ON THE EXTENSION OF CONSTRAINED OPTIMIZATION ALGORITHMS FROM DIFFERENTIABLE TO NONDIFFERENTIABLE PROBLEMS

by

E. Polak, D. Q. Mayne and Y. Wardi

Memorandum No. UCB/ERL M81/78

14 April, 1981

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

ON THE EXTENSION OF CONSTRAINED OPTIMIZATION ALGORITHMS FROM DIFFERENTIABLE TO NONDIFFERENTIABLE PROBLEMS

E. Polak, D.Q. Mayne* and Y. Wardi

Department of Electrical Engineering and Computer Sciences
and the Electronics Research Laboratory
University of California, Berkeley, California 94720

ABSTRACT

This paper presents three general schemes for extending differentiable optimization algorithms to nondifferentiable problems. It is shown that the Armijo gradient method, phase I - phase II methods of feasible directions and exact penalty function methods have conceptual analogs for problems with locally Lipschitz functions and implementable analogs for problems with semi-smooth functions. The exact penalty method required the development of a new optimality condition.

* Department of Computing and Control, Imperial College of Science and Technology, London, SW7 England

Research sponsored by the National Science Foundation Grants ECS-79-13148 ECS-79-13148/CEE-8105790 and the Joint Services Electronics Program Contract F49620-79-C-0178.

Introduction

Over the last several years, we have first witnessed systematic efforts to extend the concepts of the calculus to locally Lipschitz functions (see e.g., [C1, L1, L2]), and to extend optimality conditions for differentiable optimization problems to optimization problems with locally Lipschitz functions (see e.g., [C2, G1, L2, P7, P13]). As a result, we now have an analog of the extended F. John multiplier rule for nondifferentiable mathematical programming problems [C2], analogs of Lagrangians [C1] and an analog of the Maximum principle for nondifferentiable optimal control problems [C3].

The development of nondifferentiable optimization algorithms, for the non-convex case, has been far less systematic. Two distinct approaches have emerged: that of the Kiev school, which constructs algorithms without a monotonic descent property [S1, S2, P12], and the one favored in the West, which always insists on monotonic descent of the cost or of a surrogate cost [B2, G1, P2, P4, P7]. In this paper we are concerned with algorithms of the second type. Although the literature on nondifferentiable optimization algorithms of the second type is still extremely small, two principles seem to have emerged. The first principle (see e.g., [B2, G1, L2, D2, P7]) is that in extending a differentiable optimization algorithm to the nondifferentiable case it is necessary to replace gradients not with corresponding generalized gradients, but with bundles of generalized gradients in order to make up for the lack of continuity of the generalized gradients. The bundle-size parameter (ε) then has to be driven to zero as an optimal point is approached. The second principle was developed in [M1, L2, L5, W1, W2, P2, P7]. The gist of it is that when functions are semi-smooth, it is possible to get a good approximation to the nearest point

from the origin to their generalized gradient bundles in a finite number of operations. The importance of this fact is that it defines an important class of nondifferentiable optimization problems for which one can obtain implementable algorithms, i.e., algorithms in which all the computations that are required to be performed in each iteration can be carried out in a finite number of simple operations.

In this paper we develop three general schemes for the extension of differentiable optimization algorithms to non-differentiable problems. The first one is for unconstrained optimization, while the remaining ones are for constrained optimization algorithms. To illustrate the applicability of these schemes, we use them to construct several conceptual algorithms for optimization problems with locally Lipschitz functions. These include an extension of the Armijo gradient method (which had previously been presented in [P7]), extensions of two phase I - phase II methods of feasible directions of the type discussed in [P3], the extensions of exact penalty methods [C5, P14]. The extension of exact penalty methods required the development of a sharper optimality condition for constrained problems than the ones found in [C2]. Finally, for the semi-smooth case, we show that the conceptual algorithms give rise to implementable algorithms in a totally systematic manner. We hope that the results presented in this paper will contribute to the understanding and development of nondifferentiable optimization algorithms.

1. Preliminary Results

Our analysis of algorithms for non-smooth optimization will be based on a very small number of non-smooth analysis results. For the sake of convenience, we begin by summarizing these; for details and proofs, the reader is referred to [C1, C2, L1].

Definition 1.1 [Cl]: Let $f : \mathbb{R}^n \to \mathbb{R}^1$ be locally Lipschitz continuous. The generalized gradient of f at x is defined to be the set

$$\partial f(x) \triangleq \text{co}\{\lim_{v_i \to 0} \nabla f(x+v_i)\} \tag{1.1}$$

where co denotes the convex hull of a set, and the v_i are such that $\nabla f(x+v_i)$ exists, and $\lim_{v_i \to 0} \nabla f(x+v_i)$ exists. ¤

Definition 1.2 [Cl]: Let $f : \mathbb{R}^n \to \mathbb{R}^1$ be locally Lipschitz continuous. The generalized directional derivative of f at x in the direction h is defined to be

$$d^0 f(x;h) \triangleq \lim_{\substack{y \to 0 \\ \lambda \downarrow 0}} \frac{f(x+y+\lambda h)-f(x+y)}{\lambda} \tag{1.2}$$

Fact 1.1 [Cl]: Let $f : \mathbb{R}^n \to \mathbb{R}^1$ be locally Lipschitz continuous. Then

a) $\partial f(x)$ exists and is compact at all $x \in \mathbb{R}^n$;

b) $\partial f(x)$ is bounded on bounded sets;

c) $\partial f(\cdot)$ is u.s.c. in the sense that $\{x_i \to \hat{x},\ y_i \in \partial f(x_i)$ and $y_i \to \hat{y}\} \Rightarrow \{\hat{y} \in \partial f(\hat{x})\}$;

d) $d^0 f(x;v)$ exists for all $x,\ v \in \mathbb{R}^n$;

e) $$d^0 f(x;v) = \max_{\xi \in \partial f(x)} \langle \xi, v \rangle; \tag{1.3}$$

f) Whenever the directional derivative $df(x;v)$ exists,

$$df(x;v) \leq d^0 f(x;v), \tag{1.4}$$

furthermore, when f is $\mathbb{C}^1$ at x, equality holds;

g) if x and h are such that $d^0 f(x+sh;h) \leq -\alpha < 0$ for all $s \in [0,1]$, then

$$f(x+sh) - f(x) \leq -\alpha s \quad \forall\ s \in [0,1],\ \forall\ \alpha \in (0,1) \tag{1.5}$$

¤

Fact 1.2 (Mean Value Theorem)[L1]: Let $f : \mathbb{R}^n \to \mathbb{R}^1$ be locally Lipschitz continuous. Then, given $x, y \in \mathbb{R}^n$

$$f(y) - f(x) = \langle \xi, y - x \rangle \tag{1.6}$$

for some $\xi \in \partial f(x+s(y-x))$ and $s \in [0,1]$. ¤

Fact 1.3 [C2]: Let $f : \mathbb{R}^n \to \mathbb{R}^1$, $g^i : \mathbb{R}^n \to \mathbb{R}^1$, $i \in \underline{m} \triangleq \{1,2,\ldots,m\}$; $h^j : \mathbb{R}^n \to \mathbb{R}^1$, $j \in \underline{\ell} \triangleq \{1,2,\ldots,\ell\}$ be locally Lipschitz continuous and let $\hat{x}$ be a solution of the problem

$$\min\{f(x) | g^i(x) \leq 0,\ i \in \underline{m},\ h^j(x) = 0,\ j \in \underline{\ell}\}. \tag{1.7}$$

Then

$$0 \in \mathrm{co}\{\partial f(\hat{x}) \cup \{\partial g^i(\hat{x}) | i \in I(\hat{x})\} \cup \{t^j \partial h^i(\hat{x}) | j \in \underline{\ell}\}\}, \tag{1.8}$$

where $I(\hat{x}) \triangleq \{i \in \underline{m} | g^i(\hat{x}) = 0\}$ and $t_j \in \{+1,-1\}$.

The above result is not quite strong enough to be used in the context of exact penalty function methods and hence we had to propose the new optimality condition stated below. We wish to thank Prof. F. Clarke for supplying us with a proof (he has subsequently proved this result without requiring that the set $\{x|F(x) = 0\}$, have measure zero).

