Mathematical Programming 71 (1995) 179–194

A Gauss–Newton method for convex composite optimization [1]

J.V. Burke [a], M.C. Ferris [b,*]

[a] *Department of Mathematics, GN-50, University of Washington, Seattle, WA 98195, United States*

[b] *Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706, United States*

Received 28 September 1993; revised manuscript received December 1994

Abstract

An extension of the Gauss–Newton method for nonlinear equations to convex composite optimization is described and analyzed. Local quadratic convergence is established for the minimization of $h \circ F$ under two conditions, namely h has a set of weak sharp minima, C, and there is a regular point of the inclusion $F(x) \in C$. This result extends a similar convergence result due to Womersley (this journal, 1985) which employs the assumption of a strongly unique solution of the composite function $h \circ F$. A backtracking line-search is proposed as a globalization strategy. For this algorithm, a global convergence result is established, with a quadratic rate under the regularity assumption.

Keywords: Gauss–Newton; Convex composite optimization; Weak sharp minima; Quadratic convergence

1. Introduction

In the early nineteenth century, Gauss proposed a powerful method for solving systems of nonlinear equations that generalized the classical Newton's method for such systems. Recall that Newton's method is based on successive linearization. Unfortunately, in many applications the linearized systems can be inconsistent. To remedy this problem, Gauss proposed that the iterates be based on the least-squares solutions to the linearized problems. In making the transition from solving the linearization to solving the associated

* Corresponding author. e-mail: ferris@cs.wisc.edu.

[1] This material is based on research supported by National Science Foundation Grants CCR-9157632 and DMS-9102059, the Air Force Office of Scientific Research Grant F49620-94-1-0036, and the United States–Israel Binational Science Foundation Grant BSF-90-00455.

SSDI 0025-5610(95)00010-0

linear least-squares problem, the underlying problem changes from equation solving to minimization. Specifically, the *Gauss–Newton* method solves the associated nonlinear least-squares problem and can therefore converge to a solution to the nonlinear least-squares problem that is not a solution to the underlying system of equations. Nonetheless, the method is always implementable and can be made significantly more robust by the addition of a line-search. Other variations that enhance the robustness of the method are the addition of a quadratic term to the objective in the step-finding subproblem (see [20,27]) or the inclusion of a trust-region constraint (see [13]).

In this paper we discuss the extension of the Gauss–Newton methodology to finite-valued convex composite optimization. Convex composite optimization refers to the minimization of any extended real-valued function that can be written as the composition of a convex function with a function of class C^1:

$$(\mathcal{P}) \qquad \min_x f(x) := h(F(x)),$$

where $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is convex and $F : \mathbb{R}^n \to \mathbb{R}^m$ is of class C^1. We consider only the finite-valued case: $h : \mathbb{R}^m \to \mathbb{R}$. Obviously the nonlinear least-squares problem is precisely of this form. It is interesting to note that in their outline of the Gauss–Newton method, Ortega and Rheinboldt [28, p. 267] used the notion of a composite function. A wide variety of applications of this formulation can be found throughout the mathematical programming literature [6,16,17,21,31,39,40], e.g., nonlinear inclusions, penalization methods, minimax, and goal programming. The convex composite model provides a unifying framework for the development and analysis of algorithmic solution techniques. Moreover, it is also a convenient tool for the study of first- and second-order optimality conditions in constrained optimization [7,9,17,39].

Our extension of the Gauss–Newton methodology to finite-valued convex composite optimization is based on the development given in [5,18], which specifically address the problem of solving finite-dimensional systems of nonlinear equations and inequalities. In this case, much more can be said about the algorithmic design and this is done in the cited articles. In this article we focus on the general theory. Specifically, we extend a result due to Womersley [40] establishing the quadratic convergence of a Gauss–Newton method under the assumption of strong uniqueness. An important distinction between these results is that we do not require that the solution set be a singleton or even a bounded set. A further discussion of the relationship to Womersley's result is given at the end of Section 3.

The approach we take requires two basic assumptions: (1) the set of minima for the function h, denoted by C, is a set of *weak sharp minima* for h, and (2) there is a *regular* point for the inclusion

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

In this article, we provide a self-contained and elementary proof theory in the finite-dimensional case. The basic algorithm is discussed in Section 2. After a discussion of regularity in Section 3, we establish the local quadratic convergence of the method in

Section 4. The results of Section 4 can be extended to the setting of reflexive Banach spaces under a suitable strengthening of the regularity condition (see condition (10)). We conclude in Section 5 by establishing the convergence properties of a globalization strategy based on a backtracking line-search.

The notation that we employ is for the most part standard; however, a partial list is provided for the readers' convenience. The *inner product* on $\mathbb{R}^n$, $\langle y, x\rangle$, set addition $U \pm \beta V$ and set difference $U \setminus V$ are standard. The *polar* of $U \subset \mathbb{R}^n$ is the set $U^\circ := \{x^* \in \mathbb{R}^n \mid \langle x^*, x\rangle \leqslant 1, \forall x \in U\}$. The *relative interior* of U, $\operatorname{ri} U$, is the interior of U relative to the *affine hull* of U and the *closure* of U, $\operatorname{cl} U$, is the usual topological closure of the set U. The *cone* generated by U is $\operatorname{cone}(U) := \{\lambda u \mid \lambda > 0, u \in U\}$. The *indicator* function for U is the function $\psi_U(x)$ taking the value 0 when x is in U and $+\infty$ otherwise. The support function for U is given by $\psi_U^*(x) := \sup\{\langle x^*, x\rangle \mid x^* \in U\}$.

We denote a *norm* on $\mathbb{R}^\nu$ by $\|\cdot\|$, its closed unit ball by $\mathbb{B}$, and its dual norm by $\|x\|_\circ := \psi_{\mathbb{B}}^*(x)$. It is straightforward to show that the unit ball associated with the dual norm is $\mathbb{B}^\circ$. The distance of a point x to a set U is given by $\operatorname{dist}(x \mid U) := \inf\{\|x-u\| \mid u \in U\}$. Finally, the sets $\ker A$ and $\operatorname{im} A$ represent the kernel and image space of the linear map A, respectively, and the inverse image of a set U under the mapping A is given by $A^{-1}U := \{y \mid Ay \in U\}$.

