Recent Advances in Nonsmooth Optimization, pp. 57-86
Eds. D.-Z. Du, L. Qi and R.S. Womersley
©1995 World Scientific Publishers

Projected Gradient Methods for Nonlinear Complementarity Problems via Normal Maps

Michael C. Ferris[1]
University of Wisconsin–Madison, Computer Sciences Department, Madison, WI 53706, USA

Daniel Ralph[2]
University of Melbourne, Department of Mathematics, Melbourne, Australia

Abstract

We present a new approach to solving nonlinear complementarity problems based on the normal map and adaptations of the projected gradient algorithm. We characterize a Gauss–Newton point for nonlinear complementarity problems and show that it is sufficient to check at most two cells of the related normal manifold to determine such points. Our algorithm uses the projected gradient method on one cell and n rays to reduce the normed residual at the current point. Global convergence is shown under very weak assumptions using a property called nonstationary repulsion. A hybrid algorithm maintains global convergence, with quadratic local convergence under appropriate assumptions.

1 Introduction

The nonlinear complementarity problem is to find a vector $z \in \mathrm{R}^n$ satisfying:

$$f(z) \geq 0, \quad z \geq 0, \quad \langle f(z), z \rangle = 0, \tag{NCP}$$

[1]The work of this author was based on research supported by the National Science Foundation grant CCR-9157632 and the Air Force Office of Scientific Research grant F49620-94-1-0036.

[2]The work of this author was based on research partially supported by the U.S. Army Research Office through the Mathematical Sciences Institute, Cornell University, the National Science Foundation, the Air Force Office of Scientific Research, the Office of Naval Research, under grant DMS-8920550, and the Australian Research Council.

where $f : \mathbf{R}^n \to \mathbf{R}^n$ is a smooth function and all vector inequalities are taken component-wise.

In this paper, we will describe an algorithm for solving nonlinear complementarity problems that is computationally based on the projected gradient algorithm, and uses a reformulation of (NCP) as a system of nonsmooth equations. The algorithm is conceptually simple to implement and has a low cost per iteration; and we demonstrate its convergence properties assuming only that f is continuously differentiable.

The problem (NCP) can be reformulated using a normal map:

$$0 = f_+(x) \stackrel{\mathrm{def}}{=} f(x_+) + x - x_+, \tag{NE}$$

where x_+ is the Euclidean projection of x onto R^n_+. Note that z solves (NCP) if and only if $z - f(z)$ solves (NE), and x solves (NE) if and only if x_+ solves (NCP). Normal maps were introduced by Robinson in [32] (see also [29, 30]) and we note here simply that the formulation (NE) has some advantages over (NCP). For example, it is an equation rather than a system of inequalities and equalities, hence its examination from the viewpoint of equations may yield insight difficult to obtain otherwise. This has proven to be the case as demonstrated by recent advances on nonsmooth Newton-like algorithms for (NE) in [5, 4, 12, 28, 34]. Nonsmoothness of the normal map, however, is the difficulty assumed.

In fact, normal maps such as f_+ can be cast in a more general framework, where x_+ is replaced by $\pi_\Omega(x)$, the projection of x onto a nonempty closed convex set Ω. In this context, finding a zero of the normal map

$$f_\Omega(x) \stackrel{\mathrm{def}}{=} f(\pi_\Omega(x)) + x - \pi_\Omega(x)$$

is equivalent to a nonlinear variational inequality [11] defined by the set Ω and the function f. In the special case where $\Omega \equiv \mathrm{R}^n_+$, $f_\Omega = f_+$. For polyhedral Ω, the normal map [31, 33] f_Ω is intimately related to the normal manifold [32]. This manifold is constructed using the faces of the set Ω; it is a collection of n-dimensional polyhedral sets (called cells) which partition R^n. The normal map f_Ω is smooth in each cell of R^n; nondifferentiability only can occur as x moves from one cell to another. A cell is sometimes called a piece of linearity. In the particular example resulting from nonlinear complementarity problems where $\Omega \equiv \mathrm{R}^n_+$, the cells of the normal manifold are precisely the orthants of R^n.

Practical Newton-like methods for (NE) solve a linear or piecewise linear model based at the kth iterate, x^k, to obtain the next iterate x^{k+1}. Unfortunately, this model is not always invertible and this creates problems for defining algorithms and in computing x^{k+1}. In this paper, we are concerned with defining practical algorithms with strong global convergence properties for finding zeros of normal maps. Our goal is to obtain convergence, at least on a subsequence, to a *Gauss–Newton point* for normal maps. This generalizes the familiar notion from nonlinear equation theory where a Gauss–Newton point is a stationary point for the problem of minimizing the Euclidean norm residual of the function.

We are ultimately interested in zeros of f_Ω but finding one may be on the level of difficulty of finding zeros of general nonlinear functions. We revert to considering the residual function

$$\theta(x) \stackrel{\text{def}}{=} \min \frac{1}{2} \|f_\Omega(x)\|^2 ,$$

which gives us a measure of the violation of satisfying $f_\Omega(x) = 0$. Our aim in this paper is to develop a robust algorithm for minimizing θ that has a low cost per iteration. Note that θ is a piecewise smooth function. In order to motivate our definition of Gauss–Newton points, let us first examine the notion of a Gauss–Newton point for nonlinear equations. This corresponds to the case where $\Omega \equiv \mathrm{R}^n$, and $f_\Omega = f$. A Gauss–Newton point for the smooth function f is a point $x^* \in \mathrm{R}^n$ such that $x = x^*$ minimizes the first-order model $\frac{1}{2} \|f(x^*) + \nabla f(x^*)(x - x^*)\|^2$ of $\theta(x)$ over R^n. For general Ω, we construct a piecewise linear model of the residual function θ based on the directional derivative $f'_\Omega(x^*; \cdot)$.

There are several key ideas on which the development of this paper are based.

(i) The characterization of Gauss–Newton points for normal maps requires the stationarity of the residual function θ with respect to every cell that contains that Gauss–Newton point. Thus, for complementarity problems, we must examine up to 2^n orthants to determine whether or not x^* is a Gauss–Newton point of f_+. Our first key result is to show that it is sufficient to check at most two of these cells, independent of the magnitude of n. An alternative characterization given in this paper shows that one cell and at most n rays in neighboring cells need to be examined to verify stationarity of θ (or give a descent direction).

(ii) The inherent difficulty in defining an algorithm to determine a Gauss–Newton point is that one must be sure that the limit point of the algorithm is stationary for θ in each piece of smoothness (orthant) containing that limit point. The second key idea, motivated by the characterizations above, is to apply variants of the projected gradient method [2] simultaneously to a single cell and n rays, to reduce θ. This means that the work performed by the projected gradient algorithm at each step of the Gauss–Newton method is comparable to performing just two projected gradient steps.

(iii) Our algorithm depends heavily on the projected gradient method having *Non-Stationary Repulsion* or NSR (see Section 3). Simply stated, if an algorithm has NSR, then each nonstationary point has a neighborhood that can be visited by at most one iterate of the algorithm. The third key result is that the projected gradient algorithm and the adaptations that we use in our algorithm have NSR. This property forces our algorithm to generate a better point in a neighboring orthant if the limit point of the sequence is not stationary in such an orthant.

The paper is organized as follows. In Section 2 we define the notion of a Gauss–Newton point for f_+ and and prove several equivalent characterizations (Proposition 2.3). We give a testable regularity condition (Definition 2.4) that guarantees that such Gauss–Newton points are solutions of (NE).

Section 3 outlines the nonstationary repulsion property and shows that any algorithm having NSR possesses strong global convergence properties (Theorem 3.2). We prove several technical results that are key to the convergence of our algorithms. A special case of these results are used to show that the projected gradient algorithm has NSR (Theorem 3.6).

Section 4 contains a description of three algorithms and their convergence properties. Our main convergence result, Theorem 4.3, proves the Gauss–Newton method we present is extremely robust: assuming only continuous differentiability of f, every limit point of the method is stationary for θ. No regularity assumptions on limit points are required. However, before proving this result, we outline a basic algorithm that can easily be shown to have NSR and hence global convergence under the same assumptions. Theorem 4.3 proves convergence of an extension to the basic algorithm that is motivated by the practical considerations of reducing the number of function and Jacobian evaluations. A Newton based hybrid method with global and local quadratic convergence is given in Subsection 4.3. Some simple examples of the use of these algorithms conclude the paper.

There have been many other research papers devoted to solving nonlinear complementarity problems. Some of the more recent papers are mentioned below.

There are several types of Newton methods for solving nonsmooth equations; see Subsection 4.3 for a brief introduction. Here we mention the following references on Newton methods for nonsmooth equations and extensions, [5, 4, 6, 12, 13, 15, 16, 20, 21, 26, 27, 28, 34]. A feature shared by "pure" Newton methods is the need for an invertible model function at the current iteration; applying the inverse of this model yields the next iterate. However singularities occur in many problems, for instance see [12], causing numerical difficulties for, or outright failure of these methods.

To circumvent the singularity problem, several Gauss–Newton techniques for solving nonlinear complementarity problems have been proposed. These can be found in the following references [1, 9, 19, 23, 22, 24]. Alternative techniques can be found in [8, 10, 14, 17, 18, 36].

Most of the notation in this paper is standard. We use R^n to denote the n–dimensional real vector space, $\langle\cdot,\cdot\rangle$ for the inner product of two elements in this space, $\|\cdot\|$ for the associated Euclidean norm, and B for the corresponding ball of vectors x such that $\|x\| \leq 1$. For a differentiable function $\Phi : \mathrm{R}^n \to \mathrm{R}^m$, $\nabla\Phi(x) \in \mathrm{R}^{m\times n}$ represents the Jacobian of Φ evaluated at x, and $\nabla\Phi(x)^T$ represents the transpose of this matrix. If Φ is only directionally differentiable, we denote the directional derivative mapping at x by $\Phi'(x;\cdot)$. Calligraphic upper case letters in general represent sets of indices, upper case letters represent sets or operators. If Ω is a convex set, the normal cone to Ω at a point $x \in \Omega$ is

$$N_\Omega(x) \stackrel{\text{def}}{=} \{y : \langle y, c - x\rangle \leq 0, \quad \forall c \in \Omega\}.$$

The tangent cone at $x \in \Omega$, is defined by $T_\Omega(x) \stackrel{\text{def}}{=} N_\Omega(x)^\circ$, where for a given convex cone K, the polar cone is defined by

$$K^\circ \stackrel{\text{def}}{=} \{y \colon \langle y, k\rangle \leq 0, \quad \forall k \in K\}\,.$$

Both the tangent and normal cones are empty at points $x \notin \Omega$. The Euclidean projection of x onto the set Ω is represented by $\pi_\Omega(x)$. A function $\Phi : \Omega \to \mathrm{R}^m$ is C^1 (continuously differentiable) if it is differentiable in the relative interior of Ω and, for each sequence $\{x^k\}$ in the relative interior of Ω that converges (to a general point of Ω), $\{\nabla\Phi(x^k)\}$ is also convergent. If Ω is a polyhedral convex set, and F is a face of Ω then $N_\Omega(x)$ is the same set for every x in the relative interior of F [32]. We call this set $N_\Omega(F)$. A facet of Ω is a face that has dimension 1 less than Ω. Further definitions from convex analysis can be found in [35]. We may abuse notation, when there is no possibility of confusion, by writing $\theta_{\mathcal{O}}$ instead of $\theta|_{\mathcal{O}}$ to mean the restriction of θ to an orthant $\mathcal{O}$ (to be distinguished from a normal map involving $\pi_{\mathcal{O}}$). Finally, throughout the paper the function $f : \Omega \to \mathrm{R}^m$ is assumed to be C^1; and usually $\Omega = \mathrm{R}^n_+$.

2 Gauss–Newton Points and Regularity

As we outlined in the introduction, a Gauss–Newton point for the smooth function f is a point $x^* \in \mathrm{R}^n$ such that $x = x^*$ minimizes the first-order model

$$\frac{1}{2}\left\|f(x^*) + \nabla f(x^*)(x - x^*)\right\|^2$$

of $\theta(x) = \frac{1}{2}\|f(x)\|^2$ over R^n. Equivalently, x^* is a stationary point of θ, that is $\nabla\theta(x^*) = \nabla f(x^*)^T f(x^*) = 0$. Note again that for the remainder of this paper we assume that f is continuously differentiable on its domain (Ω or R^n_+).

In the general case, we approximate the normal map $f_\Omega(x)$ by the piecewise linear model $f_\Omega(x^*) + f'_\Omega(x^*; x - x^*)$, where the directional derivative $f'_\Omega(x^*; \cdot)$ is a piecewise linear map. We can now define the notion of a Gauss–Newton point of f_Ω, which is based on this directional derivative.

Definition 2.1 *Let $x^* \in \mathrm{R}^n$. We say x^* is a* Gauss–Newton point *for f_Ω if $x = x^*$ solves the problem*

$$\min_x \frac{1}{2}\left\|f_\Omega(x^*) + f'_\Omega(x^*; x - x^*)\right\|^2. \tag{1}$$

Equivalently, x^* is a Gauss-Newton point if

$$\frac{1}{2}\left\|f_\Omega(x^*)\right\|^2 \;\leq\; \frac{1}{2}\left\|f_\Omega(x^*) + f'_\Omega(x^*; x - x^*)\right\|^2, \quad \forall x \in \mathrm{R}^n.$$

For the remainder of this paper we will consider only the special case of nonlinear complementarity problems where $\Omega \equiv \mathrm{R}^n_+$. However, many of the results have analogues in the general polyhedral case.

