (1980) 439-455.

[12] D.Q. Mayne, "On the use of exact penalty functions to determine step length in optimization algorithms", in: G.A. Watson, ed., *Numerical analysis, Dundee 1979*, Lecture notes in mathematics 773 (Springer-Verlag, Berlin, 1980).

[13] W. Murray and M.L. Overton, "A projected Lagrangian algorithm for nonlinear minimax optimization", *SIAM Journal on Scientific and Statistical Computing* **1** (1980) 345-370.

[14] W. Murray and M.L. Overton, "A projected Lagrangian algorithm for nonlinear L_1 optimization", *SIAM Journal on Scientific and Statistical Computing* **2** (1981) 207-224.

[15] M.R. Osborne, "Algorithms for nonlinear discrete approximation", in: C.T.H. Baker and C. Phillips, eds., *The numerical solution of nonlinear problems* (Clarendon Press, Oxford, 1981).

[16] M.R. Osborne, "Strong uniqueness in nonlinear approximation", in: C.T.H. Baker and C. Phillips, eds., *The numerical solution of nonlinear problems* (Clarendon Press, Oxford, 1981).

[17] M.R. Osborne, *Finite algorithms in optimization and data analysis* (Wiley, in press).

[18] M.J.D. Powell, "The convergence of variable metric methods for nonlinearly constrained optimization calculations", in: O.L. Mangasarian, R.R. Meyer and S.M. Robinson, eds., *Nonlinear Programming 3* (Academic Press, New York, 1978).

[19] M.J.D. Powell, "General algorithms for discrete nonlinear approximation calculations", Report DAMTP 1982/NA5, University of Cambridge (to be published by Academic Press in Proceedings of the fourth Texas symposium on approximation theory, College Station, TX, 1983).

[20] M.J.D. Powell and Y. Yuan, "Conditions for superlinear convergence in l_1 and l_∞ solutions of overdetermined nonlinear equations", IMA Journal on Numerical Analysis 4 (1984) 241-251.

[21] R.S. Womersley, "Minimizing nonsmooth composite functions", Research Report No. 13-1984, Mathematical Sciences Research Centre, Australian National University (Canberra, Australia, 1984).

[22] R.S. Womersley and R. Fletcher, "An algorithm for composite nonsmooth optimization problems", Report NA/60, Department of Mathematics, University of Dundee (Dundee, Scotland, 1982).

[23] Y. Yuan, "Global convergence of trust region algorithms for nonsmooth optimization", Report DAMTP 1983/NA13, University of Cambridge (1983).

[24] Y. Yuan, "An example of only linear convergence of trust region algorithms for nonsmooth optimization", *IMA Journal of Numerical Analysis 4* (1984) 327-335.