Theorem 1.1: Let f, g^i, $i \in \underline{m}$; h^j, $j \in \underline{\ell}$, from $\mathbb{R}^n$ into $\mathbb{R}^1$ be locally Lipschitz continuous. Let $\hat{x}$ be a solution to (1.7) and let $F : \mathbb{R}^n \to \mathbb{R}^1$ be defined by

$$F(x) \triangleq \max\{f(x)-f(\hat{x});\ g^i(x)_+,\ i \in \underline{m};\ |h^j(x)|,\ j \in \underline{\ell}\}, \tag{1.9}$$

where $g^i(x)_+ \triangleq \max\{g^i(x), 0\}$. Suppose that $\{x|F(x) = 0\}$ has measure zero, then

$$0 \in \text{co}\{\partial f(\hat{x}) \cup \{\partial g^i(\hat{x})_+ \cap \partial g^i(\hat{x}) | i \in I(\hat{x})\} \cup \{t_j \partial h^j(\hat{x}) | j \in \underline{\ell}\}\} \tag{1.10}$$

where $I(\hat{x}) = \{i \in \underline{m} | \ g^i(\hat{x}) = 0\}$ and $t_j \in \{+1,-1\}$.

Proof: Although F. Clarke has proved the above result for a somewhat more general case, we shall only give a proof for the slightly restrictive case where $\hat{x}$ is also a local solution to $\min\{f(x) | g^i(x) \leq 0, i \in \underline{m}; t_j h^j(x) \leq 0, j \in \underline{\ell}\}$. (We note that (1.8) is also an optimality condition for this case). We note that $g^i(x)_+ > 0$ and $|h^j(x)| > 0$ for some $i \in \underline{m}$, $j \in \underline{\ell}$ whenever x is infeasible. Furthermore, $f(x) - f(\hat{x}) \geq 0$ for all x which are feasible. Hence, $F(x) \geq 0$ for all x. Consequently, $\hat{x} = \arg \min_{x \in \mathbb{R}^n} F(x)$ so that $0 \in \partial F(\hat{x})$. Now, by assumption $\{x | F(x) = 0\}$ has measure zero and hence (1.10) follows directly from the fact that $\partial F(\hat{x})$ involves the limit of gradients $\nabla g^i(x)$ evaluated only at points x where $g^i(x) > 0$. ¤

2. Unconstrained optimization

Let $f : \mathbb{R}^n \to \mathbb{R}^1$ be locally Lipschitz continuous. Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x). \tag{2.1}$$

We shall consider algorithms for solving (2.1) of the form

$$x_{i+1} = x_i + \lambda_i h_i, \tag{2.2a}$$

$$\lambda_i = \arg \max_{k \in \mathbb{N}_+} \{\beta^k | f(x_i + \beta^k h_i) - f(x_i) \leq \alpha \beta^k \delta_i\}, \tag{2.2b}$$

where $\alpha, \beta \in (0,1)$ $\mathbb{N}_+ = \{1,2,3,\ldots\}$, and $\delta_i < 0$. We recognize these algorithms as a generalization of the class of descent algorithms,

utilizing the Armijo step size rule [A1], that were discussed by Polak, Sargent and Sebastian in [P9], for the differentiable case. Although most, if not all, differentiable unconstrained optimization algorithms of the form considered by Polak, Sargent and Sebastian can be analysed in terms of the convergence theorem (1.3.10) in [P8], their structure permits the introduction of more readily verifiable assumptions than those found in Theorem (1.3.10) in [P8]. Consequently in [P9], we find (in a slightly different form) the following result, which is intended to be used for algorithms of the form (2.2a), (2.2b) when $\delta_i = df(x_i;h_i)$.

Theorem 2.1: Suppose that $f : \mathbb{R}^n \to \mathbb{R}^1$ is C^1 and that there exist two continuous functions N_1, $N_2 : \mathbb{R}^n \to \mathbb{R}^+$, which vanish only at points x for which $\nabla f(x) = 0$, such that for h_i in (2.2a)

$$df(x_i;h_i) = \langle \nabla f(x_i), h_i \rangle \leq -N_1(x_i), \tag{2.3}$$

$$\|h_i\| \leq N_2(x_i) \tag{2.4}$$

hold.

Then, given an $\bar{x}$ such that $\nabla f(\bar{x}) \neq 0$, there exist a $\bar{\rho} > 0$, and a $\bar{k} \in \mathbb{N}_+$ such that for all $x_i \in B(\bar{x},\bar{\rho}) \triangleq \{x \in \mathbb{R}^n \mid \|x - \bar{x}\| \leq \bar{\rho}\}$,

$$f(x_i+\lambda h_i) - f(x_i) \leq \lambda\alpha df(x_i;h_i) \leq -\lambda\alpha N_1(\bar{x})/2, \ \forall \lambda \in [0,\beta^{\bar{k}}]. \tag{2.5}$$

¤

Relation (2.5) leads to two conclusions: for all $x_i \in B(\bar{x},\bar{\rho})$

(i) $\lambda_i \geq \beta^{\bar{k}}$, and

(ii) $f(x_{i+1}) - f(x_i) \leq -\beta^{\bar{k}}\alpha N_1(\bar{x})/2$,

i.e. the algorithm map defined by (2.2a) (2.2b), with $\delta_i \triangleq df(x_i;h_i)$, is locally uniformly monotonic (see [T1]). As an immediate consequence, we see from theorem (1.3.9) in [P8] that any accumulation point $\hat{x}$ of $\{x_i\}$ satisfies $\nabla f(\hat{x}) = 0$.

Assumption 2.1: From now on, we shall assume that the function $f : \mathbb{R}^n \to \mathbb{R}^1$ is locally Lipschitz continuous. ¤

Any attempt to extend Theorem 1.1 to the case of $f(\cdot)$ locally Lipschitz only, by replacing df with d^0f in (2.3) is doomed to failure, as can be seen from the counter example in [W2]. This is due to the fact that although an h_i satisfying $d^0f(x_i;h_i) \leq -N_1(x_i)$ and (2.4) is obviously a descent direction, it is not possible to ensure that the step size λ_i is bounded from below in a ball about an $\bar{x}$ such that $0 \notin \partial f(\bar{x})$. To insure that a nonsmooth optimization algorithm is locally uniformly monotonic, it becomes necessary to "look ahead" for the "corners" of $f(\cdot)$ by "smearing" $\partial f(x)$, as follows.

Definition 2.1: For any $\varepsilon > 0$, we define the ε-smeared generalized gradient by

$$\partial_\varepsilon f(x) \triangleq co\{ \bigcup_{x' \in B(x,\varepsilon)} \partial f(x') \} \tag{2.6}$$

¤

Fact 2.1: For any $\varepsilon > 0$, $\partial_\varepsilon f(x)$ is compact, bounded on bounded sets; furthermore $\partial_\varepsilon f(\cdot)$ is upper semicontinuous (u.s.c.) (see [P7]). ¤

Definition 2.2: For any $\varepsilon > 0$, we define the ε-smeared generalized directional derivative of $f(\cdot)$ at x, in the direction h by

$$d^0_\varepsilon f(x;h) \triangleq \max_{\xi \in \partial_\varepsilon f(x)} \langle \xi, h \rangle \tag{2.7}$$

With the introduction of $d_\varepsilon^0 f(\cdot;\cdot)$, and ignoring for the moment the problem of choosing $\varepsilon > 0$, as well as that of computing $\partial_\varepsilon f(x)$ and $d_\varepsilon^0(x;h)$, we are ready to extend Theorem 1.1 to the non-smooth case. We shall refer to algorithms which assume that $\partial_\varepsilon f(x)$ and $d_\varepsilon^0(x;h)$ can be computed exactly as conceptual.

In anticipation of the application of the new theorem to conceptual optimization algorithms for non-smooth problems, we find it necessary to relax the continuity of N_1, N_2 in Theorem 1.1 to a requirement which is somewhat weaker than semi-continuity, as we shall now see.