2. The basic algorithm

Let $f(x) := h(F(x))$ be as given in $(\mathcal{P})$ with h finite-valued. The basic Gauss–Newton procedure takes a unit step along a direction selected from the following set:

$$D_\Delta(x) := \arg\min\{h(F(x) + F'(x)d) \mid \|d\| \leqslant \Delta\}, \tag{2}$$

which represents the set of solutions to the minimization problem

$$\min\{h(F(x) + F'(x)d) \mid \|d\| \leqslant \Delta\}. \tag{3}$$

There are two points to note. The first is that the "linearization" is carried out only on the smooth function F, the convex function h is treated explicitly, corresponding exactly to the Gauss–Newton methodology. The second point is that the directions are constrained to have length no greater than Δ. This is different from the standard Gauss–Newton procedure which can be recovered by setting $\Delta = \infty$. Nonetheless, from the standpoint of convergence analysis it is advantageous to take Δ finite. Observe that D_Δ is a multifunction taking points x and generating a set of directions. The basic algorithm to be considered here is as follows.

Algorithm 1. Let $\eta \geqslant 1$, $\Delta \in (0, +\infty]$ and $x^0 \in \mathbb{R}^n$ be given. Having x^k, determine x^{k+1} as follows.

If $h(F(x^k)) = \min\{h(F(x^k) + F'(x^k)d) \mid \|d\| \leqslant \Delta\}$, then stop; otherwise, choose $d^k \in D_\Delta(x^k)$ to satisfy

$$\|d^k\| \leqslant \eta \operatorname{dist}(0 \mid D_\Delta(x^k)), \tag{4}$$

and set $x^{k+1} := x^k + d^k$.

Algorithms of this type have been extensively studied in the literature [5,18,23,33]. If it is assumed that both h and the norm on $\mathbb{R}^n$ are polyhedral, then one can obtain a direction choice satisfying (4) in Algorithm 1 by computing a least-norm solution of a linear program in the sense of [24,25]. Numerical methods for obtaining least-norm solutions to linear programs have been developed in [12]. If h is piecewise linear-quadratic and the norm on $\mathbb{R}^n$ is either polyhedral or quadratic, then a two-stage procedure can be employed to obtain the step d^k. In the first stage one obtains an optimal solution to (3), d_{opt}, then in the second stage a least-norm solution is obtained by solving the linear or quadratic program $\min\{\|d\| \mid F'(x)d = F'(x)d_{\text{opt}}\}$. For example, when the norm on $\mathbb{R}^n$ is polyhedral, the algorithm given in [2] will solve (3). If h is the distance function to a nonempty closed convex set, one can apply the relaxation techniques described in [5].

In most studies, the objective function in (3) includes a quadratic term of the form $\frac{1}{2}d^{\mathrm{T}}Hd$ in order to incorporate some curvature components. Such methods are commonly referred to as *sequential quadratic programming* methods. In addition to a second-order sufficiency condition, these methods require a strong nondegeneracy condition for their local analysis. These conditions combine to imply that the point under consideration is an isolated stationary point of the problem. Consequently, this theory does not apply to the class of problems addressed in this article. Further discussion of this curvature component can be found in [13,28] for the classical Gauss-Newton method and in [17,31] for convex composite optimization. The relationship of this component to second-order optimality conditions can be found in [9,17,39]. In this article, we avoid the need for a curvature term by focusing on the local behavior of the algorithm in the neighborhood of a point $\bar{x}$ satisfying $F(\bar{x}) \in C := \arg\min h$, assumed nonempty.

Our analysis is based on two key assumptions: the set C is a set of *weak sharp minima* for the function h and the point $\bar{x}$ is a *regular* point (see Section 3) for the inclusion (1). The weak sharp minima concept was introduced in [14].

Definition 2.1. The set $C \subset \mathbb{R}^m$ is a set of *weak sharp minima* for the function $h : \mathbb{R}^m \to \mathbb{R} \cup \{\pm\infty\}$ if there is an $\alpha > 0$ such that

$$h(y) \geqslant h_{\min} + \alpha \operatorname{dist}(y \mid C), \quad \text{for all } y \in \mathbb{R}^m, \tag{5}$$

where $h_{\min} := \min_y h(y)$. The constant α and the set C are called the *modulus* and *domain of sharpness* for h over C, respectively.

Note that in finite dimensions, if inequality (5) is satisfied for one choice of norm, then it is satisfied for every other norm with perhaps a different choice of α. The prototypical example of a function h having a set of weak sharp minima is the distance function $\operatorname{dist}(\cdot \mid C)$ itself; other examples are explored in [8,14]. The notion of weak sharp minima generalizes the notion of a *sharp* [30] or *strongly unique* [11,21,29,40] minimum. These concepts have a long history in the literature and have far-reaching consequences for the convergence analysis of many iterative procedures [11,19,21,30,40]. In [8], it was shown that some of these convergence results can be extended to the case of weak sharp minima. This article continues this discussion in the context of convex composite optimization. Whereas in the fully convex case one obtains finite termination criteria, in the convex composite case we can establish quadratic convergence when regularity is also assumed.

3. Regularity

In this section, we define a notion of regularity for the inclusion (1) that can be applied at any point in $\mathbb{R}^n$. It is related to various notions of regularity that can be found in the literature [4,7,22,26,35,37,39]. In particular, in the finite-dimensional case, it is equivalent to the definition of regularity given by Maguregui [22,23]. Regularity is a convenient tool for relating the set of search directions $D_\Delta(x)$ to the set of solutions of the linearized inclusion

$$F(x) + F'(x)d \in C, \quad \text{with } \|d\| \leqslant \Delta. \tag{6}$$

Regularity also allows us to establish local bounds on the set $D_\Delta(x)$. These bounds are the key to establishing a quadratic convergence result for Algorithm 1 in the next section.

Definition 3.1. A point $\bar{x} \in \mathbb{R}^n$ is a *regular point* for the inclusion (1) if

$$\ker(F'(\bar{x})^{\mathrm{T}}) \cap \Gamma_C(F(\bar{x})) = \{0\}, \tag{7}$$

where the multifunction $\Gamma_C : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ is given by $\Gamma_C(y) := [\operatorname{cone}(C - y)]^\circ$, for all $y \in \mathbb{R}^m$.

The multifunction Γ_C is closely related to the normal cone mapping for C, $N_C(\cdot)$. Indeed, they coincide at points in C; however, $N_C(y) = \emptyset$ at points $y \notin C$. It is straightforward to show that Γ_C has the following very useful dual representation:

$$\Gamma_C(z) = \{y \mid \langle y, z\rangle - \psi_C^*(y) \geqslant 0\}. \tag{8}$$

In the context of the nonlinear least-squares problem, the set C is the origin and so $\Gamma_C(F(\bar{x})) = [F(\bar{x})]^\perp$, the subspace orthogonal to the linear span of the vector $F(\bar{x})$. Therefore condition (7) reduces to the condition $\ker(F'(\bar{x})^{\mathrm{T}}) \cap [F(\bar{x})]^\perp = \{0\}$. If $F(\bar{x}) = 0$, this can be restated as $\operatorname{im}(F'(\bar{x})) = \mathbb{R}^m$.