2.1 Gauss–Newton points of complementarity problems

Using Definition 2.1, we see that x^* is a Gauss–Newton point of f_+ if it solves (1) with $f_\Omega = f_+$. To understand this more fully, we now investigate the directional derivative f'_+ in more detail.

We can easily calculate the directional derivative of the function x_+ at x in the direction d: it is the vector $x'_+(d)$ in R^n whose ith component is given by

$$[x'_+(d)]_i = \begin{cases} d_i & \text{if } x_i > 0, \\ (d_i)_+ & \text{if } x_i = 0, \\ 0 & \text{if } x_i < 0. \end{cases}$$

In fact $x'_+(d)$ is exactly the projection of d onto the critical cone of R^n_+ at x, $K(x)$. This critical cone is the Cartesian product of n intervals in R, the ith interval being

$$K_i = \begin{cases} \mathrm{R} & \text{if } x_i > 0, \\ \mathrm{R}_+ & \text{if } x_i = 0, \\ \{0\} & \text{if } x_i < 0. \end{cases}$$

Since f is continuously differentiable, f_+ is directionally differentiable: for x, $d \in \mathrm{R}^n$,

$$f'_+(x; d) \; = \; \nabla f(x_+)\pi_K(d) + d - \pi_K(d),$$

where the notation $K = K(x)$ is used. As a function of d, the mapping on the right is exactly the normal map induced by the matrix $\nabla f(x_+)$ and the convex cone K, so

$$f'_+(x; d) \; = \; \nabla f(x_+)_K(d).$$

As mentioned above, the difficulty in determining whether a point x is a Gauss–Newton point is that we must examine potentially exponentially many pieces of smoothness of f_+, or pieces of linearity of $\nabla f(x_+)_K$. In fact, the number of pieces of linearity of $\nabla f(x_+)_K$ is the number of orthants containing x, and is given by 2^m where m is the number of components of x equal to zero. The next result removes this difficulty by showing that at most two pieces of linearity need to be considered.

We introduce some notation. Given an orthant $\mathcal{O}$, let $\mathcal{H}_i$ be the half-line $\pm\mathrm{R}_+$, $i = 1, \ldots, n$, such that

$$\mathcal{O} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_n.$$

The *complement* of $\mathcal{O}$ at a point $x \in \mathcal{O}$ is the orthant $\tilde{\mathcal{O}}$ given as the Cartesian product of half-lines $\tilde{\mathcal{H}}_i$ where

$$\tilde{\mathcal{H}}_i = \begin{cases} \mathcal{H}_i & \text{if } x_i \neq 0, \\ -\mathcal{H}_i & \text{if } x_i = 0. \end{cases}$$

It may seem odd that the complement of $\mathcal{O}$ at an interior point x is $\mathcal{O}$ itself. This is actually quite natural in the context of stationary points of θ because θ is differentiable at each interior point x of an orthant, hence the question of stationarity of θ at x is independent of other orthants.

We next introduce the formal definition of a stationary point.

Definition 2.2 *If θ is directionally differentiable and Ω is a nonempty convex set, then x^* is a* stationary point *for* $\min_{x \in \Omega} \theta(x)$ *if*

$$\theta'(x^*; d) \geq 0, \quad \forall d \in T_\Omega(x^*).$$

Note that if $\Omega \equiv \mathrm{R}^n$, then a stationary point satisfies $\theta'(x^*; d) \geq 0$, for all $d \in \mathrm{R}^n$.

Proposition 2.3 *Given $x^* \in \mathrm{R}^n$, let K be the critical cone to R^n_+ at x^*, $\mathcal{O}^*$ be any orthant containing x^* and $\tilde{\mathcal{O}}^*$ be the complement of $\mathcal{O}^*$ at x^*. Suppose f is continuously differentiable, then the function θ, defined by*

$$\theta(x) \stackrel{\text{def}}{=} \frac{1}{2} \|f_+(x)\|^2,$$

is directionally differentiable and

$$\theta'(x^*; d) = \left\langle f_+(x^*), f'_+(x^*; d) \right\rangle, \quad \forall d \in \mathrm{R}^n. \tag{2}$$

The following statements are equivalent:

1. *x^* is a Gauss–Newton point of f_+.*
2. *x^* is a stationary point of* $\min\{\theta(x) \colon x \in \mathrm{R}^n\}$.
3. $0 \in \nabla f(x^*_+)^T f_+(x^*) + K^\circ$ *and* $0 \in f_+(x^*) + K$.
4. *x^* is stationary for both* $\min\{\theta(x) \colon x \in \mathcal{O}^*\}$ *and* $\min\left\{\theta(x) \colon x \in \tilde{\mathcal{O}}^*\right\}$
5. *x^* is stationary for* $\min\{\theta(x) \colon x \in \mathcal{O}^*\}$ *and for each 1-dimensional problem*

$$\min\{\theta(x) \colon x \in x^* + N_{\mathcal{O}^*}(F)\},$$

where F is a facet of $\mathcal{O}^*$ containing x^*.

Proof If statement 1 holds, then we define

$$\gamma(x) \stackrel{\text{def}}{=} \frac{1}{2} \left\| f_+(x^*) + f'_+(x^*; x - x^*) \right\|^2,$$

and note that $\gamma'(x^*; h) = \theta'(x^*; h)$, for all h. Since x^* is a Gauss–Newton point, it follows that $\theta'(x^*; d) + o(d) \geq 0$, for all d, and hence that statement 2 holds by positive homogeneity. Conversely, if statement 2 holds, then for all d and $\mu > 0$, $0 \leq \left\langle f_+(x^*), f'_+(x^*; d) \right\rangle$, so that

$$\begin{aligned}
\gamma(x^* + \mu d) &= \frac{1}{2} \left\| f_+(x^*) + f'_+(x^*; \mu d) \right\|^2 \\
&= \frac{1}{2} \left\| f_+(x^*) + \mu f'_+(x^*; d) \right\|^2 \\
&\geq \frac{1}{2} \|f_+(x^*)\|^2 + \frac{1}{2}\mu^2 \left\| f'_+(x^*; d) \right\|^2 \\
&\geq \gamma(x^*).
\end{aligned}$$

Hence statement 1 holds.

Statement 2 means that $\left\langle f_+(x^*), \nabla f(x_+^*)_K(d)\right\rangle \geq 0$, for all $d \in \mathrm{R}^n$. If $d \in K$, then $\nabla f(x_+^*)_K(d) = \nabla f(x_+^*)d$ so that $\left\langle f_+(x^*), \nabla f(x_+^*)k\right\rangle \geq 0$, for all $k \in K$. Similarly, $\langle f_+(x^*), \nu\rangle \geq 0$, for all $\nu \in K^\circ$. This is exactly statement 3.

Conversely, let $d \in \mathrm{R}^n$, and recall from the Moreau decomposition that $d = k + \nu$ where $k = \pi_K(d)$ and $\nu = \pi_{K^\circ}(d)$. Using statement 3,

$$\left\langle f_+(x^*), \nabla f(x_+^*)_K(d)\right\rangle = \left\langle f_+(x^*), \nabla f(x_+^*)k\right\rangle + \langle f_+(x^*), \nu\rangle \geq 0.$$

Thus statement 2 holds.

Clearly statement 2 implies statement 4. Suppose statement 4 holds. Consider a facet F of $\mathcal{O}^*$ containing x^*. There is a unique index $1 \leq i \leq n$ such that neither e^i nor $-e^i$ lies in F, where e^i is the vector in R^n with component i equal to 1 and all other components equal to zero. Choose $s = \pm 1$ such that $se^i \notin \mathcal{O}^*$, then $N_{\mathcal{O}^*}(F) = \{\alpha se^i : \alpha \geq 0\}$. Note further that $se^i \in \tilde{\mathcal{O}}^*$. Thus statement 4 implies statement 5.

Suppose statement 5 holds and consider $e = \pm e^i$ for any index i. If $x_i \neq 0$, then stationarity of x^* for $\min\{\theta(x) : x \in \mathcal{O}^*\}$ yields that $\theta'(x^*; e) \geq 0$. If $x_i^* = 0$ then either $e \in \mathcal{O}^*$ or $e \in N_{\mathcal{O}^*}(F)$ for some facet F of $\mathcal{O}^*$ containing x^*. Therefore $\theta'(x^*; e) \geq 0$. It follows by linearity of $\theta'(x^*; \cdot)$ on each orthant, that $\theta'(x^*; d) \geq 0$ for each d in each orthant, hence for $d \in \mathrm{R}^n$. This is statement 2. □

The proof of the equivalence between statements 1, 2 and 3 in Proposition 2.3 can be immediately adapted to the case of a general polyhedral set Ω, with K then representing the critical cone to Ω at the point x.

2.2 Regularity

We now turn to the question of when a Gauss–Newton point for f_+ is a solution of $f_+(x) = 0$. This is commonly called regularity and we introduce a notion of regularity that is pertinent to our Gauss–Newton formulation.

Recall from Proposition 2.3 that x is a Gauss–Newton point if and only if

$$-f_+(x) \in K, \quad -\nabla f(x_+)^T f_+(x) \in K^\circ,$$

where K is the critical cone to R^n_+ at x. A simple regularity condition would be

$$-f_+(x) \in K, \quad -\nabla f(x_+)^T f_+(x) \in K^\circ \Longrightarrow f_+(x) = 0.$$

However, this condition is difficult to verify in most practical instances.

In order to generate a more testable notion of regularity, we follow the development of Moré [19]. Here, $f_+(x)$ is replaced by a general vector z and extra conditions that are satisfied by $f_+(x)$ are used to weaken the regularity assumption. Thus we define

$$\begin{aligned}\mathcal{P} &\stackrel{\text{def}}{=} \{i: x_i > 0, [f_+(x)]_i > 0\}, \\ \mathcal{N} &\stackrel{\text{def}}{=} \{i: x_i > 0, [f_+(x)]_i < 0\}, \\ \mathcal{C} &\stackrel{\text{def}}{=} \{i: [f_+(x)]_i = 0\},\end{aligned}$$

and we note that $[f_+(x)]_\mathcal{P} > 0$, $[f_+(x)]_\mathcal{N} < 0$ and $[f_+(x)]_\mathcal{C} = 0$.

Definition 2.4 *A point $x \in \mathrm{R}^n$ is said to be* regular *if the only z satisfying*

$$-z \in K, \quad -\nabla f(x_+)^T z \in K^\circ, \quad z_\mathcal{P} \geq 0, \quad z_\mathcal{N} \leq 0, \quad z_\mathcal{C} = 0$$

is $z = 0$.

This condition is closely related to [19, Definition 3.1]. This is because x is regular if and only if

$$z \neq 0, \quad -z \in K, \quad z_\mathcal{P} \geq 0, \quad z_\mathcal{N} \leq 0, \quad z_\mathcal{C} = 0 \Longrightarrow -\nabla f(x_+)^T z \notin K^\circ,$$

and the condition on the right is equivalent to existence of $p \in -K$ such that $z^T \nabla f(x_+) p > 0$. In contrast to [19, 22], the point x is not constrained to be nonnegative. Using Definition 2.4, we can prove the following result.

Lemma 2.5 *x is a regular stationary point for θ if and only if x solves (NE).*

Proof If $f_+(x) = 0$ then $\mathcal{C} = \{1, \ldots, n\}$ so $z = z_\mathcal{C} = 0$. Further, using (2), x^* is stationary for θ. Conversely, if x is stationary, then Proposition 2.3 shows that $z = f_+(x)$ satisfies all the relations required in the definition of regularity, and hence $f_+(x) = 0$. □

We turn to the question of testing whether a point x is regular. [19, 22, 28] give several conditions on the Jacobian of f to ensure that x is regular in the sense defined in the corresponding paper. For brevity we only discuss the *s-regularity* condition of Pang and Gabriel [22], and do not repeat definitions here. Moré [19] argues that s-regularity is stronger than his regularity condition; a similar comparison between Definition 2.4 and s-regularity can be made. Here we make a new observation about s-regularity. To explain this, recall that the goal of [22] is to solve $0 = \Theta(x) \stackrel{\text{def}}{=} (1/2)\|\min\{f(x), x\}\|^2$, where the min is taken component-wise; a solution of this equation solves (NCP) and vice versa. If $\bar{x}$ is nonstationary for Θ, then s-regularity of $\bar{x}$ ensures that for some direction $y \in \mathrm{R}^n$ and all x near $\bar{x}$, y is a (strict) descent direction for Θ at x, i.e. $\Theta'(x; y) < 0$; see [22, Lemmas 2, 6 and 7]. However Θ (like θ) is only piecewise smooth, and may have a nonstationary point which is a local minimum of some piece of smoothness of Θ, contradicting the existence of such a direction y. So s-regularity is too strong in the context of this investigation.

