CENTER FOR
PARALLEL OPTIMIZATION

A GAUSS-NEWTON METHOD FOR CONVEX COMPOSITE OPTIMIZATION

by

J. V. Burke & M. C. Ferris

Computer Sciences Technical Report #1176

August 1993

A Gauss–Newton Method for Convex Composite Optimization[*]

J.V. Burke[†] M.C. Ferris[‡]

August 1993

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 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.

[*] This material is based on research supported by National Science Foundation Grants CCR-9157632 and DMS-9102059 and Air Force Office of Scientific Research Grant AFOSR-89-0410

[†] Department of Mathematics, GN–50, University of Washington, Seattle, Washington 98195

[‡] Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706

176

1 Introduction

In the early nineteenth century, Gauss proposed a powerful method for solving systems of nonlinear equations which generalized the classical Newton's method for such systems. The so called *Gauss–Newton* method is easily described. Suppose one wishes to solve the system

$$F(x) = 0 \ , \tag{1}$$

where $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is of class C^1. Newton's method generates iterates of the form

$$x^{k+1} := x^k + s^k \, , \tag{2}$$

where the step s^k is a solution to the linear system

$$0 = F(x^k) + F'(x^k)s \, . \tag{3}$$

Unfortunately, the system (3) may be inconsistent, especially if the system is overdetermined ($m > n$). In order to remedy this problem, Gauss proposed taking s^k as the best approximate solution to (3) in the least–squares sense. In making the transition to a step s^k based on a least–squares solution to (3), the underlying problem has been changed from equation solving to minimization. Specifically, the algorithm is now designed to solve the minimization problem

$$\min \frac{1}{2} \|F(x)\|_2^2 \ . \tag{4}$$

In this context, it is clear that the Gauss–Newton approach can generate iterates that converge to a solution to (4) that is not a solution to (1). 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 the objective in the step finding subproblem (see [20, 28]) or the inclusion of a trust–region constraint (see [11]).

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:

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

where $h\colon \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$ is convex and $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is of class C^1. In this article, we consider only the finite–valued case; $h\colon \mathbb{R}^m \to \mathbb{R}$. Obviously the problem (4) is precisely of this form. It is interesting to note that in their outline of the Gauss–Newton method, Ortega and Rheinboldt [29, page 267] used the notion of a composite function. A wide variety of applications of this formulation can be found throughout the mathematical programming literature [3, 14, 15, 22, 35, 42, 44, 45], 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 [5, 7, 15, 18, 42]. Indeed, the deepest results on optimality conditions for a variety of problems have been obtained in this way.

Our extension of the Gauss–Newton methodology to finite–valued convex composite optimization generalizes a similar result due to Womersley [44] which uses the assumption of strong uniqueness on the composite function f. An important distinction between these results is that we do not require that the solution set be a singleton. In particular, the solution set may be unbounded. The approach we take is based on that described in [2] which is an extension of a technique due to Garcia–Palomares and Restuccia [17]. However, since [2, 17] are concerned with the more specific problem of solving nonlinear systems of equations and inequalities, we are not able to capture the full flavor of these results. In particular, [2, 17] contain results concerning active set strategies that are not considered in this article.

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

$$F(x) \in C. \tag{5}$$

The notion of weak sharp minima was introduced in [6, 12] and is reviewed in the next section. The regularity condition that we employ has been studied by many authors in various forms and contexts [1, 5, 13, 23, 27, 32, 37, 38, 39, 42, 46]. It is related to the stability of the set of solutions of the inclusion (5). In particular, the regularity hypothesis implies that the local behavior of the Gauss–Newton method presented in Section 2 mimics that of the method proposed by Maguregui [23, 24]. Maguregui studies the procedure in the more general Banach space setting and obtains a convergence result that extends Kantorovich's convergence theory for Newton's method to smooth nonlinear convex inclusions. The key ingredient in Maguregui's proof theory is the Robinson–Ursescu Theorem [39, Theorem 2], [43] on the stability of convex multifunctions. In this article, we provide a self–contained and elementary proof theory in the finite dimensional case and link this theory to a globalization strategy via the Gauss–Newton methodology for convex composite optimization. Our proof technique requires a stability result similar to the one Maguregui obtains from the Robinson–Ursescu Theorem. However, the proof we provide for this result (Proposition 3.3) is quite elementary involving a straightforward application of Fenchel duality. It is hoped that this approach will make these results more accessible to the mathematical programming community.

The regularity condition is discussed in Section 3. There we establish a few equivalent formulations of the condition useful to our study and also present our main stability result for the inclusion (5). The basic Gauss–Newton algorithm is presented in detail in Section 2 and its convergence properties are established in Section 4. We conclude in Section 5 by presenting a global version of the method and establishing its convergence properties.