Our first objective in this section is to establish several equivalent forms of regularity that are pertinent to the discussion.

Lemma 3.2. *Let $\bar{z} \in \mathbb{R}^m$ and $\bar{A} \in \mathbb{R}^{m\times n}$, and suppose that C is a nonempty closed convex subset of $\mathbb{R}^m$. Then the following statements are equivalent.*

(i) $\ker \bar{A}^{\mathrm{T}} \cap \Gamma_C(\bar{z}) = \{0\}$.

(ii) $\operatorname{im}\bar{A} + \operatorname{cone}(\operatorname{ri} C - \bar{z}) = \mathbb{R}^m$.

(iii) *There is a* $\mu > 0$ *such that*

$$0 \in \operatorname{int}(\mu \bar{A}\mathbb{B}_n + (\operatorname{ri} C - \bar{z})),$$

where $\mathbb{B}_n$ *is the unit ball of* $\mathbb{R}^n$.

(iv) *There is a* $\beta > 0$ *such that*

$$(\bar{A}^{\mathrm{T}})^{-1}\mathbb{B}_n^\circ \cap \Gamma_C(\bar{z}) \subset \beta \mathbb{B}_m^\circ. \tag{9}$$

(v) *There is an* $\epsilon > 0$ *such that each of the conditions* (i)–(iv) *above hold for all* $(z, A) \in (\bar{z}, \bar{A}) + \epsilon \mathbb{B}$ *where the unit ball in* $\mathbb{R}^m \times \mathbb{R}^{m\times n}$ *is determined by the norm*

$$\|(z, A) - (\bar{z}, \bar{A})\| = \|z - \bar{z}\| + \|A - \bar{A}\|$$

with the operator norm on $\mathbb{R}^{m\times n}$ *chosen to be compatible with the given norms on* $\mathbb{R}^n$ *and* $\mathbb{R}^m$. *In particular, the parameters* μ *in* (iii) *and* β *in* (iv) *depend only on the point* $(\bar{z}, \bar{A})$.

Proof. To obtain the equivalence of (i) and (ii), we first take the polar of the relation in (i) to see that $\operatorname{im}\bar{A} + \operatorname{cl}\operatorname{cone}(C - \bar{z}) = \mathbb{R}^m$. From this equation, the equivalence follows from a simple separation argument and the fact that $\operatorname{ri}\operatorname{cone} S = \operatorname{cone}(\operatorname{ri} S)$ for any convex set S.

Clearly, (ii) follows from (iii). The reverse implication again follows by a simple separation argument. Indeed, if this implication were false, then one could separate the origin from the set $\mu\bar{A}\mathbb{B}_n + (\operatorname{ri} C - \bar{z})$, for each $\mu > 0$. But then the cone generated by these sets, namely $\operatorname{im}\bar{A} + \operatorname{cone}(\operatorname{ri} C - \bar{z})$, would lie in a half space which would contradict (ii).

To see that (iv) follows from (iii), note that (iii) is equivalent to the statement that there exists an $\eta > 0$ such that $\eta\mathbb{B}_m \subset \mu\bar{A}\mathbb{B}_n + (\operatorname{ri} C - \bar{z})$. This implies that $(\eta/\mu)\mathbb{B}_m \subset \bar{A}\mathbb{B}_n + \operatorname{cone}(C - \bar{z})$. The polar of this last expression is precisely (iv) with $\beta = \mu\eta^{-1}$.

Clearly, (iv) implies (i) since $\ker\bar{A} = (\bar{A}^{\mathrm{T}})^{-1}0$ and the only bounded cone is the origin.

For the final statement of the lemma, it is obvious that (v) implies any one of (i)–(iv). We obtain the equivalence of (v) with any one of (i)–(iv) by showing that (iii) implies the local version of (iv). This will simultaneously establish the uniform nature of the parameter β. First, it is clear that if (iii) holds for some $\bar{A}$, $\bar{z}$ and $\bar{\mu}$, then it holds for all A, z and μ nearby. As noted above, the condition in (iii) implies the existence of an $\eta > 0$ and $\mu > 0$ such that $\eta\mathbb{B}_m \subset \mu\bar{A}\mathbb{B}_n + (\operatorname{ri} C - \bar{z})$. Hence, $\eta\mathbb{B}_m \subset \mu A\mathbb{B}_n + (\operatorname{ri} C - z) + \frac{1}{2}\eta\mathbb{B}_m$, whenever $(z, A) \in (\bar{z}, \bar{A}) + \epsilon\mathbb{B}$, for some $\epsilon > 0$. Therefore, by the Rådström Cancellation Lemma [32, Lemma 1], $\frac{1}{2}\eta\mathbb{B}_m \subset \mu A\mathbb{B}_n + (\operatorname{ri} C - z)$ which implies that $\eta/(2\mu)\mathbb{B}_m \subset A\mathbb{B}_n + \operatorname{cone}(C - z)$. Taking the polar of this last statement and setting $\beta^{-1} := \eta/(2\mu)$, we find that (iii) implies the existence of $\epsilon > 0$ and $\beta > 0$ such that the condition in (iv) holds for all $(z, A) \in (\bar{z}, \bar{A}) + \epsilon\mathbb{B}$. □

Remarks. (1) Maguregui [22,23] studies methods similar to Algorithm 1 in the Banach space setting under the regularity condition

$$0 \in \operatorname{core}((\operatorname{im} \bar{A} + C - \bar{z})). \tag{10}$$

This condition and condition (ii) of Lemma 3.2 are equivalent in the finite-dimensional setting. In the infinite-dimensional setting, (10) is stronger than the condition in (ii), but condition (ii) is not strong enough to obtain Maguregui's sensitivity results.

(2) By taking $A = F'(\bar{x})$ and $\bar{z} = F(\bar{x})$, Lemma 3.2(v) implies that $\ker(F'(x)^{\mathrm{T}}) \cap \Gamma_C(F(x)) = \{0\}$ for all points x near $\bar{x}$ at which (7) holds. That is, regularity is a local property.

The following proposition can be viewed as a local Hoffman bound for the linearized inclusions (6). This result is similar to [35, Theorem 1]. Our proof is a straightforward application of Fenchel's Duality Theorem [38, Corollary 31.2.1]. It is self-contained and significantly simplifies the proof technique employed in [35, Theorem 1] which depends on the theory of *normed convex processes* [34]. The power of this approach for the derivation of very general Hoffman bounds is illustrated further in [10].