In what follows, we give conditions that ensure x to be regular in the sense of Definition 2.4. These results are proven by adapting arguments from Moré [19]. A key construct in the results is the matrix

$$J(x) \stackrel{\text{def}}{=} T^{-1}\nabla f(x_+)T^{-1}$$

where

$$T = \mathrm{diag}(t_i), \quad t_i = \begin{cases} 1 & \text{if } i \in \mathcal{P} \\ -1 & \text{if } i \notin \mathcal{P}. \end{cases}$$

T is chosen so that every component of $\tilde{z} \stackrel{\text{def}}{=} Tz$ is nonnegative. Under this transformation, x is regular if

$$0 \neq \tilde{z}_{\mathcal{D}} \geq 0, \quad \tilde{z}_{\mathcal{C}} = 0, \quad \tilde{z} \in -TK \Longrightarrow \exists \tilde{p} \in -TK, \quad \tilde{z}^T J(x)\tilde{p} > 0,$$

where $\mathcal{D} \stackrel{\text{def}}{=} \{i\colon [f_+(x)]_i \neq 0, x_i \geq 0\}$. Note that $\tilde{z}_i = 0$ when $i \notin \mathcal{D}$. The results we now give impose conditions of $J(x)$ to guarantee regularity. We note that $A \in \mathbb{R}^{n\times n}$ is an S-matrix if there is an $x > 0$ with $Ax > 0$, see [3].

Theorem 2.6 *Let $J(x) = T^{-1}\nabla f(x_+)T^{-1}$. If $[J(x)]_{\mathcal{E}\mathcal{E}}$ is an S-matrix for some index set $\mathcal{E}$ with $\mathcal{D} \subseteq \mathcal{E} \subseteq \{i\colon x_i \geq 0\}$, then x is regular.*

Proof Since $[J(x)]_{\mathcal{E}\mathcal{E}}$ is an S-matrix, there is some $\tilde{p}_{\mathcal{E}} > 0$ such that $[J(x)]_{\mathcal{E}\mathcal{E}}\tilde{p}_{\mathcal{E}} > 0$. Let $\tilde{p}$ be the vector in $\mathbb{R}^n$ obtained by setting other elements to zero, so that $[J(x)\tilde{p}]_{\mathcal{E}} > 0$. Now $0 \neq \tilde{z}_{\mathcal{D}} \geq 0$ and $\mathcal{D} \subseteq \mathcal{E}$ so $\tilde{z}^T J(x)\tilde{p} > 0$. Also, $-TK$ is the Cartesian product of

$$-(TK)_i = \begin{cases} \mathbb{R} & \text{if } x_i > 0 \\ \mathbb{R}_+ & \text{if } x_i = 0 \\ \{0\} & \text{if } x_i < 0 \end{cases}, \quad i = 1, \ldots, n. \tag{3}$$

Thus $\tilde{p} \in -TK$ and hence x is regular. □

A is a P-matrix if all its principal minors are positive. P-matrices are S-matrices [3, Corollary 3.3.5]. The following corollary is now immediate.

Corollary 2.7 *If $[\nabla f(x_+)]_{\mathcal{D}\mathcal{D}}$ is a P-matrix, then x is regular.*

Proof The hypotheses imply that $[J(x)]_{\mathcal{D}\mathcal{D}}$ is a P-matrix and hence an S-matrix. □

To complete our discussion of tests for regularity, we give the following result. Recall that if A is partitioned in the form

$$A = \begin{bmatrix} A_{\mathcal{N}\mathcal{N}} & A_{\mathcal{N}\mathcal{M}} \\ A_{\mathcal{M}\mathcal{N}} & A_{\mathcal{M}\mathcal{M}} \end{bmatrix},$$

and the matrix $A_{\mathcal{N}\mathcal{N}}$ is nonsingular, then $(A \backslash A_{\mathcal{N}\mathcal{N}}) \stackrel{\text{def}}{=} A_{\mathcal{M}\mathcal{M}} - A_{\mathcal{M}\mathcal{N}}A_{\mathcal{N}\mathcal{N}}^{-1}A_{\mathcal{N}\mathcal{M}}$ is called the Schur complement of $A_{\mathcal{N}\mathcal{N}}$ in A. The proof of the following result is modeled after [19, Corollary 4.6].

Theorem 2.8 *If $[\nabla f(x_+)]_{\mathcal{N}\mathcal{N}}$ is nonsingular and the Schur complement of $[\nabla f(x_+)]_{\mathcal{N}\mathcal{N}}$ in $[J(x)]_{\mathcal{D}\mathcal{D}}$ is an S-matrix, then x is regular.*

Proof Let $A = [J(x)]_{\mathcal{D}\mathcal{D}}$ and partition A into

$$\begin{bmatrix} A_{\mathcal{N}\mathcal{N}} & A_{\mathcal{N}\mathcal{M}} \\ A_{\mathcal{M}\mathcal{N}} & A_{\mathcal{M}\mathcal{M}} \end{bmatrix},$$

where $A_{\mathcal{N}\mathcal{N}} = [\nabla f(x_+)]_{\mathcal{N}\mathcal{N}}$ and $\mathcal{M} \stackrel{\text{def}}{=} \mathcal{D} \setminus \mathcal{N}$. We construct $\tilde{p}_{\mathcal{N}}$, $\tilde{p}_{\mathcal{M}}$ such that

$$[J(x)]_{\mathcal{D}\mathcal{D}} \begin{bmatrix} \tilde{p}_{\mathcal{N}} \\ \tilde{p}_{\mathcal{M}} \end{bmatrix} > 0. \tag{4}$$

Let $a > 0$, then $\tilde{p}_{\mathcal{N}}$, $\tilde{p}_{\mathcal{M}}$ solve

$$\begin{bmatrix} A_{\mathcal{N}\mathcal{N}} & A_{\mathcal{N}\mathcal{M}} \\ A_{\mathcal{M}\mathcal{N}} & A_{\mathcal{M}\mathcal{M}} \end{bmatrix} \begin{bmatrix} \tilde{p}_{\mathcal{N}} \\ \tilde{p}_{\mathcal{M}} \end{bmatrix} = \begin{bmatrix} a \\ q \end{bmatrix}$$

if and only if $\tilde{p}_{\mathcal{N}}$, $\tilde{p}_{\mathcal{M}}$ solve

$$\begin{bmatrix} A_{\mathcal{N}\mathcal{N}} & A_{\mathcal{N}\mathcal{M}} \\ 0 & (A \backslash A_{\mathcal{N}\mathcal{N}}) \end{bmatrix} \begin{bmatrix} \tilde{p}_{\mathcal{N}} \\ \tilde{p}_{\mathcal{M}} \end{bmatrix} = \begin{bmatrix} a \\ q - A_{\mathcal{M}\mathcal{N}} A_{\mathcal{N}\mathcal{N}}^{-1} a \end{bmatrix}.$$

Since $(A \backslash A_{\mathcal{N}\mathcal{N}})$ is an S-matrix by assumption, there exists $\tilde{p}_{\mathcal{M}} > 0$ with $(A \backslash A_{\mathcal{N}\mathcal{N}})\, \tilde{p}_{\mathcal{M}} > 0$. Multiplying $\tilde{p}_{\mathcal{M}}$ by an appropriately large number gives $(A \backslash A_{\mathcal{N}\mathcal{N}})\, \tilde{p}_{\mathcal{M}} + A_{\mathcal{M}\mathcal{N}} A_{\mathcal{N}\mathcal{N}}^{-1} a > 0$. It follows that $q \stackrel{\text{def}}{=} (A \backslash A_{\mathcal{N}\mathcal{N}})\, \tilde{p}_{\mathcal{M}} + A_{\mathcal{M}\mathcal{N}} A_{\mathcal{N}\mathcal{N}}^{-1} a > 0$, and taking $\tilde{p}_{\mathcal{N}} = A_{\mathcal{N}\mathcal{N}}^{-1}(a - A_{\mathcal{N}\mathcal{M}} \tilde{p}_{\mathcal{M}})$ implies (4).

Let $\tilde{p} \in \mathrm{R}^n$ be the vector constructed from $\tilde{p}_{\mathcal{M}}$ and $\tilde{p}_{\mathcal{N}}$ by adding appropriate zeros. Then it is easy to see that $\tilde{p} \in -TK$, see (3). Furthermore, $\tilde{z}^T J(x) \tilde{p} = \tilde{z}_{\mathcal{D}}^T [J(x)\tilde{p}]_{\mathcal{D}} > 0$. Hence x is regular. $\square$

Note that [5, 12, 28] all assume that

$$[\nabla f(x_+)]_{\mathcal{E}\mathcal{E}} \text{ is nonsingular and } ([\nabla f(x_+)]_{\mathcal{L}\mathcal{L}} \backslash [\nabla f(x_+)]_{\mathcal{E}\mathcal{E}}) \text{ is a } P\text{-matrix.} \tag{5}$$

Here $\mathcal{E} \stackrel{\text{def}}{=} \{i : x_i > 0\}$ contains $\mathcal{N}$ and $\mathcal{L} \stackrel{\text{def}}{=} \{i : x_i \geq 0\}$ contains $\mathcal{D}$. Theorem 2.8 requires the non-singularity of a smaller matrix and a weaker assumption on the Schur complement. However, (5) guarantees regularity in the sense of Definition 2.4 as we now show.

Lemma 2.9 *If (5) holds or, equivalently, the B-derivative $f'_+(x;\cdot)$ is invertible, then*

$$-z \in K, \quad -\nabla f(x_+)^T z \in K^\circ \Longrightarrow z = 0,$$

and so x is regular.

Proof The equivalence between (5) and the existence of a Lipschitz inverse of $f'_+(x;\cdot)$ is given by [28, Proposition 12]. Since all piecewise linear functions are Lipschitz, the claimed equivalence holds.

Suppose $-z \in K$ and $-\nabla f(x_+)^T z \in K^\circ$. It follows that $z_i = 0$, $i \notin \mathcal{L}$. Also $-\nabla f(x_+)^T z \in K^\circ$ implies

$$\left[\nabla f(x_+)^T z\right]_{\mathcal{E}} = 0, \quad \left[\nabla f(x_+)^T z\right]_{\mathcal{M}} \geq 0,$$

where $\mathcal{M} \stackrel{\text{def}}{=} \mathcal{L} \setminus \mathcal{E}$. Using the invertibility assumption from (5) and $-z \in K$ again, we see that

$$\left([\nabla f(x_+)^T]_{\mathcal{M}\mathcal{M}} - [\nabla f(x_+)^T]_{\mathcal{M}\mathcal{E}}[\nabla f(x_+)^T]^{-1}_{\mathcal{E}\mathcal{E}}[\nabla f(x_+)^T]_{\mathcal{E}\mathcal{M}}\right) z_{\mathcal{M}} \geq 0, \quad z_{\mathcal{M}} \leq 0.$$

The Schur complement is a P-matrix and hence $z = 0$ follows from [3, Theorem 3.3.4]. □

3 Nonstationary Repulsion (NSR) of the Projected Gradient Method

Let Ω be a nonempty closed convex set in R^n and $\phi : \Omega \to \mathrm{R}$ be C^1. (We are thinking of Ω being an orthant and $\phi = \theta|_\Omega$.) We paraphrase the description of the projected gradient (PG) algorithm given by Calamai and Moré [2] for the problem

$$\min_{x\in\Omega} \phi(x). \tag{6}$$

For any $\alpha > 0$, the first-order necessary condition for x to be a local minimizer of this problem is that

$$\pi_\Omega(x - \alpha\nabla\phi(x)) = x.$$

When $x^k \in \Omega$ is nonstationary, a step length $\alpha_k > 0$ is chosen by searching the path

$$x^k(\alpha) \stackrel{\text{def}}{=} \pi_\Omega(x^k - \alpha\nabla\phi(x^k)), \quad \alpha > 0.$$

Given constants $\gamma_1 > 0$, $\gamma_2 \in (0,1)$, and μ_1 and μ_2 with $0 < \mu_1 \leq \mu_2 < 1$, the step length α_k must satisfy

$$\phi(x^k(\alpha_k)) \leq \phi(x^k) + \mu_1 \left\langle \nabla\phi(x^k), x^k(\alpha_k) - x^k \right\rangle \tag{7}$$

and

$$\alpha_k \geq \gamma_1 \quad \text{or} \quad \alpha_k \geq \gamma_2\bar{\alpha}_k > 0, \tag{8}$$

where $\bar{\alpha}_k$ satisfies

$$\phi(x^k(\bar{\alpha}_k)) > \phi(x^k) + \mu_2 \left\langle \nabla\phi(x^k), x^k(\bar{\alpha}_k) - x^k \right\rangle. \tag{9}$$

Condition (7) forces α_k not to be too large; it is the analogue of the condition used in the standard Armijo line search for unconstrained optimization. Condition (8) forces α_k not to be too small; in the case that $\alpha_k < \gamma_1$, this requirement is the analogue of the standard Wolfe-Goldstein [7] condition from unconstrained optimization.

The PG method is a *feasible point* algorithm in that it requires a starting point x^0 in Ω and produces a sequence of iterates $\{x^k\} \subset \Omega$. It is also *monotonic*, that is, if $x^k \in \Omega$ is nonstationary, then $\phi(x^{k+1}) \leq \phi(x^k)$. We claim that the PG method has NSR:

Definition 3.1 *An iterative feasible point algorithm for (6) has* nonstationary repulsion *(NSR) if for each nonstationary $\bar{x} \in \Omega$, there exists a neighborhood V of $\bar{x}$ such that if any iterate x^k lies in $V \cap \Omega$, then $\phi(x^{k+1}) < \phi(\bar{x})$.*

The fact that the steepest descent method, i.e. the PG method when $\Omega = \mathrm{R}^n$, has NSR is easy to see. Also, Polak [25, Chapter 1] discusses a general descent property that is similar to NSR and provides convergence results like Theorem 3.2 below.