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 $\mathbb{R}^n$ is defined as the bi-linear form

$$\langle y, x \rangle := \sum_{i=1}^{n} y_i x_i.$$

Given two subsets U and V of $\mathbb{R}^n$ and $\beta \in \mathbb{R}$, we define

$$U \pm \beta V := \{u \pm \beta v \mid u \in U,\ v \in V\}$$

and

$$U \setminus V := \{u \in U \mid u \notin V\}.$$

If $U \subset \mathbb{R}^n$ then the *polar* of U is defined to be the set

$$U^\circ := \{x^* \in \mathbb{R}^n \mid \langle x^*, x\rangle \leq 1 \ \forall x \in U\}.$$

The *indicator* and *support* functions for U are given by

$$\psi_U(x) := \begin{cases} 0 & \text{if } x \in U \\ +\infty & \text{otherwise,} \end{cases}$$

and

$$\psi_U^*(x) := \sup\{\langle x^*, x\rangle \mid x^* \in U\},$$

respectively. The *relative interior* of U, denoted by $\operatorname{ri} U$, is the interior of U relative to the *affine hull* of U which is given by

$$\operatorname{aff} U := \left\{ \sum_{k=1}^{s} \lambda_k x^k \;\middle|\; \begin{array}{c} s \in \{1, 2, \cdots\},\ x^k \in U \text{ and } \lambda_k \in \mathbb{R} \\ \text{for } k = 1, 2, \cdots, s, \text{ with } \sum_{k=1}^{s} \lambda_k = 1 \end{array} \right\}.$$

The *closure* of U, $\operatorname{cl} U$ is the usual topological closure of the set U. The *cone* generated by U is defined by

$$\operatorname{cone}(U) := \{\lambda u \mid \lambda > 0, u \in U\}.$$

For a given function f, $z \in \arg\min\{f(x) \mid x \in U\}$ means that $z \in U$ and $f(z) = \min\{f(x) \mid x \in U\}$. Furthermore, for a convex function f, ∂f signifies the subdifferential multifunction [40].

We denote a *norm* on $\mathbb{R}^\nu$ by $\|\cdot\|$. The associated closed unit ball for the given norm is denoted by $\mathbb{B}$. Each norm has an associated dual norm given by

$$\|x\|_\circ := \psi_{\mathbb{B}}^*(x).$$

It is straightforward to show that the unit ball associated with the dual norm is the set $\mathbb{B}^\circ$. The space in which the sets $\mathbb{B}$ and $\mathbb{B}^\circ$ lie and their corresponding norms is always apparent from the context of the discussion. If there is ever the possibility of confusion, we write the unit ball as $\mathbb{B}_\nu$ to specify that $\mathbb{B}_\nu \subset \mathbb{R}^\nu$.

For a given set U, the distance of a point x to U is given by

$$\operatorname{dist}(x \mid U) := \inf_{u \in U} \|x - u\|.$$

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:

$$D_\Delta(x) := \arg\min \{h(F(x) + F'(x)d) \mid \|d\| \leq \Delta\} \ . \tag{6}$$

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

(1) Choose $d^k \in D_\Delta(x^k)$ to satisfy

$$\left\|d^k\right\| \leq \eta \operatorname{dist}\big(0 \mid D_\Delta(x^k)\big) \, . \tag{7}$$

If $d^k = 0$, then stop.

(2) Set $x^{k+1} := x^k + d^k$.

Algorithms of this type have been extensively studied in the literature. However, in most studies, the objective function in the direction finding subproblem,

$$\min \{h(F(x) + F'(x)d) \mid \|d\| \leq \Delta\} \ ,$$

includes a quadratic term of the form $\frac{1}{2}d^T Hd$ in order to incorporate some curvature components. Further discussion of this curvature component can be found [11, 29] for the classical Gauss-Newton method and in [15, 35] for convex composite optimization. The relationship of this component to second–order optimality conditions can be found in [7, 15, 18, 42]. 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 $F(\bar{x}) \in C := \arg\min h$. 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 (5). The weak sharp minima concept was introduced in [12].

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*

$$h(y) \geq h_{min} + \alpha \operatorname{dist}(y \mid C) \, , \ \textit{for all } y \in \mathbb{R}^m \, , \tag{8}$$

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 (8) 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 $\text{dist}(\cdot \mid C)$ itself. Other examples are explored in [6, 12]. For example, if h is polyhedral convex, then its set of minima is necessarily a set of weak sharp minima. Moreover, in this case, if it is further assumed that the norm on $\mathbb{R}^n$ is polyhedral, then one can obtain a direction choice satisfying (7) of Step 1 in Algorithm 1 by computing a least–norm solution of a linear program in the sense of [25, 26]. Numerical methods for obtaining least–norm solutions to linear programs have been developed in [10].

The notion of weak sharp minima generalizes the notion of a *sharp* [33, 34] or *strongly unique* [9, 22, 30, 31, 44] minimum. These concepts have a long history in the literature and have far reaching consequences for the convergence analysis of many iterative procedures [9, 19, 22, 33, 34, 44]. In [6], 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 develop a notion of regularity for the inclusion (5). Regularity is the basic tool in our analysis of the linearized inclusion

$$F(x) + F'(x)d \in C, \text{ with } \|d\| \leq \Delta \ . \tag{9}$$

In particular, the regularity condition implies that the direction finding subproblem (6) is locally equivalent to solving the inclusion (9) in the sense that their solution sets coincide. Regularity also allows us to establish bounds on the local behavior to the solution set of the linearized inclusion (9). Based on these results, we establish 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 of the inclusion (5) if $F(\bar{x}) \in C$ and*

$$\ker(F'(\bar{x})^T) \cap N_C\left(F(\bar{x})\right) = \{0\} \ . \tag{10}$$

Recall that the the normal cone mapping for the convex set C is defined by the relation

$$N_C\left(y\right) := \partial\psi_C(y) \ .$$

In the context of the nonlinear least squares problem (4), the set C is the origin and so the normal cone to C at $F(\bar{x})$ is all of $\mathbb{R}^m$. Therefore the condition (10) reduces to the classical condition for this problem, namely, that the mapping $F'(\bar{x})$ is surjective. Indeed, this is the case whenever h has a unique minimizer, and in particular when h has a strongly unique minimum.