Theorem 2.2 (conceptual Algorithms): Let $\varepsilon > 0$ be given. Suppose that there exist two functions N_1, $N_2 : \mathbb{R}^n \to \mathbb{R}^+$ such that

(i) If $N_1(x)N_2(x) = 0$, then $0 \in \partial_\varepsilon f(x)$,

(ii) For every $x \in \mathbb{R}^n$ such that $0 \notin \partial_\varepsilon f(x)$,

there exist a $\rho(x) > 0$ and $b_i(x) > 0$, $i = 1,2$, such that for all $x' \in B(x,\rho(x))$

$$N_1(x') \geq b_1(x)\,, \tag{2.8a}$$

$$N_2(x') \leq b_2(x)\,. \tag{2.8b}$$

Now consider the process (2.2a) (2.2b) and suppose that for $i = 0,1,2,\ldots,$

$$d_\varepsilon^0 f(x_i;h_i) \leq -N_1(x_i)\,, \tag{2.8c}$$

$$\|h_i\| \leq N_2(x_i)\,. \tag{2.8d}$$

Then, given any $\bar{x}$ such that $0 \notin \partial_\varepsilon f(\bar{x})$, there exists a $\bar{k} \in \mathbb{N}^+$ such that for all $x_i \in B(\bar{x},\rho(\bar{x}))$ for all $\lambda \in [0,\beta^{\bar{k}}]$

$$f(x_i+\lambda h_i) - f(x_i) \leq \lambda\alpha d_\varepsilon^0 f(x_i;h_i) \leq -\lambda\alpha\, b_1(\bar{x}). \tag{2.9}$$

Proof. Let $\bar{x} \in \mathbb{R}^n$ be such that $0 \notin \partial_\varepsilon f(\bar{x})$. Let $\bar{k} \in \mathbb{N}^+$ be such that $\beta^{\bar{k}} b_2(\bar{x}) \leq \varepsilon$. Then, for all $x_i \in B(\bar{x},\rho(\bar{x}))$ and for all $\lambda \in [0,\beta^{\bar{k}}]$, $(x_i+\lambda h_i) \in B(x_i,\varepsilon)$ and hence for all such x_i and λ,

$$\begin{aligned} d^0 f(x_i+\lambda h_i;h_i) &= \max_{\xi\in\partial f(x_i+\lambda h_i)} \langle \xi,h_i\rangle \\ &\leq \max_{\xi\in\partial_\varepsilon f(x_i)} \langle \xi,h_i\rangle \\ &= d_\varepsilon^0 f(x_i;h_i) \\ &\leq -N_1(x_i) \leq -b_1(\bar{x}). \end{aligned} \tag{2.10}$$

The desired result now follows from Fact 1.1(g). ¤

Corollary 2.1 (Conceptual Algorithms): Let $\varepsilon > 0$ be given and suppose that the assumptions in Theorem 2.2 hold. Then any accumulation point $\hat{x}$ of a sequence $\{x_i\}_{i=0}^\infty$ constructed by an algorithm of the form (2.2a,b) with $d_\varepsilon^0 f(x_i,h_i) \leq \delta_i \leq -N_1(x_i)$ satisfies $0 \in \partial_\varepsilon f(\hat{x})$.

Proof: Suppose that $x_i \overset{K}{\to} \hat{x}$, with $K \subset \{0,1,2,\ldots\}$ and that $0 \notin \partial f(\hat{x})$. Then, by Theorem 2.2, there exists an i_0 and a $\hat{k} \in \mathbb{N}^+$ such that for all $i \geq i_0$, $i \in K$, $\lambda_i \geq \beta^{\hat{k}}$ and

$$\begin{aligned} f(x_{i+1}) - f(x_i) &\leq \lambda_i\alpha d_\varepsilon^0 f(x_i;h_i) \\ &\leq \lambda_i\alpha\delta_i \\ &\leq -\beta^{\hat{k}}\alpha b_1(\hat{x}). \end{aligned} \tag{2.11}$$

Now, $\{f(x_i)\}$ is monotonically decreasing and $x_i \overset{K}{\to} \hat{x}$, hence, since $f(\cdot)$ is continuous, $f(x_i) \to f(\hat{x})$. But this contradicts (2.11) and hence we are done. ¤

The simplest algorithm in the class considered in Theorem (2.2) can be viewed as an "ε-smeared" steepest descent method. It sets

$$h_i = h_\varepsilon(x_i) \triangleq -Nr(\partial_\varepsilon f(x_i)) \triangleq \arg\min\{\|h\| \mid h \in \partial_\varepsilon f(x_i)\} \tag{2.12}$$

and

$$\delta_i = -\|h_i\|^2. \tag{2.12b}$$

Hence

$$d_\varepsilon f(x_i; h_i) = -\|h_i\|^2. \tag{2.13}$$

Setting $N_1(x_i) = \|h_i\|^2$, we see that $N_1(\cdot)$ is lower semicontinuous (l.s.c.) because $\partial_\varepsilon f(x_i)$ is u.s.c. (see proof in [P7]). Next, if we define $N_2(x)$ by

$$N_2(x) = \arg\max\{\|h\| \mid h \in \partial_\varepsilon f(x)\}, \tag{2.14}$$

we see that $\|h_i\| \leq N_2(x_i)$ and that $N_2(\cdot)$ is u.s.c. because $\partial_\varepsilon f(\cdot)$ is u.s.c. (see proof in [P7]). Hence we can set $b_1(x) = N_1(x)/2$ and $b_2(x) = 2N_2(x)$ to show that this algorithm satisfies the assumptions of Theorem 2.2.

Obviously, we would prefer to have algorithms which generate accumulation points $\hat{x}$ such that $0 \in \partial f(\hat{x})$ rather than $0 \in \partial_\varepsilon f(\hat{x})$, with $\varepsilon > 0$. Hence, it is necessary to propose at least one ε-reduction scheme. The most natural thing to do is to reduce ε as x_i approaches a stationary point. This fact is not postulated in the theorem below, but unless it holds it is not possible to find a function $N_1(\cdot)$.

Theorem 2.3 (Conceptual Algorithms): Suppose that there exist three functions N_1, N_2, $N_3 : \mathbb{R}^n \to \mathbb{R}^+$ such that

(i) If $N_1(x)N_2(x)N_3(x) = 0$, then $0 \in \partial f(x)$.

(ii) For every $x \in \mathbb{R}^n$ such that $0 \notin \partial f(x)$, there exist a $\rho(x) > 0$ and $b_i(x) > 0$ $i = 1,2,3$, such that for all $x' \in B(x,\rho(x))$

$$N_1(x') \leq b_1(x) , \tag{2.15a}$$

$$N_2(x') \leq b_2(x) , \tag{2.15b}$$

$$N_3(x') \geq b_3(x) . \tag{2.15c}$$

Now consider the process (2.2a)(2.2b) and suppose that for $i = 0,1,2,\ldots,$

$$d^0_{N_3(x_i)} f(x_i;h_i) \leq -N_1(x_i) , \tag{2.15d}$$

$$\|h_i\| \leq N_2(x_i) . \tag{2.15e}$$

Then, given any $\bar{x}$ such that $0 \notin \partial f(\bar{x})$, there exists a $\bar{k} \in \mathbb{N}^+$ such that for all $x_i \in B(\bar{x},\rho(\bar{x}))$, for all $\lambda \in [0,\beta^{\bar{k}}]$,

$$f(x_i+\lambda h_i) - f(x_i) \leq \lambda\alpha d^0_{N_3(x_i)} f(x_i;h_i) \leq -\lambda\alpha b_1(\bar{x}) . \tag{2.16}$$

Furthermore, any accumulation point $\hat{x}$ of a sequence $\{x_i\}_{i=0}^N$ constructed by an algorithm of the form (2.2a,b) with $\delta_i = d^0_{N_3(x_i)} f(x_i;h_i)$ satisfies $0 \in \partial f(\hat{x})$. ¤

We omit a proof of this theorem since it is obtained by a trivial modification of the proofs of Theorem 2.2 and Corollary 2.1.

We shall now exhibit a natural candidate for $N_3(x)$ in extending the "ε-smeared" steepest descent method to one with an adjustable ε.

Thus, let $\nu \in (0,1)$, $\varepsilon_0 > 0$, $\delta > 0$ be given.

Let

$$E \triangleq \{\varepsilon | \varepsilon = \varepsilon_0 \nu^k,\ k \in \mathbb{N}^+\} \cup \{0\}. \tag{2.17}$$

Next, for any $\varepsilon \geq 0$, let

$$h_\varepsilon(x) \triangleq -\mathrm{Nr}(\partial_\varepsilon f(x)) \triangleq -\arg\min\{\|h\|^2 | h \in \partial_\varepsilon f(x)\}. \tag{2.18}$$

Then we define $\varepsilon : \mathbb{R}^n \to \mathbb{R}^1$ by

$$\varepsilon(x) \triangleq \max\{\varepsilon \in E | \|h_\varepsilon(x)\|^2 \geq \delta\varepsilon\}. \tag{2.19}$$

Proposition 2.1(a): For every $\bar{x} \in \mathbb{R}^n$ such that $0 \notin \partial f(\bar{x})$, there exist a $\rho_3(\bar{x})$ such that

$$\varepsilon(x_i) \geq \nu\, \varepsilon(\bar{x}) > 0 \quad \forall\ x_i \in B(\bar{x}, \rho_3(\bar{x})). \tag{2.20}$$

b) If $x_j \to \hat{x}$ as $j \to \infty$ with $0 \in \partial f(\hat{x})$ then $\varepsilon(x_j) \to \varepsilon(\hat{x}) = 0$ as $j \to \infty$.