Proposition 3.3. *If $\bar{x}$ is a regular point of* (1), *then for all $\Delta > \operatorname{dist}(0 \mid D_\infty(\bar{x}))$, there is some neighborhood $\mathcal{N}(\bar{x})$ of $\bar{x}$ and a $\beta > 0$ satisfying*

$$\operatorname{dist}(0 \mid D_\Delta(x)) \leqslant \beta \operatorname{dist}(F(x) \mid C), \tag{11}$$

whenever $x \in \mathcal{N}(\bar{x})$. Moreover, $\mathcal{N}(\bar{x})$ can be chosen so that there exists $d \in \Delta\mathbb{B}$ satisfying

$$F(x) + F'(x)d \in \operatorname{ri} C, \tag{12}$$

for all $x \in \mathcal{N}(\bar{x})$.

Proof. We first establish (11). Let $\epsilon > 0$ be given by Lemma 3.2(v) at the pair $(F(\bar{x}), F'(\bar{x}))$. Let $\mathcal{N}(\bar{x})$ be the neighborhood of $\bar{x}$ chosen so that $(F(x), F'(x)) \in (F(\bar{x}), F'(\bar{x})) + \epsilon\mathbb{B}$ whenever $x \in \mathcal{N}(\bar{x})$.

By Lemma 3.2(ii) and the continuity of F and F', the set $\{d \mid F(x) + F'(x)d \in C\}$ is nonempty and equals $D_\infty(x)$ for all $x \in \mathcal{N}(\bar{x})$. The relation (11) follows from the inequality

$$\operatorname{dist}(0 \mid D_\infty(x)) \leqslant \beta \operatorname{dist}(F(x) \mid C), \tag{13}$$

which we now establish via Fenchel duality.

The Fenchel dual to the problem

$$\begin{aligned}\operatorname{dist}(0 \mid D_\infty(x)) &= \inf\{\|d\| \mid F(x) + F'(x)d \in C\}\\ &= \inf\{\psi^*_{\mathbb{B}^\circ}(d) + \psi_{C-F(x)}(F'(x)d)\}\end{aligned}$$

is the problem

$$\sup\{\langle y, F(x)\rangle - \psi_C^*(y) \mid F'(x)^{\mathrm{T}} y \in \mathbb{B}^\circ\}. \tag{14}$$

The optimal values of these problems coincide with attainment in (14) if there exists $d \in \mathbb{R}^n$ satisfying (12). For each $x \in \mathcal{N}(\bar{x})$, such a d exists by parts (iii) and (v) of Lemma 3.2. Since $0 \leqslant \operatorname{dist}(0 \mid D_\infty(x))$, identity (8) can be used to further restrict the constraint region in (14) by adding the inclusion $y \in \Gamma_C(F(x))$. This observation along with parts (iv) and (v) of Lemma 3.2 yields the relation

$$\begin{aligned}\operatorname{dist}(0 \mid D_\infty(x)) &= \max\{\langle y, F(x)\rangle - \psi_C^*(y) \mid y \in (F'(x)^{\mathrm{T}})^{-1}\mathbb{B}^\circ \cap \Gamma_C(F(x))\}\\ &\leqslant \max\{\langle y, F(x)\rangle - \psi_C^*(y) \mid y \in \beta\mathbb{B}^\circ\}\\ &\leqslant \beta \operatorname{dist}(F(x) \mid C),\end{aligned}$$

for all $x \in \mathcal{N}(\bar{x})$. The last inequality follows from the fact that for every $z \in C$ and $y \in \beta\mathbb{B}^\circ$ one has

$$\langle y, F(x)\rangle \leqslant \langle y, F(x) - z\rangle + \langle y, z\rangle \leqslant \beta\|F(x) - z\| + \psi_C^*(y),$$

and so $\langle y, F(x)\rangle - \psi_C^*(y) \leqslant \beta \operatorname{dist}(F(x) \mid C)$.

We now construct d with $\|d\| \leqslant \Delta$ satisfying (12) to complete the proof. Let $\Delta > \operatorname{dist}(0 \mid D_\infty(x))$ be given and let $\Delta_0 = \frac{1}{2}\Delta + \frac{1}{2}\operatorname{dist}(0 \mid D_\infty(x))$. By (11), there is a neighborhood of $\bar{x}$ on which $D_{\Delta_0}(x) = \{d \in \Delta_0\mathbb{B} \mid F(x) + F'(x)d \in C\} \neq \emptyset$. Let $d_1 \in D_{\Delta_0}$. By parts (iii) and (v) of Lemma 3.2, there is a $d_2 \in \mathbb{R}^n$ satisfying (12). By [38, Theorem 6.1], it follows that

$$(1-t)[F(x) + F'(x)d_1] + t[F(x) + F'(x)d_2] \in \operatorname{ri} C, \quad \forall t \in (0,1],$$

and hence that $F(x) + F'(x)((1-t)d_1 + td_2) \in \operatorname{ri} C$. The required d is determined by choosing $t > 0$ small enough so that $(1-t)d_1 + td_2 \in \Delta\mathbb{B}$. □

Remark. The first part of this result can be extended to the setting of reflexive Banach spaces under Maguregui's regularity hypothesis (10) or if one assumes a regularity condition having the form of parts (ii) or (iii) of Lemma 3.2.

We now take a moment to clarify the relationship of our result to that of Womersley [40]. For this purpose, recall that Robinson [37, Theorem 1] extends the stability result (11) to smooth nonlinear systems under the same regularity conditions. His result implies that for some $\eta > 0$,

$$\operatorname{dist}(x \mid F^{-1}(C)) \leqslant \eta \operatorname{dist}(F(x) \mid C),$$

for all x in a neighborhood of the regular point $\bar{x}$. It immediately follows that regularity coupled with weak sharpness implies that the composite function $h \circ F$ is locally weak sharp with respect to the set $F^{-1}(C)$. That is,

$$f(x) = h(F(x)) \geqslant f(\bar{x}) + \gamma \operatorname{dist}(x \mid F^{-1}(C)), \tag{15}$$

for all x near $\bar{x}$. Womersley [40] assumes the relation (15) with $F^{-1}(C) = \{\bar{x}\}$, which is the key difference to our approach. Our proof theory is based on sufficient conditions to ensure weak sharpness, whereas Womersley assumes only sharpness and a unique solution. The proof we give does not assume that the set of minimizers of f is even bounded, let alone a singleton.