Our first objective in this section is to establish several equivalent forms of regularity that are pertinent to the discussion. We begin by defining an extension of the normal cone mapping to a multifunction whose domain is the whole space. Recall that the the normal cone mapping for the convex set C can be written as

$$N_C(y) = \begin{cases} [\text{cone}(C - y)]^\circ, & \text{if } y \in C, \text{ and} \\ \emptyset, & \text{otherwise.} \end{cases}$$

An extension of this multifunction to one having domain $\mathbb{R}^n$ is given by the mapping $\Gamma_C \colon \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ defined by the relation

$$\Gamma_C(y) := [\text{cone}(C - y)]^\circ, \quad \text{for all } y \in \mathbb{R}^m.$$

Clearly, this multifunction differs from the normal cone mapping only off the set C. It is straightforward to show that Γ_C has the following very useful dual representation:

$$\Gamma_C(z) = \{y \mid \langle y, z\rangle - \psi_C^*(y) \geq 0\}. \tag{11}$$

Observe that we can state the regularity condition (10) in terms of the multifunction Γ_C. By doing so, it is easy to show that the regularity condition is a local property.

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 non-empty closed convex subset of $\mathbb{R}^m$. Then the following statements are equivalent:*

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

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

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

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

$$\bar{A}^{T^{-1}} \mathbb{B}_n^\circ \cap \Gamma_C(\bar{z}) \subset \beta\, \mathbb{B}_m^\circ \,. \tag{12}$$

(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

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

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 parameter β in (iv) depends 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

$$\text{im}\, \bar{A} + \text{cl}\, \text{cone}(C - \bar{z}) = \mathbb{R}^m \ .$$

From this equation, the equivalence follows from a simple separation argument and the fact that $\text{ri}\, \text{cone}\, S = \text{cone}\,(\text{ri}\, S)$ 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 $\bar{A}\,\mathbb{B}_n + (\mathrm{ri}\, C - \bar{z})$. But then the cone generated by this set, namely $\mathrm{im}\,\bar{A} + \mathrm{cone}(\mathrm{ri}\, C - \bar{z})$, 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 $\eta\,\mathbb{B}_m \subset \bar{A}\,\mathbb{B}_n + (\mathrm{ri}\, C - \bar{z})$. This implies that $\eta\,\mathbb{B}_m \subset \bar{A}\,\mathbb{B}_n + \mathrm{cone}\,(C - \bar{z})$. The polar of this last expression is precisely (iv) with $\beta = \eta^{-1}$.

Clearly, (iv) implies (i) since $\ker \bar{A} = \bar{A}^{T^{-1}}0$ 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. As noted above, the condition in (iii) implies the existence of an $\eta > 0$ such that $\eta\,\mathbb{B}_m \subset \bar{A}\,\mathbb{B}_n + (\mathrm{ri}\, C - \bar{z})$. Hence

$$\eta\,\mathbb{B}_m \subset A\,\mathbb{B}_n + (\mathrm{ri}\, C - z) + \frac{\eta}{2}\,\mathbb{B}_m\,,$$

whenever $(z, A) \in (\bar{z}, \bar{A}) + \frac{\eta}{2}\mathbb{B}$. Therefore, by the Rådström Cancellation Lemma [36, Lemma 1],

$$\frac{\eta}{2}\,\mathbb{B}_m \subset A\,\mathbb{B}_n + (\mathrm{ri}\, C - z) \subset A\,\mathbb{B}_n + \mathrm{cone}\,(C - z)\,.$$

Taking the polar of this last statement and setting $\epsilon := \frac{\eta}{2} =: \beta^{-1}$, 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. The form of the regularity condition used by Maguregui [23, 24] most closely resembles condition (iii) above. The actual condition he employs is

$$\bar{z} \in C, \quad 0 \in \mathrm{core}\,(\mathrm{im}\,\bar{A} + C - \bar{z}), \tag{13}$$

   where the *core* of a closed convex set in a Banach space is the same as the interior of the set in the norm topology [41, pg. 31]. Condition (iii) is equivalent to this regularity condition if it is further assumed in (iii) that $\bar{z} \in C$. Observe that the use of the multifunction Γ_C avoids the need to specify that $\bar{z} \in C$ for results such as Lemma 3.2. Moreover, it should be pointed out that these conditions are not equivalent when $\bar{z} \notin C$. For example, take $C = \{2\}$, $\bar{z} = 0$ and A as the identity map on $\mathbb{R}$.

   In the infinite–dimensional setting (13) is stronger than the condition in (i). However, condition (i) is not sufficiently strong to obtain the necessary result. Condition (13) provides the link to Maguregui's application of the Robinson–Ursescu Theorem.

2. By taking $A = F'(\bar{x})$ and $\bar{z} = F(\bar{x})$, Lemma 3.2(v) implies that

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

   for all points x near $\bar{x}$ at which (10) holds. That is, regularity is a local property.