Proof: (a) Let $\bar{x}$ be such that $0 \notin \partial f(\bar{x})$. Then, since $\partial f(\cdot)$ is u.s.c. there exists an $\varepsilon_1 > 0$ such that $\|h_{\varepsilon_1}(\bar{x})\|^2 \geq \frac{1}{2}\|h_0(\bar{x})\|^2 > 0$. Hence, since $\varepsilon' < \varepsilon''$ implies that $\|h_{\varepsilon'}(\bar{x})\|^2 \geq \|h_{\varepsilon''}(\bar{x})\|^2$, it follows that

$$\varepsilon(\bar{x}) \geq \max\{\varepsilon \in E |\ \varepsilon \leq \min\{\varepsilon_1, \tfrac{1}{2\delta}\|h_0(\bar{x})\|^2\}\} > 0\ . \tag{2.21}$$

Next, since by the maximum theorem in [B1], $\|h_{\varepsilon(\bar{x})}(\cdot)\|^2$ is l.s.c., and $\|h_{\varepsilon(\bar{x})}(\bar{x})\|^2 \geq \delta\varepsilon(\bar{x})$, there exists a $\rho_3(\bar{x}) > 0$ such that

$$\|h_{\nu\varepsilon(\bar{x})}(x_i)\|^2 \geq \|h_{\varepsilon(\bar{x})}(x_i)\|^2 \geq \delta\nu\varepsilon(x) \text{ for all } x_i \in B(\bar{x}, \rho_3(x)) \tag{2.22}$$

and hence (2.20) follows directly.

(b) Suppose that $0 \in \partial f(\hat{x})$. Then $\|h_0(x)\|^2 = 0$ and for any $\varepsilon > 0$ $\|h_\varepsilon(\hat{x})\|^2 = 0$. Hence $\varepsilon(\hat{x}) = 0$. Next, suppose that $x_j \to \hat{x}$ as $j \to \infty$ and

that $\overline{\lim}\ \varepsilon(x_j) > 0$, i.e. for some $K \subset \{0,1,2,\ldots\}$ and some $\hat{\varepsilon} > 0$ $\varepsilon(x_j) \geq \hat{\varepsilon} > 0$ for all $j \in K$. Since we must have $\hat{x} \in B(x_j,\hat{\varepsilon})$ for all j sufficiently large, we must have that $0 \in \partial_{\hat{\varepsilon}} f(x_j)$, for all j sufficiently large and hence $\|h_{\varepsilon(x_j)}(x_j)\|^2 = 0 < \varepsilon(x_j)$ for all $j \in K$ sufficiently large. But this contradicts the definition of $\varepsilon(x_j)$ and hence we are done. ¤

The final version of the progressively-smeared steepest descent method is sufficiently important to be stated formally:

Algorithm 2.1 (Conceptual).

Parameters: $\alpha, \beta, \nu \in (0,1)$, $\varepsilon_0 > 0$, $\delta > 0$.

Data: $x_0 \in \mathbb{R}^n$.

Step 1: Set $i = 0$.

Step 2: Compute $h_i \triangleq h_{\varepsilon(x_i)}(x_i)$.

Step 3: Compute

$$\lambda_i = \arg\max\{\beta^k \mid f(x_i+\beta^k h_i) - f(x_i) \leq \alpha\beta^k d^0_{\varepsilon(x_i)}(x_i;h_i)\}. \tag{2.23}$$

Step 4: Set $x_{i+1} = x_i + \lambda_i h_i$, set $i = i + 1$ and go to step 2. ¤

Theorem 2.4: Suppose that $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithy 2.1. Then any accumulation point $\hat{x}$ of $\{x_i\}$ (if it exists) satisfies $0 \in \partial f(\hat{x})$.

Proof: We only need to show that the assumptions of Theorem 2.3 are satisfied. Clearly, we must set $N_3(x) = \varepsilon(x)$ and by Proposition 2.1, it has the required properties. Next, we set $N_1(x) \triangleq \|h_{\varepsilon(x)}(x)\|^2$. Then the required properties of $N_1(\cdot)$ follow from those of $\varepsilon(\cdot)$ (with $\rho_1(x) = \rho_3(x)$) and, by inspection,

$$d^0_{\varepsilon(x)} f(x, h_{\varepsilon(x)}(x)) = N_1(x) \tag{2.24}$$

Finally, we set $N_2(x) \triangleq \arg\max\{\|h\| \mid h \in \partial_{\varepsilon_0} f(x)\}$. Since $N_2(\cdot)$ is u.s.c. by the maximum theorem in [B1], we are done. ¤

Next we turn to implementable algorithms. These are characterized by the fact that they approximate the sets $\partial_\varepsilon f(x)$ by means of finite operations while retaining a great resemblance to the conceptual algorithms from which they are derived. It does not appear to be possible to construct a truly useful general convergence theorem of the form of Theorem 2.3 for such algorithms. Instead, it seems simplest to use a minor modification of theorem (1.3.10) in [P8], as follows.

Theorem 2.5: Consider algorithms of the form (2.2a,b). If for every $\bar{x} \in \mathbb{R}^n$ such that $0 \notin \partial f(\bar{x})$ there exist a $\bar{k} \in \mathbb{N}^+$, a $\bar{\delta} > 0$ and a $\bar{\rho} > 0$ such that for all $x_i \in B(\bar{x}, \bar{\rho})$,

$$f(x_i + \beta^{\bar{k}} h_i) - f(x_i) \leq -\alpha\beta^{\bar{k}}\delta_i \leq -\alpha\beta^{\bar{k}}\bar{\delta} . \tag{2.24}$$

Then any accumulation point $\hat{x}$ of a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by such an algorithm satisfies $0 \in \partial f(\hat{x})$.

Proof: Suppose $x_i \overset{K}{\to} \hat{x}$ and $0 \notin f(\hat{x})$. Then there exists an i_0 such that for all $i \in K$, $i \geq i_0$, $\lambda_i \geq \beta^{\bar{k}}$ and hence

$$f(x_{i+1}) - f(x_i) \leq -\alpha\beta^{\bar{k}}\bar{\delta} \quad \forall\, i \geq i_0,\ i \in K . \tag{2.25}$$

But $\{f(x_i)\}$ is monotonically decreasing and $f(\cdot)$ is continuous; hence $f(x_i) \to f(\hat{x})$ as $i \to \infty$. But, clearly, this contradicts (2.25) and we are done. ¤

At the present time, we only know how to construct implementable algorithms for optimization problems in which the function $f(\cdot)$ is semi-smooth (see [M1]).

Definition 2.3 [M1]: A locally Lipschitz continuous function $f(\cdot)$ is said to be semi-smooth if it is directionally differentiable and if for any $x, h \in \mathbb{R}^n$ and for any sequences $\{\lambda_k\} \subset \mathbb{R}^+$, $\{z_k\}$, $\{v_k\} \subset \mathbb{R}^n$ such that $\lambda_k \to 0$, $(1/\lambda_k)v_k \to 0$ and $z_k \in \partial\psi(x+\lambda_k h+v_k)$, the sequence $\{\langle z_k, h\rangle\}$ converges to $df(x;h)$. ¤

From our point of view, the most important property of semi-smooth functions, which does not appear in the definition, is the following one:

Proposition 2.2: Suppose that $f : \mathbb{R}^n \to \mathbb{R}^1$ is semi-smooth. Then, given any x, h, $\{\lambda_k\}$, $\{v_k\}$ as in Definition 2.3,

$$\lim_{k\to\infty} df(x+\lambda_k h+v_k;h) = df(x;h) \tag{2.25}$$

¤

We assume, until the end of this section, that $f(\cdot)$ is semi-smooth. We are now ready to construct an implementation for Algorithm 2.1, which satisfies the assumptions of Theorem 2.5. The implementation is based on the following observations derived from results of Lemarechal [L2] and Wolfe [W1, W2]. Suppose that $x_i \in \mathbb{R}^n$, $\varepsilon > 0$ are given and that $0 \notin \partial_\varepsilon f(x_i)$. Let $Y_s \subset \partial_\varepsilon f(x_i)$ be the convex hull of a finite number of points in $\partial_\varepsilon f(x)$ and let

$$\eta_s = -Nr(Y_s) \tag{2.26}$$

Now, let $k_s \in \mathbb{N}^+$ be such that

$$\beta\varepsilon < \beta^{k_s}\|\eta_s\| \leq \varepsilon . \tag{2.27}$$

Then, either

$$f(x_i+\beta^{k_s}\eta_s) - f(x_i) \leq -\alpha\beta^{k_s}\|\eta_s\|^2 \leq -\alpha\beta^{k_s}\|N_r(\partial_\varepsilon f(x_i)\|^2 \tag{2.28}$$

holds or not. If (2.28) does hold, then $h_i = \eta_s$ turns out to be an adequate approximation to $\eta_\varepsilon(x_i)$, as far as convergence is concerned. If (2.28) does not hold, then there must be a point $\bar{\mu} \in [0,\beta^{k_s}]$ such that

$$f(x_i+\bar{\mu}\eta_s) - f(x_i) = -\bar{\mu}\alpha\|\eta_s\|^2 \tag{2.29}$$

and

$$df(x_i+\bar{\mu}\eta_s) \geq -\alpha\|\eta_s\|^2. \tag{2.30}$$

Now suppose that $\mu_j \in [0,\beta^{k_s}], j = 1,2,\ldots,$ are such that $\mu_j \not\to \bar{\mu}$ and that $y_j \in \partial f(x_i+\mu_j\eta_s)$, for $j = 1,2,\ldots$. Then, because $f(\cdot)$ is semi-smooth,

$$\langle y_j,\eta_s\rangle \to df(x_i+\mu\eta_s) \text{ as } j \to \infty , \tag{2.31}$$

and, consequently, given a $\bar{\alpha} \in (\alpha,1)$, there exists a j_0 such that

$$\langle y_j,\eta_s\rangle \geq -\bar{\alpha}\|\eta_s\|^2 \ \forall j \geq j_0 . \tag{2.32}$$