It is trivial but important that NSR yields strong global convergence properties:

Theorem 3.2 *Suppose $\mathcal{A}$ is a monotonic feasible point algorithm for (6) with NSR. Let $x^0 \in \Omega$.*

1. *Any limit point of the sequence generated by $\mathcal{A}$ is stationary.*
2. *Let $\mathcal{B}$ be any monotonic feasible point algorithm for (6). Suppose $\{x^k\}$ is a sequence defined by applying either $\mathcal{A}$ or $\mathcal{B}$ to each x^k. Then any limit point of $\{x^k\}_{k\in\mathcal{K}}$ is stationary if $\mathcal{A}$ is applied infinitely many times, where $\mathcal{K} = \left\{k : x^{k+1} \text{ is generated by } \mathcal{A}\right\}$.*

Proof

1. This is a corollary of part 2 of the theorem.
2. Let $\bar{x} \in \Omega$ be nonstationary for (6) and $\mathcal{K}$ have infinite cardinality. NSR gives $\epsilon > 0$ such that $\phi(x^{k+1}) < \phi(\bar{x})$ if $x^k \in (\bar{x} + \epsilon\mathrm{B}) \cap \Omega$. If the subsequence $\{x^k\}_\mathcal{K}$ does not intersect $(\bar{x} + \epsilon\mathrm{B}) \cap \Omega$ then $\bar{x}$ is not a limit point of this subsequence. So we assume that $x^k \in (\bar{x} + \epsilon\mathrm{B}) \cap \Omega$ for some $k \in \mathcal{K}$, hence $\phi(x^{k+1}) < \phi(\bar{x})$.

   By continuity of ϕ there is $\epsilon_1 \in (0, \epsilon)$ such that $\phi(x) > \phi(x^{k+1})$ if $x \in (\bar{x} + \epsilon_1\mathrm{B}) \cap \Omega$. By monotonicity of $\mathcal{A}$ and $\mathcal{B}$, $\phi(x^{k+j}) \leq \phi(x^{k+1})$ for each $j \geq 1$; hence $\bar{x}$ is not a limit point of $\{x^{k+j}\}_{j\geq 1}$, or of $\{x^k\}_\mathcal{K}$.

□

Of course NSR is not a guarantee of convergence. To guarantee existence of a limit point of a sequence produced by a method with NSR we need additional knowledge, for instance boundedness of the lower level set

$$\left\{x : \phi(x) \leq \phi(x^0)\right\}$$

where x^0 is the starting iterate.

To prove that the PG method has NSR we need to establish that the rate of descent obtained along the path $x(\alpha) = \pi_\Omega(x - \alpha\nabla\phi(x))$ is uniform for feasible x in a neighborhood of a given nonstationary $\bar{x} \in \Omega$. The lemma below states a uniform descent property for all small perturbations ϕ about a given function $\bar{\phi}$; the reader may consider $\phi = \bar{\phi}$ for simplicity. In the case where many functions ϕ are present we use the notation

$$x_\phi(\alpha) \stackrel{\text{def}}{=} \pi_\Omega(x - \alpha\nabla\phi(x)).$$

Definition 3.3

1. *Let $\bar{x} \in \mathrm{R}^n$ and $\nu > 0$. If $\phi : \Omega \to \mathrm{R}$ is C^1, the* modulus of continuity *of $\nabla\phi$ at $\bar{x} \in \Omega$ is the function of ϕ and $\beta > 0$ (and $\bar{x}$, ν)*

$$\omega(\phi, \beta) \stackrel{\text{def}}{=} \sup\left\{\|\nabla\phi(y) - \nabla\phi(x)\| : x, y \in \Omega,\ \|x - \bar{x}\| \le \nu,\ \|x - y\| \le \beta\right\}.$$

2. *Let $\bar{x} \in \mathrm{R}^n$, $\nu > 0$, and $\bar{\phi} : \mathrm{R}^n \to \mathrm{R}$ be C^1. Given $\epsilon > 0$, let $U(\epsilon) = U(\epsilon, \bar{\phi}, \bar{x}, \nu)$ be the set of all C^1 functions $\phi : \Omega \to \mathrm{R}^n$ such that*

$$\sup\left\{|\phi(x) - \bar{\phi}(x)| \ + \ \left\|\nabla\phi(x) - \nabla\bar{\phi}(x)\right\| : x \in (\bar{x} + \nu\mathrm{B}) \cap \Omega\right\} \ < \ \epsilon, \tag{10}$$

and

$$\omega(\phi, \beta) \ \le \ (1 + \epsilon)\omega(\bar{\phi}, \beta), \quad \forall \beta \in (0, \nu). \tag{11}$$

Lemma 3.4 *Let $\bar{\phi} : \mathrm{R}^n \to \mathrm{R}$ be C^1, $\nu > 0$ and $\bar{x} \in \Omega$ be nonstationary for $\min\left\{\bar{\phi}(x) \colon x \in \Omega\right\}$. There exist positive constants ϵ and κ such that for each $\phi \in U(\epsilon) = U(\epsilon, \bar{\phi}, \bar{x}, \nu)$, $x \in (\bar{x} + \epsilon\mathrm{B}) \cap \Omega$, and $\alpha \ge 0$,*

$$\langle\nabla\phi(x), x_\phi(\alpha) - x\rangle \le -\min\{\alpha, \epsilon\}\kappa. \tag{12}$$

Proof Let $\phi : \Omega \to \mathrm{R}$ be C^1, $x \in \Omega$ and $\nu > 0$. According to [2, (2.4)],

$$\langle\nabla\phi(x), x_\phi(\alpha) - x\rangle \le -\|x_\phi(\alpha) - x\|^2/\alpha, \quad \forall x \in \Omega,\ \alpha > 0.$$

Moreover, [2, Lemma 2.2] says that, as a function of $\alpha > 0$, $\|x_\phi(\alpha) - x\|/\alpha$ is antitone (nonincreasing); in particular for any $\bar{\alpha} > 0$,

$$\|x_\phi(\alpha) - x\|/\alpha \ge \|x_\phi(\bar{\alpha}) - x\|/\bar{\alpha}, \quad \forall \alpha \in (0, \bar{\alpha}).$$

Using this with the previous inequality, we deduce for any $\bar{\alpha} > 0$ that

$$\begin{aligned}\langle\nabla\phi(x), x_\phi(\alpha) - x\rangle &\le -\alpha(\|x_\phi(\alpha) - x\|/\alpha)^2, \quad \forall x \in \Omega,\ \alpha > 0\\ &\le -\alpha(\|x_\phi(\bar{\alpha}) - x\|/\bar{\alpha})^2, \quad \forall x \in \Omega,\ 0 < \alpha < \bar{\alpha}.\end{aligned} \tag{13}$$

Fix $\bar\alpha > 0$. By hypothesis, the point $\bar x$ is such that $\left\| \bar x_{\bar\phi}(\bar\alpha) - \bar x \right\| > 0$. Also by (10), if $x \to \bar x$ and $\phi \to \bar\phi$, where convergence of ϕ means that $\phi \in U(\epsilon)$ and $\epsilon \downarrow 0$, then $x_\phi(\bar\alpha)$ converges to $\bar x_{\bar\phi}(\bar\alpha)$. Hence there are $\bar\epsilon > 0$, $\kappa > 0$ such that for $\phi \in U(\bar\epsilon)$, $x \in \bar x + \bar\epsilon \mathrm{B}$,

$$\|x_\phi(\bar\alpha) - x\| \geq \sqrt{\kappa}\bar\alpha.$$

Together with (13), this yields

$$\langle \nabla\phi(x), x_\phi(\alpha) - x\rangle \leq -\alpha\kappa, \quad \forall x \in (\bar x + \bar\epsilon \mathrm{B}) \cap \Omega,\ \alpha \in [0, \bar\alpha].$$

Let $\epsilon \stackrel{\text{def}}{=} \min\{\bar\epsilon, \bar\alpha\}$, then (12) holds for $0 \leq \alpha \leq \epsilon$. Using the well known antitone property of $\langle \nabla\phi(x), x_\phi(\alpha) - x\rangle$ in $\alpha \geq 0$, see [2, (2.6)], we see that (12) also holds for $\alpha > \epsilon$.

□

The following result gives some technical properties of the PG method that will be important for our main algorithm. We use it later to prove that the PG method has NSR, though in this case NSR follows from the simpler case in which ϕ is a fixed function.

Proposition 3.5 *Let $\bar\phi : \mathrm{R}^n \to \mathrm{R}$ be C^1, $\nu > 0$ and $\bar x \in \Omega$ be nonstationary for $\min\left\{\bar\phi(x) : x \in \Omega\right\}$. Then there is a positive constant ϵ such that for each $x \in (\bar x + \epsilon \mathrm{B}) \cap \Omega$ and $\phi \in U(\epsilon) = U(\epsilon, \bar\phi, \bar x, \nu)$:*

1. *For each $\alpha \in [0, \epsilon]$,*

$$\phi(x_\phi(\alpha)) \ \leq \ \phi(x) + \mu_1 \langle \nabla\phi(x), x_\phi(\alpha) - x\rangle .$$

2. *One step of PG on (6) from x generates $x_\phi(\alpha)$ with $\alpha \geq \epsilon$.*

Proof Suppose $\bar\phi$, $\bar x$ and ν are as stated. Let $\gamma_1 > 0$, $\gamma_2 \in (0,1)$ and $0 < \mu_1 \leq \mu_2 < 1$ be the constants of the PG method. Let ϵ_1, κ be given by Lemma 3.4, and $\phi \in U(\epsilon_1)$; and assume without loss of generality that $\epsilon_1 \in (0, \nu]$, i.e. (10) and (11) hold with $\epsilon = \nu = \epsilon_1$.

We estimate the error term

$$\varepsilon(\phi, x, y) \stackrel{\text{def}}{=} \phi(y) - \phi(x) - \langle \nabla\phi(x), y - x\rangle ,$$

where $y, x \in \Omega$. By choice of $\phi \in U(\epsilon_1)$, specifically (11) with $\epsilon = \nu = \epsilon_1$, for each $\beta \in (0, \epsilon_1)$, $x \in (\bar x + \epsilon_1 \mathrm{B}) \cap \Omega$ and $y \in (x + \beta \mathrm{B}) \cap \Omega$,

$$\|\nabla\phi(x) - \nabla\phi(y)\| \ \leq \ (1 + \epsilon_1)\omega(\bar\phi, \beta).$$

Thus, for $x \in (\bar{x} + \epsilon_1 \mathrm{B}) \cap \Omega$ and $y \in (x + \epsilon_1 \mathrm{B}) \cap \Omega$,

$$
\begin{aligned}
|\varepsilon(\phi, y, x)| &= |\int_0^1 \langle \nabla\phi(x + t(y-x)) - \nabla\phi(x), y - x\rangle \, dt| \\
&\le (1+\epsilon_1)\omega(\bar{\phi}, \|x-y\|)\frac{\|y-x\|}{2}. \qquad (14)
\end{aligned}
$$

By continuity of $\nabla\bar{\phi}$ on the compact set $(\bar{x} + \epsilon_1 \mathrm{B}) \cap \Omega$, there is a finite upper bound η on $\left\|\nabla\bar{\phi}(x)\right\|$ for $x \in (\bar{x} + \epsilon_1 \mathrm{B}) \cap \Omega$. Define $\bar{\eta} = \eta + \epsilon_1$; by choice of $\phi \in U(\epsilon_1)$, specifically (10) with $\epsilon = \nu = \epsilon_1$, $\|\nabla\phi(x)\| \le \bar{\eta}$ for $x \in (\bar{x} + \epsilon_1 \mathrm{B}) \cap \Omega$. It follows for such x and any $\alpha \ge 0$, that

$$\|x_\phi(\alpha) - x\| \le \alpha\bar{\eta}, \qquad (15)$$

because π_Ω is Lipschitz of modulus 1.