4. Quadratic convergence

We can now establish the local quadratic convergence of Algorithm 1. This result is reminiscent of the Kantorovich Theorem for the convergence of Newton's method since the existence of a solution to (1) is not assumed. The result also extends the quadratic convergence result due to Womersley [40] to the case of nonunique solution sets (for related results, see [19,23,29]).

Theorem 4.1. *Let $\bar{x} \in \mathbb{R}^n$ be a regular point of the inclusion* (1) *where C is a set of weak sharp minima for h and suppose that the conclusions of Proposition* 3.3 *are satisfied on the set $\bar{x}+\bar{\delta}\mathbb{B}$ for $\bar{\delta} > 0$, with $\Delta > \bar{\delta}$. Assume that F' is Lipschitz continuous on $\bar{x}+\bar{\delta}\mathbb{B}$ with Lipschitz constant L and h is Lipschitz continuous on $F(\bar{x}+\bar{\delta}\mathbb{B})+\frac{1}{8}L\mathbb{B}$ with Lipschitz constant M. If there exists $\delta > 0$ such that*

(a) $\delta < \min\{\frac{1}{2}\bar{\delta}, 1\}$,

(b) $\operatorname{dist}(F(\bar{x}) \mid C) < \delta/2\eta\beta$, *and*

(c) $\theta := \eta L M \delta \beta/\alpha < 1$,

then there is a neighborhood $\mathcal{M}(\bar{x})$ of $\bar{x}$ such that if Algorithm 1 *is initiated in $\mathcal{M}(\bar{x})$, then the iterates $\{x^k\}$ converge to some $x^* \in \mathbb{R}^n$ with $F(x^*) \in C$; that is, x^* solves $(\mathcal{P})$. Furthermore, $x^k \to x^*$ and $h(F(x^k)) \to h_{\min}$ at a quadratic rate.*

Proof. Since F and F' are continuous, there is a neighborhood $\mathcal{O}(\bar{x})$ such that

$$\operatorname{dist}(F(x) \mid C) \leqslant \frac{\delta}{2\eta\beta}, \tag{16}$$

and $D_\Delta(x) = \{d \in \Delta\mathbb{B} \mid F(x) + F'(x)d \in C\} \neq \emptyset$, for all $x \in \mathcal{O}(\bar{x})$. Let $\mathcal{M}(\bar{x}) := \mathcal{O}(\bar{x}) \cap (\bar{x}+\delta\mathbb{B})$. If Algorithm 1 is applied, observe that

$$\|d^0\| \leqslant \eta\beta \operatorname{dist}(F(x^0) \mid C) \leqslant \tfrac{1}{2}\delta. \tag{17}$$

Since $x^0 \in \mathcal{M}(\bar{x})$, we have

$$\|x^1 - \bar{x}\| \leqslant \|x^0 - \bar{x}\| + \|d^0\| \leqslant \delta + \tfrac{1}{2}\delta < 2\delta,$$

and $x^1 \in \bar{x} + 2\delta\mathbb{B} \subset \bar{x} + \bar{\delta}\mathbb{B}$. We claim that for $k = 1, 2, \ldots,$

$$x^k \in \bar{x} + \bar{\delta}\mathbb{B} \tag{18}$$

and

$$\|d^k\| \leqslant \eta\beta\alpha^{-1}(h(F(x^k)) - h_{\min}) \leqslant \tfrac{1}{2}\theta\|d^{k-1}\|^2 \leqslant \tfrac{1}{2}\delta 2^{[-2^k]}. \tag{19}$$

Note that (18) implies that the algorithm is well-defined for $k = 0, 1, 2, \ldots$. The proof of the claim proceeds by induction on k.

Observe that (18) has already been established for $k = 1$. To see (19) for $k = 1$, first recall that by the quadratic bound lemma [28, 3.2.12]

$$\|F(x^0) + F'(x^0)d^0 - F(x^1)\| \leqslant \tfrac{1}{2}L\|d^0\|^2 \leqslant \tfrac{1}{8}L,$$

so that $F(x^0) + F'(x^0)d^0 \in F(\bar{x} + \bar{\delta}\mathbb{B}) + \frac{1}{8}L\mathbb{B}$. Thus,

$$\begin{aligned} h(F(x^1)) &= h\left(F(x^0) + F'(x^0)d^0 + \int_0^1 (F'(x^0 + td^0) - F'(x^0))d^0\,\mathrm{d}t\right) \\ &\leqslant h(F(x^0) + F'(x^0)d^0) + M\left\|\int_0^1 (F'(x^0 + td^0) - F'(x^0))d^0\,\mathrm{d}t\right\| \\ &= h_{\min} + M\left\|\int_0^1 (F'(x^0 + td^0) - F'(x^0))d^0\,\mathrm{d}t\right\| \\ &\leqslant h_{\min} + \tfrac{1}{2}LM\|d^0\|^2, \end{aligned} \tag{20}$$

the second equality following from (12). We now have

$$\begin{aligned} \|d^1\| &\leqslant \eta\beta\alpha^{-1}(h(F(x^1)) - h_{\min}) && \text{(from (4), (11) and (5))} \\ &\leqslant \tfrac{1}{2}\theta\|d^0\|^2 && \text{(from (20))} \\ &\leqslant \tfrac{1}{2}\theta(\tfrac{1}{2}\delta)^2 && \text{(from (17))} \\ &\leqslant \tfrac{1}{2}\delta 2^{-2}. \end{aligned}$$

Next assume that (18) and (19) hold for $k = 1, \ldots, s$. We show that they also hold for $k = s + 1$. First of all, since $x^0 \in \mathcal{M}(\bar{x})$, we have

$$\|x^{s+1} - \bar{x}\| \leqslant \|x^0 - \bar{x}\| + \sum_{k=0}^{s} \|d^k\| \leqslant \delta + \tfrac{1}{2}\delta\sum_{k=0}^{s} 2^{[-2^k]} \leqslant \delta + \tfrac{1}{2}\delta\sum_{k=0}^{s} 2^{-2k} < 2\delta,$$

so that $x^{s+1} \in \bar{x} + \bar{\delta}\mathbb{B}$. Therefore, as in (20), $h(F(x^{s+1})) \leqslant h_{\min} + \frac{1}{2}LM\|d^s\|^2$, and so by the induction hypothesis we obtain

$$\|d^{s+1}\| \leqslant \eta\beta\alpha^{-1}(h(F(x^{s+1})) - h_{\min}) \leqslant \tfrac{1}{2}\theta\|d^s\|^2 \leqslant \tfrac{1}{2}\theta(\tfrac{1}{2}\delta 2^{-2^s})^2 \leqslant \tfrac{1}{2}\delta 2^{[-2^{s+1}]},$$

which concludes our induction.