We see that if we set $Y_{s+1} = co(Y_s \cup \{y_j\})$, $\eta_{s+1} = -Nr(Y_{s+1})$ is smaller then η_s in norm. We can now replace η_s by η_{s+1} and return to the test in (2.28), etc. This cycle of operations cannot continue indefinitely, because, as shown in [M1, P7], if $s \to \infty$ than $\eta_s \to 0$, which contradicts the obvious fact that $\eta_s \geq h_s(x_i) > 0$. Hence the test (2.28) will be passed in a finite number of operations. Note also that β^{k_s} is locally (w.r.t. x)

bounded both from below and from above. Hence the convergence of the algorithm below is very easily deduced from the preceeding results. Note that the algorithm below uses a bisection procedure for finding $\bar{\mu}$ and for constructing the μ_j.

Algorithm 2.2.

Parameters: $\varepsilon_0 > 0$, $\alpha,\beta,\ \nu \in (0,1)$, $\bar{\alpha} \in (\alpha,1)$.

Data: $x_0 \in \mathbb{R}^n$.

Step 0: Set $i = 0$.

Step 1: Set $\varepsilon = \varepsilon_0$, $s = 0$.

Step 2: Compute $Y_s \subset \partial_\varepsilon f(x_i)$, a convex hull of a finite number of points in $\partial_\varepsilon f(x_i)$.

Step 3: Compute $\eta_s = -Nr(Y_s)$ and $k_s \in \mathbb{N}^+$ such that $\beta\varepsilon \leq \beta^{k_s}\|\eta_s\| \leq \varepsilon$.

Step 4: If $\|\eta_s\| < \varepsilon$, set $\varepsilon = \nu\varepsilon$ and go to step 2.

Step 5: If

$$f(x_i+\beta^{k_s}\eta_s) - f(x_i) \leq -\alpha\beta^{k_s}\|\eta_s\|^2, \tag{2.33a}$$

(i) set $h_i = \eta_s$ and compute the smallest $k_i \in \mathbb{N}^+$ such that

$$f(x_i+\beta^{k_i}h_i) - f(x_i) \leq -\alpha\beta^{k_i}\|h_i\|^2; \tag{2.33b}$$

(ii) set $x_{i+1} = x_i + \beta^{k_i}h_i$;

(iii) set $i = i+1$;

(iv) go to step 1.

Step 6: Set $j = 0$.

Step 7: Set $\ell_0 = 0$, $r_0 = \beta^{k_s}\|\eta_s\|^2$, $\mu_0 = r_0/2$.

Step 8: Compute a $y_{j+1} \in \partial f(x_i + r_j \eta_s)$.

Step 9: If

$$\langle y_{j+1}, \eta_s \rangle \geq -\gamma \|\eta_s\|^2 . \tag{2.34}$$

Set

$$Y_{s+1} = \text{co}(\{y_{j+1}\} \cup Y_s) , \tag{2.35}$$

set s = s+1 and go to step 3.

Step 10: If

$$f(x_i + \mu_j \eta_s) - f(x_i) \geq -\alpha \mu_j \|\eta_s\|^2 , \tag{2.36}$$

set $r_{j+1} = \mu_j$, $\ell_{j+1} = \ell_j$, $\mu_{j+1} = (r_{j+1} + \ell_{j+1})/2$.

Else set $r_{j+1} = r_j$, $\ell_{j+1} = \mu_j$, $\mu_{j+1} = (r_{j+1} + \ell_{j+1})/2$.

Step 11: Set j = j+1 and go to step 8. ¤

Theorem 2.6: a) If Algorithm 2.2 generates a finite sequence $\{x_i\}_{i=0}^{N}$, jamming up at x_N, then $0 \in \partial f(x_N)$. b) If Algorithm 2.2 generates an infinite sequence $\{x_i\}_{i=0}^{\infty}$ then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $0 \in \partial f(\hat{x})$. ¤

The success of Algorithm 2.2 depends on the following fact, due to Wolfe [W1, W2] (see also [P7]).

Proposition 2.3: Let S' be a compact, convex subset of a compact convex set S and let $\bar{\alpha} \in (0,1)$. Let h' = Nr(S') and let $g \in S$ be such that

$$\langle g, h' \rangle \leq \bar{\alpha} \|h'\|^2 \tag{2.37a}$$

Then h" = Nr(co{g},S'}) satisfies

$$\|h''\|^2 \leq \max\{\bar{\alpha},\ 1 - (1-\bar{\alpha})^2\|h'\|^2/4C^2\}\|h'\|^2 \qquad (2.37b)$$

where $C \geq \max\{\|g\| \mid g \in S\}$. ¤

Proof of Theorem 2.6: a) Suppose that the sequence $\{x_i\}$ is finite with the algorithm jamming up at x_N, cycling indefinitely in one of the loops defined by steps 2 to 4 or steps 3 to 9 or steps 8 to 11. Suppose that $0 \notin \partial f(x_N)$.

(i) consider the loop defined by steps 2 to 4. Since $0 \notin \partial f(x_N)$, $\varepsilon(x_N) > 0$ (see (2.19)) and hence for all $\varepsilon \geq \varepsilon(x_N)$, $Y_s \subset \partial_\varepsilon f(x_N)$, $\|Nr(Y_s)\| \geq \|Nr(\partial_\varepsilon f(x_N))\| \geq \|Nr(\partial_{\varepsilon(x_N)} f(x_N))\| \geq \varepsilon(x_N) > \varepsilon$ and hence no infinite cycling can occur in this loop.

(ii) consider the loop defined by step 8 to 11. This loop is always finite because $f(\cdot)$ is semi-smooth and (2.33a) is not satisfied.

(iii) consider the loop defined by steps 3 to 9. Since $0 \notin \partial f(x_N)$, $\varepsilon \geq \varepsilon(x_N)$ while in this loop. Hence by Proposition 2.3,

$$\|\eta_{s+1}\| \leq \max\{\bar{\alpha},\ 1 - (1-\bar{\alpha})\|\eta_s\|^2/4C^2\}\|\eta_s\|^2 \qquad (2.36)$$

where $C = \max\{\|\eta\| \mid \eta \in \partial_{\varepsilon_0} f(x_N)\}$. Since $\|\eta_s\| \geq \varepsilon \geq \varepsilon(x_N)$ for all s, it is clear from (2.36) that the sequence $\{\eta_s\}$ must be finite, i.e. the loop defined by steps 3 to 9 is exited after a finite number of operations. Consequently, the algorithm jams up at x_N only if $0 \in \partial f(x_N)$. b) Now suppose that the sequence $\{x_i\}$ is infinite. Suppose that $x_i \overset{K}{\to} \hat{x}$, with $K \subset \{0,1,2,\ldots\}$ and that $0 \notin \partial f(\hat{x})$. Then, by Proposition 2.1, there exists an i_0 such that for all $i \in K$, $i \geq i_0$, $\varepsilon(x_i) \geq \nu\varepsilon(\hat{x}) > 0$. Consequently, for all $i \notin K$, $i \geq i_0$ (2.33a) is satisfied with $\|\eta_s\| \geq \nu\varepsilon(\hat{x})$ and $\beta^{k_s}\|\eta_s\| \geq \beta\varepsilon(x_i) \geq \beta\nu\varepsilon(\hat{x})$. Hence, by (2.33b), for all $i \in K$, $i \geq i_0$,

$$f(x_{i+1}) - f(x_i) \leq -\alpha\beta^{k_i}\|h_i\|^2 \leq -\alpha\beta(\nu\varepsilon(\hat{x}))^2. \tag{2.37}$$