The following proposition states that the linearized inclusions (9) are solvable in a strong sense near regular points. It is the key technical result required to establish the local quadratic convergence of Algorithm 1. We show that under the regularity hypothesis the distance to the solution set of the linearized inclusion can be bounded locally by the distance of the linearization to the set C. This result is reminiscent of several similar results due to Robinson [37, 38, 39, 43]. Indeed, the bound (14) is easily derived from the Robinson–Ursescu Theorem. On the other hand, our proof of the result is a simple application of Fenchel's Duality Theorem [40, Corollary 31.2.1].

Proposition 3.3 *If $\bar{x}$ is a regular point of (5), then for all $\Delta > 0$, there is some neighborhood $\mathcal{N}(\bar{x})$ of $\bar{x}$ and a $\beta > 0$ satisfying*

$$\operatorname{dist}(0 \mid D_{\Delta}(x)) \leq \beta \operatorname{dist}(F(x) \mid C) , \tag{14}$$

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

$$F(x) + F'(x)d \in \operatorname{ri} C \ , \tag{15}$$

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

Proof We first establish (14). Since $\operatorname{dist}(F(x) \mid C)$ is a continuous function of x and $F(\bar{x}) \in C$, the relation (14) follows from the inequality

$$\operatorname{dist}(0 \mid D_{\infty}(x)) \leq \beta \operatorname{dist}(F(x) \mid C) , \tag{16}$$

for all x sufficiently close to $\bar{x}$. This latter inequality follows easily from the following argument based on Fenchel duality.

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 the pair $(F(x), F'(x)) \in (F(\bar{x}), F'(\bar{x})) + \epsilon \mathbb{B}$ whenever $x \in \mathcal{N}(\bar{x})$.

The Fenchel dual to the problem

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

is the problem

$$\sup\{ \langle y, F(x) \rangle - \psi^*_C(y) \mid F'(x)^T y \in \mathbb{B}^\circ \} \ , \tag{17}$$

and the optimal values of these problems coincide with attainment in (17) if there exists $d \in \mathbb{R}^n$ such that

$$F(x) + F'(x)d \in \operatorname{ri} C \ . \tag{18}$$

The existence of such a vector d is guaranteed by Parts (iii) and (v) of Lemma 3.2 for all $x \in \mathcal{N}(\bar{x})$. Hence

$$\operatorname{dist}(0 \mid D_{\infty}(x)) = \max\{ \langle y, F(x) \rangle - \psi^*_C(y) \mid F'(x)^T y \in \mathbb{B}^\circ \} \ . \tag{19}$$

Moreover, since $0 \leq \operatorname{dist}(0 \mid D_\infty(x))$, the constraint region on the right–hand side of (19) can be further restricted by adding the inclusion $y \in \Gamma_C(F(x))$. This follows from the identity (11). This observation along with Parts (iv) and (v) of Lemma 3.2 yields the relation

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

for all $x \in \mathcal{N}(\bar{x})$, where the last equality follows from [4, Lemma 2.9].

To complete the proof, we now establish (15). Let $\Delta > 0$ be given. By (14), there is a neighborhood of $\bar{x}$ on which $D_{\frac{\Delta}{2}}(x) \neq \emptyset$. Let $d_1 \in D_{\frac{\Delta}{2}}(x)$. By Parts (iii) and (v) of Lemma 3.2, there is a $d_2 \in \mathbb{R}^n$ satisfying (18). By [40, 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 $(1-t)d_1 + td_2 \in \Delta\, \mathbb{B}$.
□

The remainder of this section is not critical to the development of the convergence theory for our algorithm. We show that regularity combined with weak sharpness implies that the composite function is also weak sharp in a local sense (see Proposition 3.4). In the proposition above, we have obtained (14), essentially a linear result, by way of Fenchel duality. However, a stronger result is obtained by Robinson [38, Theorem 1], where it is shown that the same regularity assumption implies

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

for all x in a neighborhood of the regular point $\bar{x}$ and some $\eta > 0$. It is immediate from this fact that regularity of $\bar{x}$ for the inclusion $F(x) \in C$ coupled with an assumption that h is sharp on C implies that the composite function $h \circ F$ is locally sharp [6] with respect to the set $F^{-1}(C)$. This observation is summarized in the following proposition.

Proposition 3.4 *Let $\bar{x} \in \mathbb{R}^n$ be a regular point of the inclusion (5) where C is a set of weak sharp minima for h. Then there is a $\gamma > 0$ and a neighborhood $\mathcal{N}(\bar{x})$ of $\bar{x}$ such that*

$$f(x) = h(F(x)) \geq f(\bar{x}) + \gamma \operatorname{dist}\big(x \mid F^{-1}(C)\big), \tag{20}$$

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

This result allows us to relate the hypotheses used in the current paper to those used by Womersley [44]. Womersley assumes that the compostite function f has a strongly unique minimum, and uses this assumption to derive a quadratic convergence rate for his Gauss–Newton procedure. In this paper, we make the assumption that $\bar{x} \in \mathbb{R}^n$ is a regular point of the inclusion (5) where C is a set of weak sharp minima for h and show in the next section that this guarantees a quadratic convergence rate for our Gauss–Newton procedure. This is the fundamental difference between the two approaches: Proposition 3.4 shows that our assumptions guarantee local sharpness of f, whereas Womersley assumes that f has a strongly unique minimum. Consequently, our proof theory is based on sufficient conditions to ensure 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. For this we require the following technical result whose proof is found in the Appendix.