Therefore, the sequence is Cauchy, and so must converge to some x^* satisfying $h(F(x^*)) = h_{\min}$. To prove the quadratic rate of convergence for $\{x^k\}$, we note from the triangle inequality that

$$\|x^{k+1} - x^k\| - \|x^{k+i} - x^{k+1}\| \leqslant \|x^{k+i} - x^k\|,$$

$$\|x^{k+i} - x^{k+1}\| \leqslant \sum_{j=1}^{i-1} \|x^{k+j+1} - x^{k+j}\| \leqslant \sum_{j=1}^{\infty} \|x^{k+j+1} - x^{k+j}\|.$$

For large k, $\|x^{k+1}-x^k\| = \epsilon < 1/(3\theta)$. It follows from (19) that $\|x^{k+2}-x^{k+1}\| \leqslant \theta\epsilon^2$, and in general that $\|x^{k+j+1}-x^{k+j}\| \leqslant \theta^{2^j-1}\epsilon^{2^j} \leqslant (\theta\epsilon)^j\epsilon$. Thus, $\|x^{k+i}-x^{k+1}\| \leqslant \theta\epsilon^2/(1-\theta\epsilon)$, and hence $\|x^{k+i}-x^k\|^2 \geqslant \epsilon^2((1-2\theta\epsilon)/(1-\theta\epsilon))^2$. From these estimates,

$$\frac{\|x^{k+1}-x^{k+i}\|}{\|x^k-x^{k+i}\|^2} \leqslant \frac{\theta\epsilon^2/(1-\theta\epsilon)}{\epsilon^2((1-2\theta\epsilon)/(1-\theta\epsilon))^2} < 6\theta,$$

and the quadratic rate for $\{x^k\}$ follows. The quadratic rate of convergence for $\{h(F(x^k))\}$ is obvious from (19). □

Observe that the above result can also be viewed as a domain of attraction result by assuming that point $\bar{x}$ referred to in the hypotheses actually solves the inclusion (1). In this case the inequality in assumption (b) is trivially satisfied so that a $\delta > 0$ satisfying assumptions (a)–(c) is guaranteed to exist.

5. A globalization strategy

In this section, we propose a globalization strategy for Algorithm 1, based on a backtracking line-search. The algorithm is simply stated as follows.

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

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

(2) Set $x^{k+1} := x^k + t_k d^k$ where t_k is the maximum value of γ^s, for $s = 0, 1, \ldots$, such that

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

Algorithm 2 is an instance of the class of algorithms studied in [6], so the global convergence properties of the method follow from [6, Theorems 2.4 and 5.3]. These theorems specify the behavior of sequences generated by Algorithm 2 in terms of the first-order optimality conditions for the problem $(\mathcal{P})$. Recall that a point $\bar{x}$ is a first-order stationary point for the problem $(\mathcal{P})$ if

$$f'(\bar{x}; d) \geqslant 0, \quad \text{for all } d \in \mathbb{R}^n, \tag{21}$$

where $f'(\bar{x}; \cdot)$ (the usual directional derivative of f at the point $\bar{x}$) exists and is finite-valued on $\mathbb{R}^n$. By [6, Lemma 4.5 and Theorem 3.6], condition (21) is equivalent to the condition

$$h(F(\bar{x}) + F'(\bar{x})d) - h(F(\bar{x})) = 0, \quad \text{for all } d \in D_\Delta(\bar{x}), \tag{22}$$

which can be more simply stated as

$$0 \in D_\Delta(\bar{x}). \tag{23}$$

Conditions (22) and (23) are particularly important in light of the search direction and step-length choice specified in Algorithm 2.

The key global properties for Algorithm 2, established in [6], are stated as the following theorem.

Theorem 5.1. *Let $x^0 \in \mathbb{R}^n$ and let $f = h \circ F$ be as in $(\mathcal{P})$. Suppose that F' is uniformly continuous on the closed convex hull of the set $\{x \in \mathbb{R}^n \colon f(x) \leqslant f(x^0)\}$ and that h is Lipschitz on the set $\{y \in \mathbb{R}^m \colon h(y) \leqslant f(x^0)\}$. If $\{x^k\}$ is the sequence generated by Algorithm 2 with initial point x^0, then at least one of the following must occur.*

(i) *The iterates terminate finitely at a first-order stationary point for the problem $(\mathcal{P})$.*

(ii) *The sequence of values $\{f(x^k)\}$ decreases to $-\infty$.*

(iii) *The sequence $\{\|d^k\|\}$ diverges to $+\infty$.*

(iv) *For every subsequence $K \subset \{1,2,\ldots\}$ for which the search directions $\{d^k\}_K$ remain bounded, one has*

$$\lim_{k\in K}[h(F(x^k) + F'(x^k)d^k) - h(F(x^k))] = 0.$$

Moreover, every cluster point of the subsequence $\{x^k\}_K$ is a first-order stationary point for $(\mathcal{P})$.

An immediate consequence of the above result is that if the set $C = \arg\min h$ is nonempty and $\Delta < +\infty$, then

$$\lim_{k}[h(F(x^k) + F'(x^k)d^k) - h(F(x^k))] = 0,$$

and every cluster point of the sequence $\{x^k\}$ is a first-order stationary point for $(\mathcal{P})$. We now further analyze the convergence behavior at cluster points satisfying the sharpness and regularity hypotheses.

Theorem 5.2. *Let $f := h \circ F$ be as in $(\mathcal{P})$ with h finite-valued and F' locally Lipschitz continuous. Suppose that $\{x^k\}$ is a sequence generated by Algorithm 2 and that $\bar{x}$ is a cluster point of this sequence. If $\bar{x}$ is a regular point of the inclusion (1) where C is a set of weak sharp minima for h, then $F(\bar{x}) \in C$ and both $x^k \to \bar{x}$ and $f(x^k) \to h_{\min}$ at a quadratic rate.*

Proof. We first show that $F(\bar{x}) \in C$. By Theorem 5.1, we have $0 \in F'(\bar{x})^{\mathrm{T}}\partial h(F(\bar{x}))$. If $F(\bar{x}) \notin C$, then $0 \notin \partial h(F(\bar{x}))$ in which case there is a nonzero $z \in \partial h(F(\bar{x})) \cap \ker F'(\bar{x})^{\mathrm{T}}$. But $z \in \Gamma_C(F(\bar{x}))$ since $0 \geqslant h(y) - h(F(\bar{x})) \geqslant \langle z, y - F(\bar{x})\rangle$, for all $y \in C$, so that

$$z \in \{v \colon \langle v, F(\bar{x})\rangle - \psi_C(v) \geqslant 0\} = \Gamma_C(F(\bar{x})) \quad \text{(by (8))}.$$

This contradicts the assumption that $\bar{x}$ is a regular point of the inclusion (1), hence $F(\bar{x}) \in C$.