Now $f(x_i) \overset{K}{\to} f(\hat{x})$ by continuity and $\{f(x_i)\}$ is monotonic decreasing. Hence, we must have $f(x_i) \to f(\hat{x})$, which contradicts (2.37). This completes our proof. ¤

3. Constrained Optimization: Conceptual Algorithms

We begin by examining the easiest case, viz., problems of the form

$$\min\{f(x) | g^j(x) \leq 0,\ j \in \underline{m}\} \tag{3.1}$$

where $f, g^j : \mathbb{R}^n \to \mathbb{R}^1$ are locally Lipschitz continuous. For the purpose of conceptual algorithms, it is convenient to define the function

$$\psi(x) \triangleq \max_{j \in \underline{m}} g^j(x) \tag{3.2}$$

and to treat problem (3.1) in the simpler form

$$\min\{f(x) | \psi(x) \leq 0\} \tag{3.3}$$

In implementable algorithms, since we may not be able to obtain a formula for the set $\partial_\varepsilon \psi(x)$, we may have to use the possibly bigger set

$$M_\varepsilon(x) \triangleq \text{co}\{\bigcup_{j \in I_\varepsilon(x)} \partial_\varepsilon g^j(x)\} \tag{3.4a}$$

with

$$I_\varepsilon(x) \triangleq \{j \in \underline{m} | g^j(x) \geq \psi(x) - \varepsilon\}. \tag{3.4b}$$

It is quite easy to construct an appropriate counterpart to Theorem 2.3, for algorithms which generate sequences $\{x_i\}$ by a construction of the phase I - phase II feasible directions type [P3], using parameters $\alpha, \beta \in (0,1)$ viz:

$$x_{i+1} = x_i + \lambda_i h_i, \quad i = 0,1,2,... \tag{3.5a}$$

$$\lambda_i = \begin{cases} \arg\max\{\beta^k \mid \\ \quad \psi(x_i+\beta^k h_i) - \psi(x) \leq \alpha\beta^k\delta_i < 0\} \text{ if } \psi(x_i) > 0; \\ \arg\max\{\beta^k \mid \\ \quad f(x_i+\beta^k h_i) - f(x_i) \leq \alpha\beta^k\delta_i < 0;\ \psi(x_i+\beta^k h_i) \leq 0\} \text{ if } \psi(x_i) \leq 0 \end{cases} \tag{3.5b}$$

Since "ε-smearing" was needed for the unconstrained case, it is a foregone conclusion that it is also needed for the constrained case and we shall not go into any further justifications of the case of "ε-smearing." Also, for the phase I part of the algorithms to work we need the following

Assumption 3.1: For all $x \in \mathbb{R}^n$ such that $\psi(x) \geq 0$, $0 \notin \partial\psi(x)$. ¤

This assumption ensures that a feasible point can be computed by means of an unconstrained optimization algorithm in a finite number of iterations.

Theorem 3.1 (Conceptual Algorithms):

1. Suppose that Assumption 3.1 holds.
2. Suppose that there exist three functions $N_1, N_2, N_3 : \mathbb{R}^n \to \mathbb{R}^+$ such that

(i) If $N_1(x)N_2(x)N_3(x) = 0$, then
either $\psi(x) = 0$ and $0 \in \text{co}(\partial f(x) \cup \partial\psi(x))$;
or $\psi(x) < 0$ and $0 \in \partial f(x)$.

(ii) For every $x \in \mathbb{R}^n$ such that $N_1(x)N_2(x)N_3(x) > 0$, there exist a $\rho(x) > 0$ and $b_i(x) > 0$, $i = 1,2,3$, such that for all $x' \in B(x,\rho(x))$

$$N_1(x') \geq b_1(x), \tag{3.7a}$$

$$N_2(x') \leq b_2(x), \tag{3.7b}$$

$$N_3(x') \geq b_3(x). \tag{3.7c}$$

Now consider the process (3.5a), (3.5b) and suppose that for all i,

$$d^0_{N_3(x_i)}\psi(x_i;h_i) \leq \delta_i \leq -N_1(x_i), \text{ if } \psi(x_i) \geq -N_3(x_i), \tag{3.7d}$$

$$d^0_{N_3(x_i)}f(x_i;h_i) \leq \delta_i \leq -N_1(x_i) \text{ if } \psi(x_i) \leq 0, \tag{3.7e}$$

$$\|h_i\| \leq N_2(x_i). \tag{3.7f}$$

If $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence constructed by this process, then any accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $\psi(\hat{x}) \leq 0$ and $0 \in \partial f(\hat{x})$ if $\psi(\hat{x}) < 0$, otherwise $0 \in co\{\partial f(\hat{x}) \cup \partial\psi(\hat{x})\}$.

Proof: We note that we can distinguish between two cases: a) $\psi(x_i) > 0$ for all i, and b) there exists an i_0 such that $\psi(x_i) \leq 0$ for all $i \geq i_0$.
a) Suppose that $\psi(x_i) > 0$ for all i, that $x_i \overset{K}{\to} \hat{x}$, with $K \subset \{0,1,2,3,\ldots\}$, and that $N_1(\hat{x})N_2(\hat{x})N_3(\hat{x}) > 0$. Then, the process (3.5a,b) reduces to the one considered in Theorem 2.3, and hence we conclude that $\psi(x_i) \downarrow -\infty$. But this contradicts the fact that, by continuity of ψ, $\psi(\hat{x}) \geq 0$, and hence this case is impossible.
b1) Suppose that $\psi(x_i) \leq 0$ for all $i \geq i_0$ and that $x_i \overset{K}{\to} \hat{x}$, with $\psi(\hat{x}) < 0$, and $N_1(\hat{x})N_2(\hat{x})N_3(\hat{x}) > 0$. Then, because of our assumptions, there exist i_1, $\bar{k} \in \mathbb{N}^+$, $i_1 \geq i_0$, such that $\psi(x_i+\beta^{\bar{k}}h_i) \leq 0$ for all $i \geq i$, $i \in K$. Similarly, as in the proof of Theorem 2.3, there exist i_2, $\hat{k} \in \mathbb{N}^+$, with $i_2 \geq i_1$ and $\hat{k} \geq \bar{k}$, such that $\lambda_i \geq \beta^{\hat{k}}$ for all $i \geq i_2$, $i \in K$. Hence, for all $i \in K$, $i \geq i_2$

$$f(x_{i+1}) - f(x_i) \leq \alpha\beta^{k_i}\delta_i \leq -\alpha\beta^{\hat{k}}b_3(\hat{x}) < 0. \tag{3.8}$$

But $f(x_i) \searrow$ for $i \geq i_0$ and hence (3.8) implies that $f(x_i) \searrow -\infty$, which contradicts our assumption that $x_i \to \hat{x}$. Hence this case is not possible.

b2) Suppose that $\psi(x_i) \leq 0$ for all $i \geq i_0$ and that $x_i \overset{K}{\to} \hat{x}$, with $\psi(\hat{x}) = 0$ and $N_1(\hat{x})N_2(\hat{x})N_3(\hat{x}) > 0$. Then our assumptions lead us to the conclusion that there exists an $i_1 \geq i_0$ and a $\hat{k} \in \mathbb{N}^+$ such that

$$f(x_i+\beta^{\hat{k}}h_i) - f(x_i) \leq \alpha\beta^{\hat{k}}d^0_{N_3(x_i)}f(x_i;h_i) \leq \alpha\beta^{\hat{k}}\delta_i \tag{3.9a}$$

$$\psi(x_i+\beta^{\hat{k}}h_i) - \psi(x_i) \leq \alpha\beta^{\hat{k}}d^0_{N_3(x_i)}\psi(x_i;h_i) \leq \alpha\beta^{\hat{k}}\delta_i \tag{3.9b}$$

and consequently, $\lambda_i \geq \beta^{\hat{k}}$. Therefore, (3.8) holds for all $i \geq i_1$, $i \in K$ and the contradiction follows exactly as for case b1). We have thus shown that if $x_i \overset{K}{\to} \hat{x}$, then $N_1(\hat{x})N_2(\hat{x})N_3(\hat{x}) = 0$ must hold and hence the desired conclusion follows from assumption (i) on N_1, N_2, N_3.