Lemma 4.1 *Suppose $\{x^k\} \subset \mathbb{R}^n$, $x^k \to x^*$ and*

$$\limsup_{k\to\infty} \frac{\left\|x^{k+1} - x^k\right\|}{\left\|x^k - x^{k-1}\right\|^2} \le M.$$

Then, $x^k \to x^*$ quadratically, that is

$$\limsup_{k\to\infty} \frac{\left\|x^{k+1} - x^*\right\|}{\left\|x^k - x^*\right\|^2} \le \hat{M}.$$

The following result extends the standard domain of attraction theory for Gauss–Newton methods to finite–valued convex composite optimization under the weak sharpness and regularity assumptions. In particular, this extends the quadratic convergence result under strong uniqueness due to Womersley [44]. Our result is local and guarantees a quadratic rate of convergence. Related results can be found in [24, 19, 30, 31].

Theorem 4.2 *Let $\bar{x} \in \mathbb{R}^n$ be a regular point of the inclusion (5) where C is a set of weak sharp minima for h and suppose F' is locally Lipschitz at $\bar{x}$. Then there is a neighborhood $\mathcal{M}(\bar{x})$ of $\bar{x}$ such that if the algorithm 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 Let $\mathcal{N}(\bar{x})$ be the neighborhood of $\bar{x}$ postulated by Proposition 3.3 and assume without loss that F' is Lipschitz on $\mathcal{N}(\bar{x})$ with Lipschitz constant L. Note that h (being finite–valued and convex) is Lipschitzian on the bounded set $F(\mathcal{N}(\bar{x})) + \frac{L}{8}\mathbb{B}$. Denote the corresponding Lipschitz constant by M. Choose $1 \ge \delta > 0$ such that

$$\frac{\eta L M \delta \beta}{\alpha} < 1 \text{ and } \bar{x} + 2\delta\,\mathbb{B} \subset \mathcal{N}(\bar{x}). \tag{21}$$

Since $\operatorname{dist}(F(x) \mid C)$ is continuous, there is a neighborhood $\mathcal{O}(\bar{x})$ such that

$$\operatorname{dist}(F(x) \mid C) \le \frac{\delta}{2\eta\beta}, \text{ for all } x \in \mathcal{O}(\bar{x}). \tag{22}$$

Let $\mathcal{M}(\bar{x}) := \mathcal{N}(\bar{x}) \bigcap \mathcal{O}(\bar{x}) \bigcap (\bar{x} + \delta\,\mathbb{B})$. We claim that for $k = 0, 1, 2, \ldots$

$$x^k \in \mathcal{N}(\bar{x}), \tag{23}$$

$$\left\|d^k\right\| \le \eta\beta\alpha^{-1}(h(F(x^k)) - h_{min}) \le \frac{\eta LM\beta}{2\alpha}\left\|d^{k-1}\right\|^2 \le \frac{\delta}{2}2^{[-2^k]}. \tag{24}$$

Note that (23) implies that the algorithm is well defined for $k = 0, 1, 2, \ldots$. The proof of the claim proceeds by induction on k. Note that

$$\left\|d^0\right\| \le \eta\beta \operatorname{dist}\left(F(x^0) \mid C\right) \le \frac{\delta}{2}, \tag{25}$$

the first inequality following from (7) and (14) and the second inequality from (22). Since $x^0 \in \mathcal{M}(\bar{x})$ we have

$$\left\|x^1 - \bar{x}\right\| \le \left\|x^0 - \bar{x}\right\| + \left\|d^0\right\| \le \delta + \frac{\delta}{2} < 2\delta.$$

Hence $x^1 \in \bar{x} + 2\delta\,\mathbb{B} \subset \mathcal{N}(\bar{x})$. Furthermore, by the quadratic bound lemma [29, 3.2.12]

$$\left\|F(x^0) + F'(x^0)d^0 - F(x^1)\right\| \le \frac{L}{2}\left\|d^0\right\|^2 \le \frac{L}{8},$$

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

$$\begin{aligned} h(F(x^1)) &= h(F(x^0) + F'(x^0)d^0 + \textstyle\int_0^1 (F'(x^0 + td^0) - F'(x^0))d^0 dt) \\ &\le h(F(x^0) + F'(x^0)d^0) + M\left\|\textstyle\int_0^1 (F'(x^0 + td^0) - F'(x^0))d^0 dt\right\| \\ &= h_{min} + M\left\|\textstyle\int_0^1 (F'(x^0 + td^0) - F'(x^0))d^0 dt\right\| \\ &\le h_{min} + \tfrac{LM}{2}\left\|d^0\right\|^2, \end{aligned} \tag{26}$$

the second equality following from (15). We now have that

$$\begin{aligned} \|d^1\| &\le \eta\beta\alpha^{-1}(h(F(x^1)) - h_{min}) && \text{(from (7), (14) and (8))} \\ &\le \tfrac{\eta LM\beta}{2\alpha}\left\|d^0\right\|^2 && \text{(from (26))} \\ &\le \tfrac{\eta LM\beta}{2\alpha}\left(\tfrac{\delta}{2}\right)^2 && \text{(from (25))} \\ &\le \tfrac{\delta}{2}2^{-2}. && \text{(from (21))} \end{aligned}$$