We now establish the convergence rates. Since the regularity hypothesis at $\bar{x}$ implies that $F(\bar{x}) \in C$, the result will follow immediately from Theorem 4.1 if it can be shown that there is a neighborhood N of $\bar{x}$ such that

$$h(F(x+d)) - h(F(x)) \leqslant c[h(F(x) + F'(x)d) - h(F(x))], \tag{24}$$

for all $x \in N$ and $d \in \mathbb{R}^n$ satisfying $\|d\| \leqslant \eta\,\mathrm{dist}(0 \mid D_\Delta(x))$. Indeed, if (24) holds, then Algorithms 1 and 2 generate identical iterates sufficiently close to $\bar{x}$. Hence, by Theorem 4.1, these iterates remain close to $\bar{x}$ and converge to a solution of $(\mathcal{P})$. Since $\bar{x}$ is a cluster point of this sequence, the entire sequence must converge to $\bar{x}$ with $x^k \to \bar{x}$ and $f(x^k) \to h_{\min}$ quadratically.

Suppose to the contrary that (24) does not hold near $\bar{x}$. Then there is a sequence $\{\bar{x}^k\}$ converging to $\bar{x}$ such that

$$c[h(F(\bar{x}^k) + F'(\bar{x}^k)\bar{d}^k) - h(F(\bar{x}^k))] < h(F(\bar{x}^k + \bar{d}^k)) - h(F(\bar{x}^k)) \tag{25}$$

at each $\bar{x}^k$ for some $\bar{d}^k \in \mathbb{R}^n$ satisfying

$$\|\bar{d}^k\| \leqslant \eta\,\mathrm{dist}(0 \mid D_\Delta(\bar{x}^k)). \tag{26}$$

In particular, we obtain from (11) that

$$\|\bar{d}^k\| \to 0. \tag{27}$$

Let N_1 be a compact neighborhood of $\bar{x}$ containing the set $\bar{x} + 2\Delta\mathbb{B}$ and let K and M be Lipschitz constants for h on $F(N_1)$ and F' on N_1, respectively. Let $\Delta > \delta > 0$ be chosen so that the conclusions of Proposition 3.3 hold for this choice of δ. We suppose with no loss of generality that $\{\bar{x}^k\} \subset \bar{x} + \delta\mathbb{B}$. Then for all k we have from [28, 3.2.12] that

$$\begin{aligned} h(F(\bar{x}^k + \bar{d}^k)) - h_{\min} &= \|h(F(\bar{x}^k + \bar{d}^k)) - h(F(\bar{x}^k) + F'(\bar{x}^k)\bar{d}^k)\| \\ &\leqslant K\|F(\bar{x}^k + \bar{d}^k) - F(\bar{x}^k) - F'(\bar{x}^k)\bar{d}^k\| \\ &\leqslant \tfrac{1}{2}KM\|\bar{d}^k\|^2. \end{aligned}$$

Therefore, by (25),

$$\begin{aligned} c[h_{\min} - h(F(\bar{x}^k))] &= c[h(F(\bar{x}^k) + F'(\bar{x}^k)\bar{d}^k) - h(F(\bar{x}^k))] \\ &< h(F(\bar{x}^k + \bar{d}^k)) - h(F(\bar{x}^k)) \\ &\leqslant h_{\min} - h(F(\bar{x}^k)) + \tfrac{1}{2}KM\|\bar{d}^k\|^2. \end{aligned}$$

Consequently,

$$\begin{aligned} 0 &< (1-c)[h_{\min} - h(F(\bar{x}^k))] + \tfrac{1}{2}KM\|\bar{d}^k\|^2 \\ &\leqslant (c-1)\alpha\,\mathrm{dist}(F(\bar{x}^k) \mid C) + \tfrac{1}{2}KM\|\bar{d}^k\|^2 \qquad \text{(from (5))} \\ &\leqslant (c-1)\alpha(\beta\eta)^{-1}\|\bar{d}^k\| + \tfrac{1}{2}KM\|\bar{d}^k\|^2 \qquad \text{(from (11) and (26))}. \end{aligned}$$

After dividing this expression through by $\|\bar{d}^k\|$ and using (27) while taking the limit in k, we obtain the contradiction $0 \leqslant (c-1)\alpha(\beta\eta)^{-1}$. □

6. Concluding remarks

We have shown that under the assumptions of weak sharpness and regularity one obtains the local quadratic convergence of a Gauss–Newton method for convex composite optimization. Moreover, the method can be "globalized" with the addition of a backtracking line-search that does not inhibit the local rate. A similar result can be established for a trust-region-based globalization strategy.

Let us briefly consider the implications for the case of the nonlinear least squares (NLLS). Consider Algorithm 1 under the assumptions of weak sharpness and regularity. Recall that the regularity condition implies that $\operatorname{im}(F'(x)) = \mathbb{R}^m$ so that $n \geqslant m$. This instance of the NLLS problem seems to have received little study [13,15,28]. Nonetheless, it is of great significance in the description of constraint regions in nonlinear programming. In this case, condition (4) with $\eta = 1$ in Algorithm 1 is easily satisfied with the aid of either a QR factorization or a singular-value decomposition. In either case, the algorithm is locally equivalent to the Ben-Israel iteration for NLLS [1,3]. In their analyses of this iteration, neither Ben-Israel [1] or Boggs [3] establish a rate of convergence for the method nor do they provide a globalization strategy. The results of this article can be applied to fill in these gaps.

In the context of solving the more general problem (1), recall that near a regular solution to this inclusion, the direction-finding subproblem in Algorithm 1 corresponds to locating a least-norm solution to the linearized problem $F(x^k) + F'(x^k)d = 0$. Thus, the method is locally equivalent to the procedure described in [23]. It is also equivalent to the procedure given in [33] if C is a cone. Robinson [33] (if C is a cone) and Maguregui [23] obtain a Kantorovich-type convergence result. Theorem 4.1 is similar to these results since it can also be viewed as an existence result for the inclusion (1). Robinson's approach appeals to the theory of normed convex processes [34], while Maguregui makes use of the Robinson–Ursescu Theorem [36] along with the perturbation theory in [22]. The proof theory provided in this article is more elementary and accessible since it only requires a straightforward application of Fenchel duality. Moreover, we improve the applicability of the theory by providing a simple globalization strategy that preserves the local rate.