¤

We are now ready to apply this theorem to two phase I - phase II methods in the class of the ones presented in [P3] for differentiable optimization. We begin with the simpler one. We shall need the following definitions. Let

$$\psi(x)_+ \triangleq \max\{0, \psi(x)\}. \tag{3.10}$$

Let $\varepsilon_0 > 0$ and $\nu \in (0,1)$ be given and let

$$E = \{\varepsilon | \varepsilon = \varepsilon_0\nu^k, k \in \mathbb{N}^+\} \cup \{0\}. \tag{3.11}$$

Next, let $\gamma > 0$ be given and let $\Gamma : \mathbb{R}^n \to \mathbb{R}^1$ be defined by

$$\Gamma(x) \triangleq \exp(-\gamma\psi(x)_+). \tag{3.12}$$

Finally,

for any $\varepsilon \geq 0$, $\delta > 0$, we define

$$\partial_\varepsilon^+ \psi(x) \triangleq \begin{cases} \partial_\varepsilon \psi(x) & \text{if } \psi(x) \geq -\varepsilon \\ \phi & \text{if } \psi(x) < -\varepsilon , \end{cases} \tag{3.13a}$$

$$h_\varepsilon^f(x) \triangleq -Nr(co\{\partial_\varepsilon f(x), \partial_\varepsilon^+ \psi(x)\}) , \tag{3.13b}$$

$$h_\varepsilon^\psi(x) \triangleq -Nr(\partial_\varepsilon \psi(x)) , \tag{3.13c}$$

$$\theta_\varepsilon^1(x) \triangleq -\max\{\|\Gamma(x) h_\varepsilon^f(x)\|^2, \|(1-\Gamma(x) h_\varepsilon^\psi(x)\|^2\} , \tag{3.13d}$$

$$h_\varepsilon^1(x) \triangleq \Gamma(x) h_\varepsilon^f(x) + (1-\Gamma(x)) h_\varepsilon^\psi(x) , \tag{3.13e}$$

$$\varepsilon^1(x) \triangleq \max\{\varepsilon \in \dot{E} | \theta_\varepsilon'(x) \leq -\delta\varepsilon\} . \tag{3.13f}$$

We recognize $h_\varepsilon^\psi(x)$ as a "steepest descent" direction for $\psi(\cdot)$ at an infeasible point, while $h_\varepsilon^f(x)$ is a "usable" feasible direction when x is feasible. The vector $h_\varepsilon(x)$ moves from $h_\varepsilon^\psi(\cdot)$ to $h_\varepsilon^f(\cdot)$ as x moves from the infeasible into the feasible region. This type of construction is the essence of the algorithms presented in [P3] and ensures that the possible increase in cost is kept in check as the feasible region is approached.

Algorithm 3.1 (Conceptual).

Parameters: $\alpha, \beta, \nu \in (0,1)$, ε_0, δ, $\gamma > 0$.

Data: $x_0 \in \mathbb{R}^n$.

Step 0: Set $i = 0$.

Step 1: Compute $h_i = h_{\varepsilon(x_i)}(x_i)$. Stop if $h_i = 0$.

Step 2: If $\psi(x_i) > 0$, compute the largest stepsize β^{k_i}, $k_i \in \mathbb{N}^+$ such that

$$\psi(x_i+\beta^{k_i}h_i) - \psi(x_i) \leq -\alpha\beta^{k_i}\|h_i\|^2. \tag{3.14a}$$

If $\psi(x_i) \leq 0$, compute the largest step size β^{k_i}, $k_i \in \mathbb{N}^+$, such that

$$f(x_i+\beta^{k_i}h_i) - f(x_i) \leq -\alpha\beta^{k_i}\|h_i\|^2 \tag{3.14b}$$

and

$$\psi(x_i+\beta^{k_i}h_i) \leq 0. \tag{3.14c}$$

Step 3: Set $x_{i+1} = x_i + \beta^{k_i}h_i$, set $i = i + 1$ and go to step 1. ¤

To bring this algorithm into correspondence with Theorem 3.1, we define

$$N_1(x) \triangleq -\theta_\varepsilon^1 1_{(x)}(x), \tag{3.15a}$$

$$N_2(x) \triangleq \arg\max\{\|h\| \,|\, h \in \mathrm{co}\{\partial_{\varepsilon_0} f(x),\ \partial^+_{\varepsilon_0}\psi(x)\}, \tag{3.15b}$$

$$N_3(x) \triangleq \varepsilon^1(x), \tag{3.15c}$$

and we set

$$\delta_i \triangleq -\|h_i\|^2 \text{ for } i = 0,1,2,\ldots \tag{3.15d}$$

Lemma 3.1: For every $\varepsilon \geq 0$ and any $x \in \mathbb{R}^n$,

$$\|h_\varepsilon(x)\|^2 \geq -\theta_\varepsilon^1(x) \tag{3.16}$$

Proof: Case 1: Suppose that $\psi(x) < -\varepsilon$. Then $\|h_\varepsilon(x)\|^2 = -\theta_\varepsilon^1(x)$. Hence, consider

Case 2: $\psi(x) \geq -\varepsilon$. Consider the function $g : [0,1] \to \mathbb{R}^1$ defined by

$$g(t) \triangleq \|th_\varepsilon^f(x) + (1-t)h_\varepsilon^\psi(x)\|^2 - (1-t)^2\|h_\varepsilon^\psi(x)\|^2 \tag{3.17}$$

Then $g(0) = 0$, $g(1) = \|h_\varepsilon^f(x)\|^2 \geq 0$ and

$$\frac{d^2}{dt^2} g(t) = 2\{\|h_\varepsilon^f(x) - h_\varepsilon^\psi(x)\|^2 - \|h_\varepsilon^\psi(x)\|^2\}$$

$$= 2\{\|h_\varepsilon^f(x)\|^2 - 2\langle h_\varepsilon^f(x), h_\varepsilon^\psi(x)\rangle\}$$

$$\leq 0, \tag{3.18}$$

because $\langle h_\varepsilon^f(x), h_\varepsilon^\psi(x)\rangle \geq \|h_\varepsilon^f(x)\|^2$, by construction of $h_\varepsilon^f(x)$ and $h_\varepsilon^\psi(x)$. Hence $g(\cdot)$ is concave on $[0,1]$ and, since $g(0) = 0$ and $g(1) \geq 0$, $g(t) \geq 0$ for all $t \in [0,1]$. Consequently,

$$\|h_\varepsilon(x)\|^2 \geq (1-\Gamma(x))^2 \|h_\varepsilon^\psi(x)\|^2. \tag{3.19}$$

Similar reasoning gives that

$$\|h_\varepsilon(x)\|^2 \geq \Gamma(x)^2 \|h_\varepsilon^f(x)\|^2 \tag{3.20}$$

and we are done. ¤

Corollary 3.1: With δ_i defined by (3.15d) and $N_1(x_i)$ defined by (3.15a), we have $\delta_i \leq -N_1(x_i)$ for all i. ¤

Proposition 3.1: Consider the functions $\theta_\varepsilon^1(\cdot)$ defined in (3.13d). (a) For any $x \in \mathbb{R}^n$, if $\varepsilon' > \varepsilon'' \geq 0$, then $\theta_{\varepsilon'}^1(x) \geq \theta_{\varepsilon''}^1(x)$. (b) For any $\varepsilon \geq 0$, $\theta_\varepsilon^1(\cdot)$ is u.s.c.

Proof: a) Since $\varepsilon' > \varepsilon''$ implies that $\partial_{\varepsilon'}\psi(x) \supset \partial_{\varepsilon''}\psi(x)$ and $\partial_{\varepsilon'}f(x) \supset \partial_{\varepsilon''}f(x)$, this part is obvious.

b) Since for any $\varepsilon \geq 0$, $\partial_\varepsilon^+\psi(\cdot)$ and $\partial_\varepsilon f(\cdot)$ are both u.s.c., it follows from the maximum theorem in [B1] that $\|h_\varepsilon^f(\cdot)\|^2$ and $\|h_\varepsilon^\psi(\cdot)\|^2$ are l.s.c. Hence $\theta_\varepsilon^1(\cdot)$ is u.s.c. ¤

Lemma 3.2: For every $\bar{x} \in \mathbb{R}^n$ such that $\theta_0^1(\bar{x}) \neq 0$, $\varepsilon^1(x) > 0$ and there exists a $\bar{\rho} > 0$ such that

$$N_3(x') \triangleq \varepsilon^1(x') \geq \nu\varepsilon^1(\bar{x}) \triangleq b_3(x) > 0 \text{ for all } x' \in B(\bar{x},\bar{\rho}) \tag{3.21}$$