Next we assume that (23) and (24) hold for $k = 0, 1, \ldots, s$ and show that it also holds for $k = s + 1$. First of all, since $x^0 \in \mathcal{M}(\bar{x})$ we have

$$\left\|x^{s+1} - \bar{x}\right\| \le \left\|x^0 - \bar{x}\right\| + \sum_{k=0}^{s}\left\|d^k\right\| \le \delta + \frac{\delta}{2}\sum_{k=0}^{s} 2^{[-2^k]} \le \delta + \frac{\delta}{2}\sum_{k=0}^{s} 2^{-2k} < 2\delta,$$

so that $x^{s+1} \in \mathcal{N}(\bar{x})$. Therefore, as in (26), $h(F(x^{s+1})) \leq h_{min} + \frac{LM}{2}\|d^s\|^2$, and so by the induction hypothesis we obtain

$$\begin{aligned} \left\|d^{s+1}\right\| &\leq \eta\beta\alpha^{-1}(h(F(x^{s+1})) - h_{min}) \\ &\leq \frac{\eta LM\beta}{2\alpha}\|d^s\|^2 \\ &\leq \frac{\eta LM\beta}{2\alpha}\left(\frac{\delta}{2}2^{-2^s}\right)^2 \\ &\leq \frac{\delta}{2}(2^{[-2^s]})^2 = \frac{\delta}{2}2^{[-2^{s+1}]}, \end{aligned}$$

which concludes our induction.

Therefore, the sequence is Cauchy, and so must converge to some x^* satisfying $h(F(x^*)) = h_{min}$. Moreover, the quadratic rate of convergence for $\{x^k\}$ follows from (24) and Lemma 4.1 while the quadratic rate of convergence for $\{h(F(x^k))\}$ is obvious from (24). □

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

(1) Choose $d^k \in D_\Delta(x^k)$ to satisfy

$$\left\|d^k\right\| \leq \eta \,\mathrm{dist}\left(0 \mid D_\Delta(x^k)\right), \tag{27}$$

where

$$D_\Delta(x) := \arg\min\{h(F(x) + F'(x)d) \mid \|d\| \leq \Delta\} \;. \tag{28}$$

If $d^k = 0$, then stop.

(2) Set

$$\begin{aligned} t_k \;:=\; & \max \gamma^s \\ & \text{subject to } s \in \{0, 1, \dots\}, \text{ and} \\ & h(F(x^k + \gamma^s d^k)) - h(F(x^k)) \\ & \qquad \leq c\gamma^s[h(F(x^k) + F'(x^k)d^k) - h(F(x^k))] \end{aligned}$$

(3) Set $x^{k+1} := x^k + t_k d^k$.

Algorithm 2 is an instance of the class of algorithms studied in [3]. Hence the global convergence properties of the method follow from Theorems 2.4 and 5.3 in [3]. These theorems specify the behavior of sequences generated by Algorithm 2 in terms of the first–order optimality conditions for the problem $\mathcal{P}$. A point $\bar{x}$ is a first–order stationary point for the problem $\mathcal{P}$ if

$$f'(\bar{x}; d) \geq 0 \text{ for all } d \in \mathbb{R}^n , \tag{29}$$

where $f'(\bar{x}; \cdot)$ is the usual directional derivative of f at the point $\bar{x}$. This condition can be equivalently stated in terms of the convex subdifferential of h as

$$0 \in F'(\bar{x})^T \partial h(F(\bar{x})) . \tag{30}$$

Moreover, by [3, Lemma 4.5 and Theorem 3.6], the conditions (29) and (30) are equivalent to the conditions

$$h(F(\bar{x}) + F'(\bar{x})d) - h(F(\bar{x})) = 0 \quad \text{for all } d \in D_\Delta(\bar{x}) \tag{31}$$

and

$$0 \in D_\Delta(\bar{x}) . \tag{32}$$

These latter conditions are particularly important in light of the search direction and step–length choice specified in Algorithm 2.

The key global properties established in [3] for Algorithm 2 are recalled in 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 $\{x \in \mathbb{R}^n \colon f(x) \leq f(x^0)\}$ and that h is Lipschitz on the set $\{y \in \mathbb{R}^m \colon h(y) \leq f(x^0)\}$. 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 $\{f(x^k)\}$ decrease 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 non–empty 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 show that if C is a set of weak sharp minima for h and there is a cluster point of the sequence $\{x^k\}$ that is a regular point of the inclusion (5), then the entire sequence must converge to this cluster point at a quadratic rate.

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 (5) where C is a set of weak sharp minima for h, then $x^k \to \bar{x}$ and $f(x^k) \to h_{min}$ at a quadratic rate.*

Proof Since the regularity hypothesis at $\bar{x}$ implies that $F(\bar{x}) \in C$, the result will follow immediately from Theorem 4.2 if it can be shown that there is a neighborhood, N of $\bar{x}$ such that

$$h(F(x+d)) - h(F(x)) \le c[h(F(x) + F'(x)d) - h(F(x))], \tag{33}$$

for all $x \in N$ and $d \in \mathbb{R}^n$ satisfying $\|d\| \le \eta \operatorname{dist}(0 \mid D_\Delta(x))$. Indeed, if (33) holds, then Algorithms 1 and 2 generate identical iterates sufficiently close to $\bar{x}$. Hence, by Theorem 4.2, 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 (33) does not hold near $\bar{x}$. Then there is a sequence $\{\bar{x}^k\}$ converging to $\bar{x}$ such that

$$c[h(F(\bar{x}^k) + F'(\bar{x}^k)\bar{d}^k) - h(F(\bar{x}^k))] < h(F(\bar{x}^k + \bar{d}^k)) - h(F(\bar{x}^k)) \tag{34}$$