References

[1] A. Ben-Israel, "A Newton–Raphson method for the solution of systems of equations," *Journal of Mathematical Analysis and its Applications* 15 (1966) 243–252.

[2] S.C. Billups and M.C. Ferris, "Solutions to affine generalized equations using proximal mappings," Mathematical Programming Technical Report 94-15 (Madison, WI, 1994).

[3] P.T. Boggs, "The convergence of the Ben-Israel iteration for nonlinear least squares problems," *Mathematics of Computation* 30 (1976) 512–522.

[4] J.M. Borwein, "Stability and regular points of inequality systems," *Journal of Optimization Theory and Applications* 48 (1986) 9–52.

[5] J.V. Burke, "Algorithms for solving finite dimensional systems of nonlinear equations and inequalities that have both global and quadratic convergence properties," Report ANL/MCS-TM-54, Mathematics and Computer Science Division, Argonne National Laboratory (Argonne, IL, 1985).

[6] J.V. Burke, "Descent methods for composite nondifferentiable optimization problems," *Mathematical Programming* 33 (3) (1985) 260–279.

[7] J.V. Burke, "An exact penalization viewpoint of constrained optimization," *SIAM Journal on Control and Optimization* 29 (1991) 968–998.

[8] J.V. Burke and M.C. Ferris, "Weak sharp minima in mathematical programming," *SIAM Journal on Control and Optimization* 31 (1993) 1340–1359.

[9] J.V. Burke and R.A. Poliquin, "Optimality conditions for non-finite valued convex composite functions," *Mathematical Programming* 57 (1) (1992) 103–120.

[10] J.V. Burke and P. Tseng, "A unified analysis of Hoffman's bound via Fenchel duality," *SIAM Journal on Optimization*, to appear.

[11] L. Cromme, "Strong uniqueness. A far reaching criterion for the convergence analysis of iterative procedures," *Numerische Mathematik* 29 (1978) 179–193.

[12] R. De Leone and O.L. Mangasarian, "Serial and parallel solution of large scale linear programs by augmented Lagrangian successive overrelaxation," in: A. Kurzhanski et al., eds., *Optimization, Parallel Processing and Applications*, Lecture Notes in Economics and Mathematical Systems, Vol. 304 (Springer, Berlin, 1988) pp. 103–124.

[13] J.E. Dennis and R.B. Schnabel, *Numerical Methods for Unconstrained Optimizations and Nonlinear Equations* (Prentice-Hall, Englewood Cliffs, NJ, 1983).

[14] M.C. Ferris, "Weak sharp minima and penalty functions in mathematical programming," Ph.D. Thesis, University of Cambridge (Cambridge, 1988).

[15] R. Fletcher, "Generalized inverse methods for the best least squares solution of non-linear equations," *The Computer Journal* 10 (1968) 392–399.

[16] R. Fletcher, "Second order correction for nondifferentiable optimization," in: G.A. Watson, ed., *Numerical Analysis*, Lecture Notes in Mathematics, Vol. 912 (Springer, Berlin, 1982) pp. 85–114.

[17] R. Fletcher, *Practical Methods of Optimization* (Wiley, New York, 2nd ed., 1987).

[18] U.M. Garcia-Palomares and A. Restuccia, "A global quadratic algorithm for solving a system of mixed equalities and inequalities," *Mathematical Programming* 21 (3) (1981) 290–300.

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

[20] K. Levenberg, "A method for the solution of certain nonlinear problems in least squares," *Quarterly Applied Mathematics* 2 (1944) 164–168.

[21] K. Madsen, "Minimization of nonlinear approximation functions," Ph.D. Thesis, Institute of Numerical Analysis, Technical University of Denmark (Lyngby, 1985).

[22] J. Maguregui, "Regular multivalued functions and algorithmic applications," Ph.D. Thesis, University of Wisconsin (Madison, WI, 1977).

[23] J. Maguregui, "A modified Newton algorithm for functions over convex sets," in: O.L. Mangasarian, R.R. Meyer and S.M. Robinson, eds., *Nonlinear Programming 3* (Academic Press, New York, 1978) pp. 461–473.

[24] O.L. Mangasarian, "Least-norm linear programming solution as an unconstrained minimization problem," *Journal of Mathematical Analysis and Applications* 92 (1) (1983) 240–251.

[25] O.L. Mangasarian, "Normal solutions of linear programs," *Mathematical Programming Study* 22 (1984) 206–216.

[26] O.L. Mangasarian and S. Fromovitz, "The Fritz John necessary optimality conditions in the presence of equality and inequality constraints," *Journal of Mathematical Analysis and its Applications* 17 (1967) 37–47.

[27] D.W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal of Applied Mathematics* 11 (1963) 431–441.

[28] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables* (Academic Press, New York, 1970).

[29] M.R. Osborne and R.S. Womersley, "Strong uniqueness in sequential linear programming," *Journal of the Australian Mathematical Society. Series B* 31 (1990) 379–384.

[30] B.T. Polyak, *Introduction to Optimization* (Optimization Software, New York, 1987).

[31] M.J.D. Powell, "General algorithm for discrete nonlinear approximation calculations," in: C.K. Chui, L.L. Schumaker and J.D. Ward, eds., *Approximation Theory IV* (Academic Press, New York, 1983) pp. 187–218.

[32] H. Rådström, "An embedding theorem for spaces of convex sets," *Proceedings of the American Mathematical Society* 3 (1952) 165–169.

[33] S.M. Robinson, "Extension of Newton's method to nonlinear functions with values in a cone," *Numerische Mathematik* 19 (1972) 341–347.

[34] S.M. Robinson, "Normed convex processes," *Transactions of the American Mathematical Society* 174 (1972) 127–140.

[35] S. Robinson, "Stability theory for systems of inequalities, Part I: linear systems," *SIAM Journal on Numerical Analysis* 12 (1975) 754–769.

[36] S. Robinson, "Regularity and stability for convex multivalued functions," *Mathematics of Operations Research* 1 (1976) 130–143.

[37] S. Robinson, "Stability theory for systems of inequalities, Part II: Differentiable nonlinear systems," *SIAM Journal on Numerical Analysis* 13 (1976) 497–513.

[38] R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, NJ, 1970).

[39] R.T. Rockafellar, "First- and second-order epi-differentiability in nonlinear programming," *Transactions of the American Mathematical Society* 307 (1988) 75–108.

[40] R.S. Womersley, "Local properties of algorithms for minimizing nonsmooth composite functions," *Mathematical Programming* 32 (1) (1985) 69–89.