Furthermore, since $\nabla\bar{\phi}$ is uniformly continuous on compact sets, $\omega(\bar{\phi}, \beta) \downarrow 0$ as $\beta \downarrow 0$. Thus, using the fact that $\omega(\bar{\phi}, \cdot)$ is nondecreasing, there exists $\epsilon_2 \in (0, \epsilon_1)$ such that for $x \in (\bar{x} + \epsilon_1 \mathrm{B}) \cap \Omega$ and $\alpha \in (0, \epsilon_2)$, both $\alpha\bar{\eta} \le \epsilon_1$ and

$$(1+\epsilon_1)\omega(\bar{\phi}, \alpha\bar{\eta})\frac{\bar{\eta}}{2} \le \kappa(1-\mu_2).$$

From these inequalities and the inequalities (14) and (15), we see that for $x \in (\bar{x} + \epsilon_1 \mathrm{B}) \cap \Omega$ and $\alpha \in (0, \epsilon_2)$,

$$\varepsilon(\phi, x_\phi(\alpha), x) \;\le\; \alpha\kappa(1-\mu_2). \qquad (16)$$

Now for such x and α,

$$
\begin{aligned}
\phi(x_\phi(\alpha)) &= \phi(x) \;+\; [\mu_2 + (1-\mu_2)]\,\langle \nabla\phi(x), x_\phi(\alpha) - x\rangle \;+\; \varepsilon(\phi, x_\phi(\alpha), x) \\
&\le \phi(x) \;+\; \mu_2\,\langle \nabla\phi(x), x_\phi(\alpha) - x\rangle \;-\; \alpha(1-\mu_2)\kappa \;+\; \alpha(1-\mu_2)\kappa
\end{aligned}
$$

where the second inequality relies on the uniform descent property of Lemma 3.4 and (16). Thus

$$\phi(x_\phi(\alpha)) \le \phi(x) + \mu_2\,\langle \nabla\phi(x), x_\phi(\alpha) - x\rangle\,, \quad \forall x \in (\bar{x} + \epsilon_1 \mathrm{B}) \cap \Omega,\; \alpha \in (0, \epsilon_2),$$

and this inequality with μ_1 replacing μ_2 also holds. Finally, for any $x^k \in (\bar{x} + \epsilon_1 \mathrm{B}) \cap \Omega$, the auxiliary scalar $\bar{\alpha}_k$ satisfying (9) is bounded below by ϵ_2; hence the step size α_k is bounded below by $\epsilon \stackrel{\text{def}}{=} \min\{\gamma_1, \gamma_2\epsilon_2\}$. Since $0 < \gamma_2 < 1$, then $\epsilon < \epsilon_2 < \epsilon_1$, parts 1 and 2 of the proposition hold.

□

Theorem 3.6 *The PG method applied to (6) has NSR.*

Proof Let $\bar{x} \in \mathrm{R}^n$ be nonstationary, so according to Proposition 3.5 and Lemma 3.4, if $x^k \in (\bar{x} + \epsilon\mathrm{B}) \cap \Omega$ then $\alpha_k \geq \epsilon$ and

$$\begin{aligned} \phi(x^{k+1}) &\leq \phi(x^k) + \mu_1 \left\langle \nabla\phi(x^k), x^k(\alpha_k) - x^k \right\rangle \\ &\leq \phi(x^k) - 2\delta, \end{aligned} \tag{17}$$

where $\delta = \mu_1 \epsilon \kappa / 2 > 0$. Now by continuity of ϕ there is $\bar{\epsilon} \in (0, \epsilon)$ such that

$$|\phi(x) - \phi(\bar{x})| \leq \delta, \quad \forall x \in (\bar{x} + \bar{\epsilon}\mathrm{B}) \cap \Omega.$$

Take $V \stackrel{\text{def}}{=} (\bar{x} + \bar{\epsilon}\mathrm{B}) \cap \Omega$, and use the above inequality with (17) to see that for any $x^k \in V$,

$$\phi(x^{k+1}) \ \leq \ \phi(\bar{x}) - \delta.$$

The NSR property of Definition 3.1 follows. $\square$

4 Projected Gradient Algorithms for NCP

Our main goal here is to present a method for minimizing θ that has a low computational cost, and has NSR. Before proceeding we will make a few comments on guaranteeing convergence, at least on a subsequence.

Existence of a (stationary) limit point of a sequence produced by a method with NSR follows from boundedness of the lower level set

$$\left\{ x \in \mathrm{R}^n : \|f_+(x)\| \leq \left\| f_+(x^0) \right\| \right\},$$

where x^0 is the initial point. This boundedness property holds in many cases, for instance if f is a uniform P-function, see Harker and Xiao [12]; hence if f is strongly monotone. However the uniform P-function property implies that $f'_+(x; \cdot)$ is invertible for each x, a condition that we believe is too strong in general (c.f. Lemma 2.9). A weaker condition yielding boundedness of the above level set is that f_+ is proper, namely that the inverse image $f_+^{-1}(S)$ of any compact set $S \subset \mathrm{R}^n$ is compact.

4.1 A simple globally convergent algorithm

Given statement 4 of Proposition 2.3, it is tempting to use the following steepest descent idea in algorithms for minimizing θ. Given the kth iterate $x^k \in \mathrm{R}^n$, an orthant $\mathcal{O}^k$ containing x^k and the complement $\tilde{\mathcal{O}}^k$ of $\mathcal{O}^k$ at x^k, let d^k solve

$$\min_d \theta'(x^k; d) \quad \text{subject to} \quad \|d\| \leq 1, \ d \in \mathcal{O}^k \cup \tilde{\mathcal{O}}^k.$$

This essentially requires two n-dimensional convex quadratic programs to be solved (a polyhedral norm on d may be used), one for each orthant. If $d = 0$ is a solution, then x^k is stationary for θ. Otherwise $\theta'(x^k; d^k) < 0$, and we can perform a line search to establish $\alpha_k > 0$ such that for $x^{k+1} = x^k + \alpha_k d^k$, $\theta(x^k + \alpha_k d^k)$ is strictly less than $\theta(x^k)$.

However if θ is nonsmooth there seems to be little global convergence theory for algorithms based on this idea. For instance, it is not known if the step length α_k can be chosen to be uniformly large in a neighborhood of a nonstationary point, while still retaining a certain rate of descent; hence it is hard to show that the sequence produced will not accumulate at a nonstationary point. Pang, Han and Rangaraj [23, Corollary 1] give an additional smoothness assumption at a limit point that is required to prove stationarity.

Alternatively given the stationarity characterization of Proposition 2.3.5, we can design a naive steepest descent algorithm for minimizing θ, each iteration of which is based on a projected gradient step over an orthant $\mathcal{O}^k$ containing the current iterate x^k, and an additional m projected gradient steps on 1-dimensional problems corresponding to moving in directions normal to the m facets of $\mathcal{O}^k$ that contain x^k (so m is the number of zero components of x^k). It is significant that to obtain global convergence, we only need to increase the number of 1-dimensional subproblems at each iteration from m to n, i.e. normals to all facets of $\mathcal{O}^k$ must be examined.

The algorithm below introduces notation not strictly required for its statement; this notation is presented in preparation for the main algorithm, Algorithm 2, which appears in the next subsection. By $\theta_{\mathcal{O}}$ we mean the restriction $\theta|_{\mathcal{O}}$ of θ to $\mathcal{O}$.

Algorithm 1. Let $x^0 \in \mathrm{R}^n$. Given $k \in \{0, 1, 2, \ldots,\}$ and $x^k \in \mathrm{R}^n$, define x^{k+1} as follows.

Choose any orthant $\mathcal{O}^k$ containing x^k, let $y^0(\alpha) \stackrel{\mathrm{def}}{=} \pi_{\mathcal{O}^k}[x^k - \alpha \nabla \theta_{\mathcal{O}^k}(x^k)]$, and α_0 be the step size determined by one step of the projected gradient algorithm applied to $\min\left\{\theta(x) \colon x \in \mathcal{O}^k\right\}$ from x^k. Suppose $F_1, \ldots, F_n$ are the facets of $\mathcal{O}^k$. For $j = 1, \ldots, n$, let $y^j \stackrel{\mathrm{def}}{=} \pi_{F_j}(x^k)$, $N_j \stackrel{\mathrm{def}}{=} N_{\mathcal{O}^k}(F_j)$, $y^j(\alpha) \stackrel{\mathrm{def}}{=} y^j + \alpha \pi_{N_j}[-\nabla\theta(y^j)]$, and α_j be the step size determined by one step of the projected gradient algorithm applied to

$$\min\left\{\theta(x) \colon x \in y^j + N_j\right\}, \tag{18}$$

starting from y^j. Let

$$x^{k+1} \stackrel{\mathrm{def}}{=} y^{\hat{\jmath}}(\alpha_{\hat{\jmath}}), \quad \text{where} \quad \hat{\jmath} \in \operatorname{argmin}\left\{\theta(y^j(\alpha_j)) \colon j = 0, 1, 2, \ldots, n\right\}.$$

If $\theta(x^{k+1}) = \theta(x^k)$ then STOP; x^k is a Gauss–Newton point of f_+.

Remark. In Algorithm 1 the projected gradient method is used as a subroutine. Therefore we assume that if the starting point of a subproblem is stationary, then the projected gradient method merely returns this point; the decision of whether or not the main algorithm should continue is made elsewhere.

Theorem 4.1 *Algorithm 1 is well defined and has NSR.*

Proof Since the projected gradient method is well defined, for each k and x^k the algorithm produces x^{k+1}. If $\theta(x^{k+1}) = \theta(x^k)$ then none of the subproblems of the form (18) produced a point with a lower function value than $\theta(x^k)$. So x^k is stationary for each subproblem for which F_j is a facet of $\mathcal{O}^k$ containing x^k, and by Proposition 2.3, x^k is also a Gauss-Newton point of f_+. Thus Algorithm 1 is well defined.

We show that the algorithm has NSR. Suppose $\bar{x}$ is not a Gauss-Newton point of f_+. For x^k sufficiently close to $\bar{x}$, $\bar{x} \in \mathcal{O}^k$. So consider the case when $\mathcal{O}^k = \bar{\mathcal{O}}$ for some fixed orthant $\bar{\mathcal{O}}$ containing $\bar{x}$. By Proposition 2.3 $\bar{x}$ is nonstationary either for $\min\left\{\theta(x) : x \in \bar{\mathcal{O}}\right\}$ or for $\min\{\theta(x) : x \in \bar{x} + N_{\bar{\mathcal{O}}}(F)\}$, where F is some facet of $\bar{\mathcal{O}}$ containing $\bar{x}$.

In the former case, for some $\epsilon = \epsilon(\bar{\mathcal{O}}) > 0$ and each $x^k \in \bar{x} + \epsilon\mathrm{B}$, we have from Theorem 3.6 with $\phi = \theta_{\bar{\mathcal{O}}}$ and $\Omega = \bar{\mathcal{O}}$, that the candidate $y^0(\alpha_0)$ for the next iterate x^{k+1} yields $\theta(y^0(\alpha_0)) < \theta(\bar{x})$. Hence our choice of x^{k+1} also yields $\theta(x^{k+1}) < \theta(\bar{x})$. In the latter case, we can apply Proposition 3.5 by reformulating the subproblem (18) as $\min\{\theta(y^j + d) : d \in N_{\bar{\mathcal{O}}}(F)\}$, i.e. define $\phi(d) = \theta(y^j + d)$, $\bar{\phi}(d) = \theta(\bar{x} + d)$, $\Omega = N_{\bar{\mathcal{O}}}(F)$, and ν as any positive constant, and let $\epsilon_1 > 0$ be the constant given by Proposition 3.5. Given the simple form of ϕ, it is easy to check that there is $\epsilon = \epsilon(\bar{\mathcal{O}}) > 0$ such that if $x^k \in \bar{x} + \epsilon\mathrm{B}$, then $\phi \in U(\epsilon_1, \bar{\phi}, \bar{x}, \nu)$. For such x^k, Proposition 3.5 says that the candidate iterate $y^j(\alpha_j)$ yields $\theta(y^j(\alpha_j)) < \theta(\bar{x})$, hence $\theta(x^{k+1}) < \theta(\bar{x})$.

Since there are only finitely many orthants, we conclude that for some $\epsilon > 0$ independent of $\mathcal{O}^k$, and each $x^k \in \bar{x} + \epsilon\mathrm{B}$, we have $\theta(x^{k+1}) < \theta(\bar{x})$. □

This algorithm is extremely robust: under the single assumption that f is C^1 on R^n_+, the method is well defined and accumulation points are always Gauss–Newton points. It is also reasonably simple, using the projected gradient method as the work horse.

A serious drawback of Algorithm 1 is that we need at least $n+1$ function and Jacobian evaluations per iteration, in order to carry out the projected gradient method on the $n+1$ subproblems. By contrast, the use of 1-dimensional subproblems means the linear algebra performed by Algorithm 1 is only around twice as expensive as the linear algebra needed to perform one projected gradient step on an orthant.

4.2 An efficient globally convergent algorithm

We present a globally convergent method for finding Gauss–Newton points of f_+ based on the PG method. It is efficient in the sense that per iteration, the number of function evaluations is comparable to that needed for the PG method applied to minimizing a smooth function over an orthant, and the linear algebra computation involves about double the work required for linear algebra in the PG method.

At each iteration, we approximate θ by linearizing f about x_+^k. Let

$$A^k(x) \stackrel{\text{def}}{=} \frac{1}{2}\left\|L_+^k(x)\right\|^2$$
$$\text{where} \quad L_+^k(x) \stackrel{\text{def}}{=} f(x_+^k) + \nabla f(x_+^k)(x_+ - x_+^k) + x - x_+. \tag{19}$$

The "linearization" L_+^k is a local *point-based approximation* [34] when ∇f is locally Lipschitz, and more generally a *uniform first-order approximation* near x^k [28]; such approximations are more powerful than directional derivatives in that they approximate f_+ uniformly well for all x near x^k. In [5, 4, 28, 34] these approximation properties have been exploited to give strong convergence results for Newton methods applied to nonsmooth equations like $f_+(x) = 0$. Our main algorithm, below, and its extremely robust convergence behavior also rely on these approximation properties.