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

$$\left\|\bar{d}^k\right\| \le \eta \operatorname{dist}\Big(0 \mid D_\Delta(\bar{x}^k)\Big). \tag{35}$$

In particular, we obtain from (14) that

$$\left\|\bar{d}^k\right\| \to 0. \tag{36}$$

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 [29, 3.2.12] that

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

Therefore, by (34)

$$\begin{array}{rcl} c[h_{min} - h(F(\bar{x}^k))] & = & c[h(F(\bar{x}^k) + F'(\bar{x}^k)\bar{d}^k) - h(F(\bar{x}^k))] \\ & < & h(F(\bar{x}^k + \bar{d}^k)) - h(F(\bar{x}^k)) \\ & \leq & h_{min} - h(F(\bar{x}^k)) + \frac{KM}{2} \left\|\bar{d}^k\right\|^2 . \end{array}$$

Consequently,

$$\begin{array}{rll} 0 & < & (1-c)[h_{min} - h(F(\bar{x}^k))] + \frac{KM}{2} \left\|\bar{d}^k\right\|^2 \\ & \leq & (c-1)\alpha \operatorname{dist}\left(F(\bar{x}^k) \mid C\right) + \frac{KM}{2} \left\|\bar{d}^k\right\|^2 \qquad \text{(from (8))} \\ & \leq & (c-1)\alpha(\beta\eta)^{-1} \left\|\bar{d}^k\right\| + \frac{KM}{2} \left\|\bar{d}^k\right\|^2 , \qquad \text{(from (14) and (35))} . \end{array}$$

After dividing this expression through by $\left\|\bar{d}^k\right\|$ and using (36) while taking the limit in k, we obtain the contradiction

$$0 \leq (c-1)\alpha(\beta\eta)^{-1} .$$

□

A Appendix

We now establish Lemma 4.1. Our proof is based on [16, Lemma 3.7].
Lemma 4.1 Suppose $\{x^k\} \subset \mathbb{R}^n$, $x^k \rightarrow x^*$ and

$$\limsup_{k\to\infty} \frac{\left\|x^{k+1} - x^k\right\|}{\left\|x^k - x^{k-1}\right\|^2} \leq M.$$

Then, $x^k \rightarrow x^*$ quadratically, that is

$$\limsup_{k\to\infty} \frac{\left\|x^{k+1} - x^*\right\|}{\left\|x^k - x^*\right\|^2} \leq \hat{M}.$$

Proof Let $0 < \mu < 1/2$. Then there is some N depending on μ such that

$$\left\|x^{k+1} - x^k\right\| \leq M \left\|x^k - x^{k-1}\right\|^2 \leq \mu \left\|x^k - x^{k-1}\right\|$$

for all $k \geq N$. Now

$$\begin{array}{rcl} \left\|x^{k+1} - x^{k+i+1}\right\| & \leq & \left\|x^{k+1} - x^{k+2}\right\| + \ldots + \left\|x^{k+i} - x^{k+i+1}\right\| \\ & \leq & (1 + \mu + \ldots + \mu^i) \left\|x^{k+1} - x^{k+2}\right\| \\ & = & \frac{1-\mu^{i-1}}{1-\mu} \left\|x^{k+1} - x^{k+2}\right\| . \end{array}$$

In the limit as $i \to \infty$ we have

$$\left\|x^{k+1} - x^*\right\| \leq \frac{\left\|x^{k+1} - x^{k+2}\right\|}{1-\mu}. \tag{37}$$

Similarly,

$$\begin{aligned} \left\|x^k - x^{k+i}\right\| &\geq \left\|x^k - x^{k+1}\right\| - \left\|x^{k+1} - x^{k+2}\right\| - \ldots - \left\|x^{k+i-1} - x^{k+i}\right\| \\ &\geq (1 - (\mu - \mu^i)/(1-\mu)) \left\|x^k - x^{k+1}\right\|, \end{aligned}$$

which in the limit as $i \to \infty$ gives

$$\left\|x^k - x^*\right\| \geq \frac{1-2\mu}{1-\mu} \left\|x^k - x^{k+1}\right\|. \tag{38}$$

Therefore, it follows from (37) and (38) that for all $k \geq N$,

$$\begin{aligned} \frac{\left\|x^{k+1} - x^*\right\|}{\left\|x^k - x^*\right\|^2} &\leq \frac{1-\mu}{(1-2\mu)^2} \frac{\left\|x^{k+1} - x^{k+2}\right\|}{\left\|x^k - x^{k+1}\right\|^2} \\ &\leq \frac{1-\mu}{(1-2\mu)^2} M, \end{aligned}$$

as required. □

References

[1] J.M. Borwein Stability and regular points of inequality systems. *J. Optim. Theory Appl.*, 48:9–52, 1986.

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

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

[4] J.V. Burke. A sequential quadratic programming method for potentially infeasible mathematical programs. *Journal of Mathematical Analysis and its Applications*, 139(2):319–351, 1989.

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

[6] J.V. Burke and M.C. Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5), 1993.

[7] J.V. Burke and R.A. Poliquin. Optimality conditions for non finite–valued convex composite functions. *Mathematical Programming*, 57:103–120, 1992.