Lemma 4.2 *Let $\bar{x} \in \mathrm{R}^n$ and $\epsilon > 0$. There is a non-decreasing function $\varepsilon : \mathrm{R}_+ \to \mathrm{R}_+$ such that $\varepsilon(\beta) = o(\beta)$ as $\beta \downarrow 0$, and for each $x^k, x \in \bar{x} + \epsilon\mathrm{B}$,*

$$\left\|\theta(x) - A^k(x)\right\| \quad \le \quad \varepsilon(\left\|x - x^k\right\|).$$

Proof We have

$$|\theta(x) - A^k(x)| = (1/2)\left\langle f_+(x) + L_+^k(x), f_+(x) - L_+^k(x)\right\rangle$$
$$\le c\left\|f_+(x) - L_+^k(x)\right\|,$$

where $c \in (0, \infty)$ is the maximum value of $(1/2)\left\|f_+(x) + L_+^k(x)\right\|$ for $x^k, x \in \bar{x} + \epsilon\mathrm{B}$. Let ω be the modulus of continuity of ∇f on $\mathrm{R}_+^n \cap (\bar{x}_+ + \epsilon\mathrm{B})$ (see Definition 3.3). Similar to (14) in the proof of Proposition 3.5,

$$\left\|f_+(x) - L_+^k(x)\right\| \le \omega(\left\|x - x^k\right\|)\left\|x - x^k\right\|/2,$$

where $\omega(\beta) \to 0$ as $\beta \downarrow 0$. Take $\varepsilon(\beta) \stackrel{\text{def}}{=} c\omega(\beta)\beta/2$. □

We will search several paths during an iteration but, unlike Algorithm 1, our criteria for choosing the path parameter α will use derivatives of the approximation A^k rather than of θ. Let $\mu_0 \in (0, \mu_1)$ and $\sigma \in (0, 1)$. Suppose we are also given an orthant $\mathcal{O}$ containing a point y (but not necessarily x^k), and a path $y : [0, \infty) \to \mathrm{R}^n$ with $y(0) = y$. Given $\alpha > 0$, $y(\alpha)$ is a candidate for x^{k+1} if

$$\theta(y(\alpha)) \le A^k(y) + \mu_0\left\langle \nabla A_{\mathcal{O}}^k(y), y(\alpha) - y\right\rangle. \tag{20}$$

Here $A_{\mathcal{O}}^k$ is the restriction $A^k|_{\mathcal{O}}$. If α fails the above test, we can try $\alpha = \sigma\alpha$. Note that if $y = x^k$ then $\nabla A_{\mathcal{O}}^k(y) = \nabla\theta_{\mathcal{O}}(x^k)$, and the obvious choice for $y(\alpha)$ is $\pi_{\mathcal{O}}(x^k - \alpha\nabla\theta_{\mathcal{O}}(x^k))$. In this case (20) is equivalent to (7) with $\phi = \theta_{\mathcal{O}}$ and $\Omega = \mathcal{O}$.

The first part of Algorithm 2 is a single step of Algorithm 1 applied to A^k instead of θ. The second part determines the path and the corresponding step length that will define the next iterate x^{k+1}.

Algorithm 2. Let $x^0 \in \mathrm{R}^n$ and (in addition to the constants used for the PG method), $\mu_0 \in (0, \mu_1)$, $\sigma \in (0,1)$. Given $k \in \{0,1,2,\ldots,\}$ and $x^k \in \mathrm{R}^n$, define x^{k+1} as follows.

Part I. Choose any orthant $\mathcal{O}^k$ containing x^k, let $y^0 \stackrel{\text{def}}{=} x^k$,

$$y^0(\alpha) \stackrel{\text{def}}{=} \pi_{\mathcal{O}^k}[x^k - \alpha \nabla \theta_{\mathcal{O}^k}(x^k)],$$

and α_0 be the step size determined by one step of the projected gradient algorithm applied to $\min\left\{A^k(x) \colon x \in \mathcal{O}^k\right\}$ from y^0. Suppose $F_1, \ldots, F_n$ are the facets of $\mathcal{O}^k$. For $j = 1, \ldots, n$, let $y^j \stackrel{\text{def}}{=} \pi_{F_j}(x^k)$, $N_j \stackrel{\text{def}}{=} N_{\mathcal{O}^k}(F_j)$, $\mathcal{O}_j \stackrel{\text{def}}{=} F_j + N_j$,

$$y^j(\alpha) \stackrel{\text{def}}{=} y^j + \alpha \pi_{N_j}[-\nabla A^k_{\mathcal{O}_j}(y^j)],$$

and α_j be the step size determined by one step of the projected gradient algorithm applied to

$$\min\left\{A^k(x) \colon x \in y^j + N_j\right\} \tag{21}$$

from y^j.

Part II. Path search: Let $\mathcal{M} \stackrel{\text{def}}{=} \{0, \ldots, n\}$, $\hat{\jmath} \stackrel{\text{def}}{=} 0$ and $\alpha_0 \stackrel{\text{def}}{=} \alpha_0/\sigma$.
REPEAT

> Let $\alpha_{\hat{\jmath}} \stackrel{\text{def}}{=} \sigma \alpha_{\hat{\jmath}}$. If $\alpha_{\hat{\jmath}} \le \left\| y^{\hat{\jmath}} - x^k \right\|$ then $\mathcal{M} \stackrel{\text{def}}{=} \mathcal{M} \setminus \{\hat{\jmath}\}$.
> If $\mathcal{M} = \emptyset$ STOP; x^k is a Gauss–Newton point of f_+.
> Else let
>
> $$\hat{\jmath} \in \operatorname{argmin}\left\{A^k(y^j) + \mu_0 \left\langle \nabla A^k_{\mathcal{O}_j}(y^j), y^j(\alpha_j) - y^j \right\rangle : j \in \mathcal{M}\right\}.$$

UNTIL (20) holds for $y(\alpha) = y^{\hat{\jmath}}(\alpha_{\hat{\jmath}})$, $y = y^{\hat{\jmath}}$ and $\mathcal{O} = \mathcal{O}_{\hat{\jmath}}$.

Let $x^{k+1} \stackrel{\text{def}}{=} y^{\hat{\jmath}}(\alpha_{\hat{\jmath}})$.

Remark. For the algorithm to work properly, we assume that part I returns $\alpha_j = 0$ if y^j is already stationary for the corresponding subproblem.

Theorem 4.3 *Algorithm 2 is well defined and has NSR.*

Proof First we show that each step of the algorithm is well defined. Consider one step of the algorithm given $k \in \{0,1,2,\ldots\}$ and $x^k \in \mathrm{R}^n$. Part I is well defined because the projected gradient method is well defined. For part II we see that each iteration of the REPEAT loop is well defined; we claim that the loop terminates after finitely many iterations. Certainly if $j \in \{0, \ldots, n\}$ and $y^j \neq x^k$, then after a finite number of loop iterations in which $\hat{\jmath} = j$ and $\alpha_j \stackrel{\text{def}}{=} \sigma \alpha_j$, we have $\alpha_j \le \left\| y^j - x^k \right\|$;

hence in any subsequent loop iterations $j \notin \mathcal{M}$ and $\hat{\jmath} \neq j$. Instead suppose j is such that $y^j = x^k$. Either y^j is stationary for the jth subproblem hence $\alpha_j = 0$ and, by construction of $\mathcal{M}$, $\hat{\jmath}$ equals j for at most one loop iteration; or using Proposition 3.5.1, initially $\alpha_j > 0$ and after finitely many loop iterations in which $\hat{\jmath} = j$ and $\alpha_j \stackrel{\text{def}}{=} \sigma\alpha_j$, (20) holds, terminating the loop.

It is only left to check that x^k is a Gauss-Newton point of f_+ if $\mathcal{M} = \emptyset$. In this case, $\alpha_j \leq \left\| y^j - x^k \right\|$ for each j, in particular $\alpha_j = 0$ if $y^j = x^k$, i.e. for $j = 0$ and each j in $\mathcal{M}' = \left\{ j: 1 \leq j \leq n, x^k \in F_j \right\}$. This is only possible if x^k is stationary for each subproblem $\min\left\{ A^k(x): x \in \mathcal{O}^k \right\}$ and $\min\left\{ A^k(x): x \in x^k + N_{\mathcal{O}_j}(F_j) \right\}$ where $j \in \mathcal{M}'$. Since for each orthant $\mathcal{O}$ containing x^k we have $\nabla A^k_{\mathcal{O}}(x^k) = \nabla\theta_{\mathcal{O}}(x^k)$, it follows that x^k is also stationary for $\min\left\{ \theta(x): x \in \mathcal{O}^k \right\}$ and $\min\left\{ \theta(x): x \in x^k + N_{\mathcal{O}_j}(F_j) \right\}$ where $j \in \mathcal{M}'$. Proposition 2.3 says x^k is indeed a Gauss-Newton point of f_+.

We now prove the NSR property. Suppose that $\bar{x}$ is nonstationary for θ. As in the proof of Theorem 4.1 we assume $\mathcal{O}^k = \bar{\mathcal{O}}$ for some fixed orthant $\bar{\mathcal{O}}$ containing $\bar{x}$. Observe from Proposition 2.3, that either $\bar{x}$ is nonstationary for $\min\left\{ \theta(x): x \in \bar{\mathcal{O}} \right\}$ or $\bar{x}$ is nonstationary for $\min\{\theta(x): x \in \bar{x} + N_{\bar{\mathcal{O}}}(F)\}$, for some facet F of $\bar{\mathcal{O}}$ containing $\bar{x}$. Below we assume the latter, and deduce for x^k near $\bar{x}$ that $\theta(x^{k+1}) < \theta(\bar{x})$.

Let $\bar{\mathcal{O}}$ be an orthant containing $\bar{x}$ and F be a facet of $\bar{\mathcal{O}}$ containing $\bar{x}$. Assume $\bar{x}$ is nonstationary for $\min\{\theta(x): x \in \bar{x} + N_{\bar{\mathcal{O}}}(F)\}$. Assume further that x^k is some iterate with $\mathcal{O}^k = \bar{\mathcal{O}}$, so if $F_1, \ldots, F_n$ are the facets of $\mathcal{O}^k$, then $F = F_{\tilde{\jmath}}$ for some index $\tilde{\jmath}$. To simplify notation we omit the superscript or subscript $\tilde{\jmath}$ where possible. Let $N = N_{\bar{\mathcal{O}}}(F)$, $\mathcal{O} = F + N$, $y = \pi_F(x^k)$, $y(\alpha) = \pi_N(y - \alpha\nabla A^k_{\mathcal{O}}(y))$, and

$$\bar{A}(x) \quad \stackrel{\text{def}}{=} \quad (1/2)\left\| f(\bar{x}_+) + \nabla f(\bar{x}_+)(x_+ - \bar{x}_+) + x - x_+ \right\|^2.$$

Observe, since $\nabla\theta_{\bar{\mathcal{O}}}(\bar{x}) = \nabla\bar{A}_{\bar{\mathcal{O}}}(\bar{x})$, that $\bar{x}$ is nonstationary for $\min\left\{ \bar{A}(x): x \in y + N \right\}$.

Rewriting the $\tilde{\jmath}$th subproblem, $\min\left\{ \bar{A}_{\mathcal{O}}(x): x \in y + N \right\}$, as

$$\min\left\{ \bar{A}_{\mathcal{O}}(y + d): d \in N \right\},$$

defining $\phi(d) = \mathcal{A}^k_N(y + d)$, $\bar{\phi}(d) = \bar{A}_N(y + d)$, $\Omega = N$ and choosing $\nu > 0$, enables us to apply Lemma 3.4 and Proposition 3.5. Then there exist $\epsilon_1 > 0$, $\kappa > 0$ such that if $\left\| x^k - \bar{x} \right\| \leq \epsilon_1$ and $\phi \in U(\epsilon_1) = U(\epsilon_1, \bar{\phi}, \bar{x}, \nu)$ (see Definition 3.3), then

$$A^k(y(\alpha)) - A^k(y) \leq \mu_1 \left\langle \nabla A^k_{\mathcal{O}}(y), y(\alpha) - y \right\rangle, \tag{22}$$

$$\left\langle \nabla A^k_{\mathcal{O}}(x^k), y(\alpha) - y \right\rangle \leq -\min\{\alpha, \epsilon_1\}\kappa, \tag{23}$$

and the initial step size $\alpha_{\tilde{\jmath}}$ chosen in Part I of the algorithm is bounded below by ϵ_1. Now $\bar{A}_N$ and A^k_N are quadratic functions defined on the half-line N, hence, by continuity of ∇f, it follows easily that there exists $\epsilon_2 \in (0, \epsilon_1]$ such that $\phi \in U(\epsilon_1)$ if $\left\| x^k - \bar{x} \right\| \leq \epsilon_2$. Thus (22) and (23) hold for such x^k and $\alpha \in [0, \epsilon_2]$.

Let $\left\|x^k - \bar{x}\right\| \le \epsilon_2$ and $0 \le \alpha \le \epsilon_2$. We have

$$\begin{aligned}
&\theta(y(\alpha)) - \theta(x^k) \\
&= [A^k(y(\alpha)) - A^k(y)] \ + \ [A^k(y) - \theta(x^k)] \ + \ [\theta(y(\alpha)) - A^k(y(\alpha))] \\
&\le \mu_1 \left\langle \nabla A^k_{\mathcal{O}}(y), y(\alpha) - y \right\rangle \ + \ [A^k(y) - \theta(x^k)] \ + \ [\theta(y(\alpha)) - A^k(y(\alpha))],
\end{aligned} \tag{24}$$

using (22). Let L be an upper bound on $\left\|\nabla A^k_{\mathcal{O}}(y)\right\|$ for $x^k \in \bar{x} + \epsilon_2\mathrm{B}$, and observe

$$\left\|y(\alpha) - x^k\right\| \le \|y(\alpha) - y\| + \left\|y - x^k\right\| \ \le \ \alpha L + \left\|y - x^k\right\|.$$

Also $y = \pi_F(x^k)$ is bounded on $\bar{x}+\epsilon_2\mathrm{B}$, therefore Lemma 4.2 provides a non-decreasing error bound $\varepsilon(t) = o(t)$ such that for each $x^k \in \bar{x} + \epsilon_2\mathrm{B}$, $\alpha \in [0, \epsilon_2]$,

$$\theta(y(\alpha)) - A^k(y(\alpha)) \le \varepsilon(\alpha L + \left\|y - x^k\right\|). \tag{25}$$

Let $\hat{\kappa} = (\mu_1 - \mu_0)\sigma\kappa/2$ and choose $\bar{\alpha} \in (0, \epsilon_2)$ such that $\varepsilon(2\bar{\alpha}L) \le \hat{\kappa}\bar{\alpha}$. Now choose $\epsilon_3 \in (0, \epsilon_2)$ such that if $\left\|x^k - \bar{x}\right\| \le \epsilon_3$, then both $\left\|y - x^k\right\| \le \bar{\alpha}\min\{\sigma, L\}$ and $|A^k(y) - \theta(x^k)| \le \hat{\kappa}\bar{\alpha}$. Let $x^k \in \bar{x} + \epsilon_3\mathrm{B}$. For $\alpha \in (0, \bar{\alpha}]$, (24) and (25) yield

$$\begin{aligned}
\theta(y(\alpha)) - \theta(x^k) \ &\le \ \mu_1 \left\langle \nabla A^k_{\mathcal{O}}(y), y(\alpha) - y \right\rangle \ + \ \hat{\kappa}\bar{\alpha} \ + \ \varepsilon(\alpha L + \bar{\alpha}L) \\
&\le \ \mu_1 \left\langle \nabla A^k_{\mathcal{O}}(y), y(\alpha) - y \right\rangle \ + \ (\mu_1 - \mu_0)\sigma\bar{\alpha}\kappa \\
&\le \ \mu_1 \left\langle \nabla A^k_{\mathcal{O}}(y), y(\alpha) - y \right\rangle \ + \ (\mu_1 - \mu_0)\alpha\kappa, \quad \forall \alpha \in [\sigma\bar{\alpha}, \bar{\alpha}].
\end{aligned}$$

From (23), if $\alpha \le \epsilon_1$ then $\alpha\kappa \le -\left\langle \nabla A^k_{\mathcal{O}}(y), y(\alpha) - y \right\rangle$; therefore

$$\theta(y(\alpha)) - \theta(x^k) \le \mu_0 \left\langle \nabla A^k_{\mathcal{O}}(y), y(\alpha) - y \right\rangle, \quad \forall \alpha \in [\sigma\bar{\alpha}, \bar{\alpha}]. \tag{26}$$

From above, the initial step size $\alpha_{\tilde{\jmath}}$ and the point $y^{\tilde{\jmath}} = y$ are such that $\alpha_{\tilde{\jmath}} \ge \bar{\alpha}$, and $\sigma\bar{\alpha} > \left\|y - x^k\right\|$. We claim it follows from (26) that, during the REPEAT loop of part II, $\tilde{\jmath} \in \mathcal{M}$ and $\alpha_{\tilde{\jmath}} \ge \sigma\bar{\alpha}$. To see this suppose that $\alpha_{\tilde{\jmath}}$ decreases in some loop iteration after the first loop iteration. Then at the end of the previous loop iteration, ($\hat{\jmath} = \tilde{\jmath}$ and) the condition (20) fails for $y(\alpha) = y^{\tilde{\jmath}}(\alpha_{\tilde{\jmath}})$, $y = y^{\tilde{\jmath}}$ and $\mathcal{O} = \mathcal{O}_{\tilde{\jmath}}$; so it follows from (26) that $\alpha_{\tilde{\jmath}} > \bar{\alpha}$. Thus the new value $\sigma\alpha_{\tilde{\jmath}}$ of $\alpha_{\tilde{\jmath}}$ is bounded below by $\sigma\bar{\alpha}$, hence also $\sigma\alpha_{\tilde{\jmath}} > \left\|y - x^k\right\|$ and $\tilde{\jmath}$ is not deleted from $\mathcal{M}$.

Therefore after the REPEAT loop terminates, $\tilde{\jmath} \in \mathcal{M}$ and $\alpha_{\tilde{\jmath}} \ge \sigma\bar{\alpha}$; and the selection of x^{k+1}, whether or not using $y^{\tilde{\jmath}}(\cdot) = y(\cdot)$, satisfies

$$\begin{aligned}
\theta(x^{k+1}) &\le \min\left\{A^k(y^j) + \mu_0 \left\langle \nabla A^k_{\mathcal{O}_j}(y^j), y^j(\alpha_j) - y^j \right\rangle : j \in \mathcal{M}\right\} \\
&\le A^k(y) + \mu_0 \left\langle \nabla A^k_{\mathcal{O}}(y), y(\alpha_{\tilde{\jmath}}) - y \right\rangle \\
&\le A^k(y) - \mu_0 \min\{\alpha_{\tilde{\jmath}}, \epsilon_1\}\kappa \quad \text{(from (23))} \\
&\le A^k(y) - \delta,
\end{aligned}$$

where $\delta \stackrel{\text{def}}{=} \sigma\mu_0\bar{\alpha}\kappa$ is a positive constant independent of x^k. As noted above, $A^k(y) \to \theta(\bar{x})$ as $x^k \to \bar{x}$, so $\theta(x^{k+1}) < \theta(\bar{x})$ for x^k sufficiently close to $\bar{x}$.

A similar argument can be made for the case when $\mathcal{O}^k = \bar{\mathcal{O}}$ and $\bar{x}$ is nonstationary for $\min\left\{\theta(x): x \in \bar{\mathcal{O}}\right\}$. In this case $\tilde{\jmath} = 0$, $y = x^k$ and $y(\alpha) = \pi_{\bar{\mathcal{O}}}(x^k - \alpha\nabla A^k_{\bar{\mathcal{O}}}(x^k))$. We do not give details, but only note that this process is somewhat simpler than that above because the inequality corresponding to (24) only has two summands on the right:

$$\theta(y(\alpha)) - \theta(x^k) \leq \mu_1 \left\langle \nabla A^k(x^k), y(\alpha) - x^k \right\rangle \; + \; [\theta(y(\alpha)) - A^k(y(\alpha))].$$

Since there are only finitely many choices of $\bar{\mathcal{O}}$, the NSR property of Algorithm 2 is established. □

4.3 A hybrid algorithm with quadratic local convergence

Both of the algorithms given above have at best a linear rate of convergence because the projected gradient method is only a first-order method. However, if an algorithm for finding a Gauss-Newton point of f_+ has NSR (such as Algorithms 1 and 2), then this lends itself to hybrid methods that alternate between steps of the original algorithm and Newton-like steps and therefore admit the possibility of quadratic local convergence. For such a hybrid algorithm, let $\mathcal{K}$ be the set of indices k for which the original algorithm determines x^{k+1}. If $\mathcal{K}$ has infinitely many elements and monotonicity of the algorithm is maintained, accumulation points of the subsequence $\{x^k\}_{k\in\mathcal{K}}$ are Gauss–Newton points of f_+. If such a limit point x^* is in fact a point of attraction of a Newton method, and a Newton step is taken every ℓth iteration, then convergence will be ℓ-step superlinear, or ℓ-step quadratic if ∇f is Lipschitz. See [2] for details on a related hybrid algorithm in the context of quadratic programming.

We briefly sketch three popular Newton methods for solving the nonsmooth equation $f_+(x) = 0$, which often produce Q-quadratically convergent sequences of iterates. To make comparisons easy, we use the general notion of a *Newton path* [28] which, given the iterate x^k, is some function $p^k : [0,1] \to \mathrm{R}^n$ with $p^k(0) = x^k$; the next iterate x^{k+1} is defined as $p^k(\alpha)$ for some $\alpha \in [0,1]$ (details are given below). We say a *Newton iterate* or *Newton step* is taken if $x^{k+1} = p^k(1)$. We may not take a Newton step, however, if it does not yield "sufficient progress". A simple damping strategy is used to ensure sufficient progress: recall the constants $\mu_0, \sigma \in (0,1)$, and define α as the largest member of $\{1, \sigma, \sigma^2, \ldots\}$ such that

$$\left\| f_+(p^k(\alpha)) \right\| \;\leq\; (1 - \mu_0\alpha) \left\| f_+(x^k) \right\|. \tag{27}$$

Then $x^{k+1} \stackrel{\text{def}}{=} p^k(\alpha)$; this is the damped Newton iterate.

Newton path 1. Given k and x^k, let $\mathcal{O}^k$ be an orthant containing x^k, $M^k = \nabla f_+|_{\mathcal{O}^k}$,

$$d^k = -(M^k)^{-1} f_+(x^k), \quad p^k(\alpha) = x^k + \alpha d^k.$$

Newton path 2. Given k and x^k, let $B^k = f'_+(x^k;\cdot)$, the B-derivative of f_+ at x^k,

$$d^k = (B^k)^{-1}[-f_+(x^k)], \quad p^k(\alpha) = x^k + \alpha d^k.$$

Newton path 3. Given k and x^k, let L^k_+ be the linearization of f_+ at x^k given by (19),

$$p^k(\alpha) \ = \ (L^k_+)^{-1}[(1-\alpha)f_+(x^k)].$$

Let Newton 1 be Newton's method with Newton path 1, etc. We give sufficient conditions for Q-quadratic convergence of these algorithms.

Assumption (A). The point $\bar{x}$ is such that first, for each orthant $\bar{\mathcal{O}}$ containing $\bar{x}$, $\nabla f_+|_{\bar{\mathcal{O}}}(\bar{x})$ is invertible; and second, for some neighborhood V of $\bar{x}_+$, ∇f is Lipschitz in $V \cap \mathrm{R}^n_+$.

Assumption (B). The point $\bar{x}$ is such that first, $f'_+(\bar{x};\cdot)$ is invertible; and second, for some neighborhood V of $\bar{x}_+$, ∇f is Lipschitz in $V \cap \mathrm{R}^n_+$.

Suppose $f_+(\bar{x}) = 0$. Assumption (A) is sufficient for Newton 1 to produce a sequence $\{x^k\}$ such that the Newton step is taken at each iteration $k' \geq k$, if some x^k is near enough to $\bar{x}$ [15, 26]. Furthermore, $\{x^k\}$ converges Q-quadratically to $\bar{x}$. However (A) may not be sufficient for Newton 2 and Newton 3 to be well-defined, even though some x^k is arbitrarily close to $\bar{x}$. We need condition (B): if some x^k is near enough to $\bar{x}$, then the same conclusion holds for Newton 2 [26] and Newton 3 [28, 34] as holds above for Newton 1. Assumption (B) implies assumption (A) but not vice versa; assumption (A) is called BD-regularity [27]. Also note that the condition $f_+(\bar{x}) = 0$ is implied, assuming (B) holds, by $\bar{x}$ being a Gauss-Newton point of f_+ (Lemma 2.9). Further work on Newton 2 is given in [12, 20, 27] and on Newton 3 in [5, 4].

We define a hybrid algorithm involving Algorithm 2, for global convergence properties under no regularity assumptions, and Newton's method for fast local convergence under one of the regularity assumptions (A) or (B).

Hybrid. Let $x^0 \in \mathrm{R}^n$ and (in addition to the constants used in Algorithm 2) $\bar{\alpha} \in (0, 1]$. Fix "Newton's method" as one of Newton 1, Newton 2 and Newton 3. Given $k \in \{0, 1, 2, \ldots,\}$ and $x^k \in \mathrm{R}^n$, define x^{k+1} as follows.

If $f_+(x^k) = 0$ then STOP.
If Newton's method produces $p^k(\alpha)$ such that $\alpha \geq \bar{\alpha}$, then accept this as the next iterate: $x^{k+1} \stackrel{\text{def}}{=} p^k(\alpha)$. Otherwise, let Algorithm 2 produce x^{k+1} from x^k.

Observe that given x^k it may not be possible to carry out a step of Newton's method, for example in Newton 1 it may be that M^k is a singular matrix; yet the Hybrid algorithm is still well defined (via Algorithm 2). On the other hand, if Newton's method is feasible and provides sufficient descent, for instance by taking a Newton step, then Algorithm 2 is not invoked. Also, by setting $\bar{\alpha} = 1$, Newton's method is restricted to taking a Newton step; this is rejected by the Hybrid algorithm if (27) is violated for $\alpha = 1$.

The proof of the next result follows easily from Theorem 4.3 and the above remarks on convergence properties of algorithms Newton 1-3.