[8] F.H. Clarke. *Optimization and Nonsmooth Analysis.* John Wiley & Sons, New York, 1983.

[9] L. Cromme. Strong uniqueness. *Numerische Mathematik*, 29:179–193, 1978.

[10] R. De Leone and O.L. Mangasarian. Serial and parallel solution of large scale linear programs by augmented Lagrangian successive overrelaxation. In A. Kurzhanski, K. Neumann, and D. Pallaschke, editors, *Optimization, Parallel Processing and Applications*, pages 103–124, Berlin, 1988. Springer–Verlag. Lecture Notes in Economics and Mathematical Systems 304.

[11] J.E. Dennis and R.B. Schnabel. *Numerical Methods for Unconstrained Optimizations and Nonlinear Equations.* Prentice Hall, New Jersey, 1983.

[12] M.C. Ferris. *Weak Sharp Minima and Penalty Functions in Mathematical Programming.* PhD thesis, University of Cambridge, 1988.

[13] A.V. Fiacco. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming.* Academic Press, New York, 1983.

[14] R. Fletcher. Second order correction for nondifferentiable optimization. In G.A. Watson, editor, *Numerical Analysis*, pages 85–114. Springer–Verlag, Berlin, 1982.

[15] R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, New York, second edition, 1987.

[16] U.M. Garcia–Palomares. Superlinearly convergent algorithms for linearly constrained optimization. In R.R. Meyer O.L. Mangasarian and S.M. Robinson, editors, *Nonlinear Programming 2*, pages 101–119, New York, 1975. Academic Press.

[17] U.M. Garcia–Palomares and A. Restuccia. A global quadratic algorithm for solving a system of mixed equations. *Mathematical Programming*, 21:290–300, 1981.

[18] A.D. Ioffe. Variational analysis of a composite function: A formula for the lower second–order directional derivative. *Journal of Mathematical Analysis and its Applications*, 1993. Forthcoming.

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

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

[21] D.G. Luenberger. *Optimization by Vector Space Methods.* John Wiley & Sons, New York, 1969.

[22] K. Madsen. *Minimization of Nonlinear Approximation Functions.* PhD thesis, Institute of Numerical Analysis, Technical University of Denmark, Lyngby, Denmark, 1985.

[23] J. Maguregui. *Regular Multivalued Functions and Algorithmic Applications.* PhD thesis, University of Wisconsin, Madison, Wisconsin, 1977.

[24] J. Maguregui. A modified Newton algorithm for functions over convex sets. In O.L. Mangasarian, R.R. Meyer, and S.M. Robinson editors, *Nonlinear Programming 3*, pages 461–473, Academic Press, 461–473, 1978

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

[26] O.L. Mangasarian. Normal solutions of linear programs. *Mathematical Programming Study*, 22:206–216, 1984.

[27] O.L. Mangasarian and S. Fromovitz. The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. *J. Math. Anal. Appl.*, 17:37–47, 1967.

[28] D.W. Marquardt. An algorithm for least–squares estimation of nonlinear parameters. *SIAM Journal of Applied Mathematics*, 11:431–441, 1963.

[29] J.M. Ortega and W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, 1970.

[30] 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.

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

[32] J.–P. Penot. On regularity conditions in mathematical programming. *Math. Prog. Studies*, 17:28–66, 1982.

[33] B.T. Polyak. Sharp minima. A Talk given at the IIASA Workshop on Generalized Lagrangians and their Applications, IIASA, Laxenburg, Austria, 1979.

[34] B.T. Polyak. *Introduction to Optimization.* Optimization Software, Inc., Publications Division, New York, 1987.

[35] M.J.D. Powell. General algorithm for discrete nonlinear approximation calculations. In L.L. Schumaker C.K. Chui and J.D. Ward, editors, *Approximation Theory IV*, pages 187–218. Academic Press, New York, 1983.

[36] H. Rådström An embedding theorem for spaces of convex sets. *Proc. Amer. Math. Soc.*, 3:165–169, 1952.

[37] S.M. Robinson Stability theory for systems of inequalities, Part I: linear systems. *SIAM J. Numer. Anal.*, 12:754–769, 1975.

[38] S.M. Robinson Stability theory for systems of inequalities, Part II: differentiable nonlinear systems. *SIAM J. Numer. Anal.*, 13:497–513, 1976.

[39] S.M. Robinson Regularity and stability for convex multivalued functions. *Math. of O.R.*, 1:130–143, 1976.

[40] R.T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

[41] R.T. Rockafellar. *Conjugate Duality and Optimization*, volume 16, National Science Foundation CBMS Series. SIAM, Philadelphia, PA, 1974.

[42] R.T. Rockafellar. First– and second–order epi–differentiabilty in nonlinear programming. *Transactions of the American Mathematical Society*, 307:75–108, 1988.

[43] C. Ursescu. Multifunctions with closed convex graph. *Czech. Math. J.*, 35:438–441, 1975.

[44] R.S. Womersley. Local properties of algorithms for minimizing composite functions. *Mathematical Programming*, 32:69–89, 1985.

[45] Y. Yuan. On the superlinear convergence of a trust region algorithm for nonmooth optimization. *Mathematical Programming*, 31:269–285, 1985.

[46] J. Zowe and S. Kurcyusz. Regularity and stability for mathematical programming in Banach spaces. *Appl. Math. Optim.*, 5:49–62, 1979.