Theorem 4.4 *The Hybrid algorithm either terminates after finitely many steps at a Gauss-Newton point $\bar{x}$ of f_+, or produces an infinite sequence $\{x^k\}$ each limit point $\bar{x}$ of which is a Gauss-Newton point of f_+. Moreover, if either of the following conditions hold:*

1. *$f_+(\bar{x}) = 0$, the Hybrid uses Newton 1 and assumption (A),*
2. *the Hybrid uses any of Newton 1-3 and assumption (B),*

then $\bar{x}$ is a zero of f_+ such that after finitely many iterations each iteration is a Newton step, $\{x^k\}$ actually converges to $\bar{x}$, and the rate of convergence is Q-quadratic.

4.4 Examples

We now give some examples of the algorithms described in this section. The first example is a linear complementarity problem, to find $z \geq 0$ with $f(z) \stackrel{\text{def}}{=} Mz + q \geq 0$, where

$$M = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad q = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

The unique solution of this problem is $\bar{z} = (0, 1)$. The corresponding zero of the normal map $f_+(x)$ is $\bar{x} = (-1, 1)$.

The difficulty with this example is that if we start at any point $x \geq 0$, the corresponding piece of the normal map is not invertible. We will first show how to apply Algorithm 1 to this problem. In all the calculations below we will use exact minimization to choose the step length α rather than the less stringent conditions (7)-(9). Also note that A^k and θ are identical for each k, because f is an affine function.

Suppose that we start at the point $x^0 = (1/2, 1/2)$ at which $\theta(x^0) = 1/2$. Then $\mathcal{O}^0 = \mathrm{R}^2_+$ and $\nabla\theta_{\mathcal{O}^0}(x^0) = (1, 1)$. Thus $y^0(\alpha) = (1/2 - \alpha, 1/2 - \alpha)_+$. The value of α that minimizes $\theta(y^0(\alpha))$ over $\alpha \geq 0$ is $\alpha_0 = 1/4$, giving $y^0(\alpha_0) = (1/4, 1/4)$ corresponding to $\theta(y^0(\alpha_0)) = 1/4$.

We enumerate the facets of $\mathcal{O}^0$ by setting the corresponding $x_i = 0$. Thus $F_1 = 0 \times \mathrm{R}_+$, $N_1 = \mathrm{R}_- \times 0$, $\mathcal{O}_1 = \mathrm{R}_- \times \mathrm{R}_+$, $y^1 = (0, 1/2)$ and $\nabla\theta_{\mathcal{O}_1}(y^1) = (1/2, 0)$. Hence

$y^1(\alpha) = (-\alpha/2, 1/2)$, and $\theta(y^1(\alpha))$ is minimized over $\alpha \geq 0$ at $\alpha = 1$. This gives $\alpha_1 = 1$, $y^1(\alpha_1) = (-1/2, 1/2)$ and $\theta(y^1(\alpha_1)) = 1/8$.

Since $F_2 = \mathrm{R}_+ \times 0$ it follows that $N_2 = 0 \times \mathrm{R}_-$, $\mathcal{O}_2 = \mathrm{R}_+ \times \mathrm{R}_-$, $y^2 = (1/2, 0)$, and $\nabla\theta_{\mathcal{O}_2}(y^2) = (0, -1/2)$. Then $\pi_{N_2}(-\nabla\theta_{\mathcal{O}_2}(y^2)) = 0$, and $y^2(\alpha) = y^2$ for each $\alpha \geq 0$. So we take the optimal value of α as $\alpha_2 = 0$, yielding $y^2(\alpha_2) = (1/2, 0)$ corresponding to $\theta(y^2(\alpha_2)) = 1/4$.

Thus, in Algorithm 1, $x^1 = y^1(\alpha_1)$ and we proceed to the second iteration. In fact, $\mathcal{O}^2 = \mathrm{R}_- \times \mathrm{R}_+$ remains the chosen orthant for the remaining iterations, and it can be easily calculated that $\{x^k\}$ converges at a linear rate to $\bar{x} = (-1, 0)$. However, in $\mathcal{O}^2$ the linear map is invertible, so if the Hybrid were employed at this point, then the algorithm with any of Newton 1–3 would move to the solution in a single step.

If instead, we start at the point $x^0 = (1/4, 1/4)$, then $y^0 = (1/4, 1/4)$, $\nabla\theta_{\mathcal{O}_0}(y^0) = (0, 0)$, hence $\alpha_0 = 0$ and $y^0(\alpha_0) = (1/4, 1/4)$ with $\theta(y^0(\alpha_0)) = 1/4$. It is easy to calculate that $y^1 = (0, 1/4)$, $\nabla\theta_{\mathcal{O}_1}(y^1) = (1/4, -1/2)$, $y^1(\alpha) = (-\alpha/4, 1/4)$, hence the optimal value of α is $\alpha_1 = 1$ which gives $y^1(\alpha_1) = (-1/4, 1/4)$ with $\theta(y^1(\alpha_1)) = 9/32$. Also, $\alpha_2 = 0$ implying that $y^2(\alpha_2) = (1/4, 0)$ with $\theta(y^2(\alpha_2)) = 5/16$. Hence the algorithm terminates at $(1/4, 1/4)$ which is a Gauss–Newton point. Clearly, it is not regular. Note that the hypotheses of Theorem 2.6, Corollary 2.7, Theorem 2.8 and Lemma 2.9 are all violated at this point.

If $x^0 = (1/2, 3/4)$, then it can also be seen that $x^1 = (-3/4, 3/4)$ and so the algorithm similarly to the first example and converges to the solution of the complementarity problem. In order to avoid converging to a non-regular Gauss–Newton point, we need to get enough descent on one of the rays to leave the nonnegative orthant.

In all the above examples, $A^k = \theta$ and hence the steps of Algorithm 1 are identical to those of Algorithm 2. We now adapt the example to a nonlinear complementarity problem.

Let

$$f(z) = \begin{bmatrix} \frac{2}{3}z_1^3 + z_1 z_2 + \frac{1}{2}z_2 + \frac{5}{12} \\ z_1^2 + z_2^2 - \frac{1}{2} \end{bmatrix}$$

and consider starting Algorithm 2 at $x^0 = (1/2, 1/2)$. Note that the unique solution of the corresponding nonlinear complementarity problem is $z^* = (0, 1/\sqrt{2})$ with corresponding $x^* = z^* - f(z^*) = (-0.7702, 0.7071)$. We first construct a piecewise linear model of f_+ at x^0, which results in the linear complementarity problem given above. Hence applying the projected gradient algorithm to the cell and the two rays leads to

$$y^0(\alpha_0) = (1/4, 1/4), \quad y^1(\alpha_1) = (-1/2, 1/2), \quad y^2(\alpha_2) = (1/2, 0),$$

with corresponding values of $\alpha_0 = 1/4$, $\alpha_1 = 1/2$, $\alpha_2 = 0$. Proceeding to Part II of Algorithm 2, we note that $\mathcal{M} = \{0, 1\}$ and that for $\mu_0 = 0.1$, the value of the models linearized around $y^0 = (1/2, 1/2)$ and $y^1 = (0, 1/2)$ are

$$j = 0 1/2 + 0.1 \langle (1,1), (1/4, 1/4) - (1/2, 1/2) \rangle = 0.45$$
$$j = 1 1/4 + 0.1 \langle (1/2, 0), (-1/2, 1/2) - (0, 1/2) \rangle = 0.225$$

respectively. Thus we evaluate $\theta(y^1(\alpha_1)) = 0.045$ which is below the linear model value of 0.225, hence $x^1 = (-1/2, 1/2)$. Algorithm 2 now proceeds from this point by relinearizing. Applying the Hybrid algorithm in the new cell would lead to the solution in very few iterations.

Acknowledgement. We thank an anonymous referee for drawing our attention to reference [16].

References

[1] J. V. Burke and M. C. Ferris, A Gauss–Newton method for convex composite optimization, Technical Report, 1176, Computer Sciences Department, University of Wisconsin (Madison, Wisconsin 53706, 1993).

[2] P. H. Calamai and J. J. Moré, Projected gradient methods for linearly constrained problems, *Mathematical Programming* 39 (1987) 93–116.

[3] R. W. Cottle, J. S. Pang and R. E. Stone, *The Linear Complementarity Problem* (Academic Press, Boston, 1992).

[4] S. P. Dirkse and M. C. Ferris, A pathsearch damped Newton method for computing general equilibria, Mathematical Programming Technical Report 94-03, Computer Sciences Department, University of Wisconsin (Madison, Wisconsin 53706, 1994).

[5] S. P. Dirkse and M. C. Ferris, The PATH solver: A non-monotone stabilization scheme for mixed complementarity problems, *Optimization Methods & Software* (1994, forthcoming).

[6] M. C. Ferris and S. Lucidi, Nonmonotone stabilization methods for nonlinear equations, *Journal of Optimization Theory and Applications* 81 (1994) 53–74.

[7] R. Fletcher, *Practical Methods of Optimization* (John Wiley & Sons, New York, second edn., 1987).

[8] M. Fukushima, Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems, *Mathematical Programming* 53 (1992) 99–110.

[9] S. A. Gabriel and J. S. Pang, An inexact NE/SQP method for solving the nonlinear complementarity problem, *Computational Optimization and Applications* 1 (1992) 67–91.

[10] P. T. Harker, *Lectures on Computation of Equilibria with Equation–Based Methods*, CORE Lecture Series (CORE Foundation, Louvain–la–Neuve, Université Catholique de Louvain, 1993).

[11] P. T. Harker and J. S. Pang, Finite–dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications, *Mathematical Programming* 48 (1990) 161–220.

[12] P. T. Harker and B. Xiao, Newton's method for the nonlinear complementarity problem: A B–differentiable equation approach, *Mathematical Programming* 48 (1990) 339–358.

[13] N. H. Josephy, Newton's method for generalized equations, Technical Summary Report 1965, Mathematics Research Center, University of Wisconsin (Madison, Wisconsin, 1979).

[14] C. Kanzow, Some equation–based methods for the nonlinear complementarity problem, *Optimization Methods and Software* 3 (1994) 327–340.

[15] M. Kojima and S. Shindo, Extensions of Newton and quasi-Newton methods to systems of PC^1 equations, *Journal of Operations Research Society of Japan* 29 (1986) 352–374.

[16] B. Kummer, Newton's method for non–differentiable functions, in: *Advances in Mathematical Optimization* (Akademie–Verlag, Berlin, 1988) pp. 114–125.

[17] O. L. Mangasarian, Equivalence of the complementarity problem to a system of nonlinear equations, *SIAM Journal on Applied Mathematics* 31 (1976) 89–92.

[18] P. Marcotte and J.-P. Dussault, A note on a globally convergent Newton method for solving monotone variational inequalities, *Operations Research Letters* 6 (1987) 35–42.

[19] J. J. Moré, Global methods for nonlinear complementarity problems, Technical Report, MCS-P429-0494, Argonne National Laboratory (Argonne, Illinois, 1994).

[20] J. S. Pang, Newton's method for B–differentiable equations, *Mathematics of Operations Research* 15 (1990) 311–341.

[21] J. S. Pang, A B–differentiable equation based, globally and locally quadratically convergent algorithm for nonlinear programs, complementarity and variational inequality problems, *Mathematical Programming* 51 (1991) 101–132.

[22] J. S. Pang and S. A. Gabriel, NE/SQP: A robust algorithm for the nonlinear complementarity problem, *Mathematical Programming* 60 (1993) 295–338.

[23] J. S. Pang, S.-P. Han and N. Rangaraj, Minimization of locally Lipschitzian functions, *SIAM Journal on Optimization* 1 (1991) 57–82.

[24] J. S. Pang and L. Qi, Nonsmooth equations: Motivation and algorithms, *SIAM Journal on Optimization* 3 (1993) 443–465.

[25] E. Polak, *Computational Methods in Optimization; A Unified Approach* (Academic Press, New York, 1971).

[26] L. Qi, Convergence analysis of some algorithms for solving nonsmooth equations, *Mathematics of Operations Research* 18 (1993) 227–244.

[27] L. Qi and J. Sun, A nonsmooth version of Newton's method, *Mathematical Programming* 58 (1993) 353–368.

[28] D. Ralph, Global convergence of damped Newton's method for nonsmooth equations, via the path search, *Mathematics of Operations Research* 19 (1994) 352–389.

[29] S. M. Robinson, Mathematical foundations of nonsmooth embedding methods, *Mathematical Programming* 48 (1990) 221–229.

[30] S. M. Robinson, An implicit–function theorem for a class of nonsmooth functions, *Mathematics of Operations Research* 16 (1991) 292–309.

[31] S. M. Robinson, Homeomorphism conditions for normal maps of polyhedra, in: A. Ioffe, M. Marcus and S. Reich eds., *Optimization and Nonlinear Analysis*, Pitman Research Notes in Mathematics Series No. 244 (Longman, Harlow, Essex, England, 1992) 240–248.

[32] S. M. Robinson, Normal maps induced by linear transformations, *Mathematics of Operations Research* 17 (1992) 691–714.

[33] S. M. Robinson, Nonsingularity and symmetry for linear normal maps, *Mathematical Programming* 62 (1993) 415–425.

[34] S. M. Robinson, Newton's method for a class of nonsmooth functions, *Set Valued Analysis* (1993, forthcoming).

[35] R. T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, New Jersey, 1970).

[36] P. K. Subramanian, Gauss–Newton methods for the complementarity problem, *Journal of Optimization Theory and Applications* 77 (1993) 467–482.