Proof: First, because the set valued maps $\partial f(\cdot)$ and $\partial\psi(\cdot)$ are u.s.c., and $\theta_0^1(x) < 0$, there must exist an $\bar{\varepsilon} \in E$, $\bar{\varepsilon} > 0$, such that $\theta_{\bar{\varepsilon}}^1(\bar{x}) \leq -\delta\bar{\varepsilon}$. Hence $\varepsilon^1(\bar{x}) > 0$. Now, for the sake of contradiction, suppose that there is no $\bar{\rho} > 0$ such that (3.21) holds. Then there must exist a sequence $\{x_i\}$, $x_i \to \bar{x}$ such that

$$\theta^1_{\nu\varepsilon'(x)}(x_i) > -\delta\nu\varepsilon^1(\bar{x}) \text{ for all } i \tag{3.22}$$

Since by Lemma 3.2 $\theta^1_{\nu\varepsilon^1(x)}(\cdot)$ is u.s.c., we conclude from (3.22) that

$$-\delta\nu\varepsilon^1(\bar{x}) \leq \overline{\lim}\ \theta^1_{\nu\varepsilon^1(\bar{x})}(x_i) \leq \theta^1_{\nu\varepsilon^1(\bar{x})}(\bar{x}) \tag{3.23a}$$

But, by Lemma 3.2, $\theta^1_{\varepsilon^1(\bar{x})}(\bar{x}) \geq \theta^1_{\nu\varepsilon^1(\bar{x})}(\bar{x})$ and hence (3.22a) implies that

$$-\delta\varepsilon^1(\bar{x}) < \theta^1_{\varepsilon^1(\bar{x})}(\bar{x}) \tag{3.23b}$$

Which contradicts the definition of $\varepsilon^1(\bar{x})$. ¤

Theorem 3.2: Let $\{x_i\}_{i=0}^{\infty}$ be any sequence constructed by Algorithm 3.1. Then any accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $\psi(\hat{x}) \leq 0$ and $0 \in \text{co}\{\partial f(\hat{x}) \cup \partial\psi(\hat{x})\}$.

Proof: With N_1, N_2, N_3, δ_i defined as in (3.15a) - (3.15d), we see that at any $\bar{x}$ such that $N_1(\bar{x})N_2(\bar{x})N_3(\bar{x}) \neq 0$, By Lemma 3.2, there exists a $\bar{\rho} > 0$ such that $b_1(\bar{x}) = b_3(\bar{x}) = \nu\varepsilon^1(\bar{x}) > 0$ satisfy (3.7a) and (3.7c) for all $x' \in B(\bar{x},\bar{\rho})$. Since $\partial_{\varepsilon_0} f(x)$ and $\partial_{\varepsilon} f(x)$ are both u.s.c., it is clear

that a required $b_2(\bar{x}) > 0$ exists for (3.7b) to hold in $B(\bar{x},\bar{\rho})$. Finally, by Corollary 3.1, we have that $\delta_i \leq -N_1(x_i)$ for all i. Furthermore, Assumption 3.1 and Lemma 3.2 ensure that $N_1(x)N_2(x)N_3(x) = 0$ implies that condition (i) of Theorem 3.1 is satisfied. Consequently, the desired result follows directly from Theorem 3.1. ¤

Our second algorithm has exactly the same structure as Algorithm 3.1 except that h_i is computed by evaluating a different optimality function, $\theta^2_{\varepsilon(x)}(x)$. It is a direct extension of the most efficient phase I-phase II method of feasible directions known. ([P3]) We need the following notation. Given $\gamma > 0$, for any $\varepsilon \geq 0$ and $x \in \mathbb{R}^n$ we define

$$\theta^2_\varepsilon(x) \triangleq \min_{h\in\mathbb{R}^n}\{\tfrac{1}{2}\|h\|^2 + \max\{\langle \xi_f, h\rangle - \gamma\psi_+(x),\ \xi_f \in \partial_\varepsilon f(x);$$

$$\langle \xi_\psi, h\rangle,\ \xi_\psi \in \partial^+_\varepsilon\psi(x)\}\} \tag{3.24a}$$

and

$$h^2_\varepsilon(x) = \arg\min_{h\in\mathbb{R}^n}\{\tfrac{1}{2}\|h\|^2 + \max\{\langle \xi_f, h\rangle - \gamma\psi_+(x),\ \xi_f \in \partial_\varepsilon f(x);$$

$$\langle \xi_\psi, h\rangle,\ \xi_\psi \in \partial^+_\varepsilon\psi(x)\}\}. \tag{3.24b}$$

It follows by duality that when $\psi_+(x) = 0$, for all $\varepsilon \geq 0$, $\theta^1_\varepsilon(x) = \theta^2_\varepsilon(x)$ and $h^2_\varepsilon(x) = h^f_\varepsilon(x)$. Hence, the behavior of the two algorithms can differ only in the infeasible region. We now define

$$\varepsilon^2(x) \triangleq \max\{\varepsilon \in \textstyle\sum | \theta^2_\varepsilon(x) \leq -\delta\varepsilon\} \tag{3.25}$$

where ε and δ are as in (3.13f).

Not surprisingly, the conclusions of Lemma 3.1, Propositions 3.1, Lemma 3.2 and Corollary 3.1 remain valid when $\varepsilon^2(x)$, $h^2_\varepsilon(x)$ and $\theta^2_\varepsilon(x)$ are substituted for $\varepsilon^1(x)$, $h^1_\varepsilon(x)$ and $\theta^1_\varepsilon(x)$ in the appropriate definitions.

Consequently, we may state, without proof the following

Theorem 3.3: Suppose that Algorithm 3.1 is modified so that $h_i = h^2_{\varepsilon_l(x_i)}(x_i)$ in Step 1. If $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence constructed by this modified algorithm then any accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $\psi(\hat{x}) \leq 0$ and $0 \in co\{\partial f(\hat{x}) \cup \partial_0^+ \psi(\hat{x})\}$ ¤

Finally we turn to problems with both inequality and equality constraints, i.e., problems of the form

$$P : \min\{f(x) | g^i(x) \leq 0,\ i \in \underline{m};\ h^j(x) = 0,\ j \in \underline{\ell}\} \tag{3.26}$$

where f, g^i, $i \in \underline{m}$ and h^j, $j \in \underline{\ell}$, from $\mathbb{R}^n$ into $\mathbb{R}$ are all locally Lipschitz continuous. In the differentiable case, i.e. when f, g^i and h^j, $i \in \underline{m}$, $j \in \underline{\ell}$ are all continuously differentiable, there are two major approaches, based on exact penalty functions, for solving (3.26). The first is due to Mayne and Polak([M3]). It replaces the problem P with P_c^1, below, $c > 0$

$$P_c^1 : \min\{f(x) - c \sum_{j \in \underline{\ell}} h^j(x) | g^i(x) \leq 0,\ i \in \underline{m};\ h^j(x) \leq 0,\ j \in \underline{\ell}\} \tag{3.27}$$

and, under mild assumptions, computes a finite $\bar{c}$ which makes P_c^1 and P "locally equivalent" in the vicinity of Kuhn-Tucker points of P for all $c \geq \bar{c}$. The second approach, see e.g. [C5, P14], replaces P with P_c^2 below, $c > 0$,

$$P_c^2 : \min_{x \in \mathbb{R}^n} f_c(x) \tag{3.28a}$$

where

$$f_c(x) \triangleq f(x) + c[\max_{i \in \underline{m}} g^i(x)_+ + \max_{j \in \underline{\ell}} |h^j(x)|] \tag{3.28b}$$

Again, it can be shown that, under mild assumptions, P and P_2^c are

"locally equivalent" for c sufficiently large, in the vicinity of feasible Kuhn-Tucker points of P (see [P14]).

In the nondifferentiable case both approaches tend to break down when equality constraints are present because stationary points of P_2^c which are feasible for P cannot be shown to be also stationary for P. Furthermore, arbitrary feasible points of P may be stationary for P_2^c. Thus, consider the problem P_2^c. Suppose, for simplicity, that there are no inequality constraints in P, and that $\ell = 1$, i.e., that there is only one equality constraint. Then (3.26) and (3.28a) become

$$P : \min_{x \in \mathbb{R}^n} \{f(x) | h(x) = 0\} \tag{3.29}$$

and

$$P_c^2 : \min_{x \in \mathbb{R}^n} \{f(x) + c\{|h(x)|\} \tag{3.30}$$

respectively. Suppose that for some $c > 0$, $\hat{x} \in \mathbb{R}^n$ satisfies the necessary optimality condition for P_c^2, and that $h(\hat{x}) = 0$. Then

$$0 \in \partial f(\hat{x}) + co\{\partial h(\hat{x}) \cup -\partial h(\hat{x})\} \tag{3.31}$$

Now, from (3.31) we would like to conclude that (1.8) holds, i.e., that either

$$0 \in co\{\partial f(\hat{x}) \cup \partial h(\hat{x})\} \tag{3.32a}$$

or

$$0 \in co\{\partial f(\hat{x}) \cup -\partial h(\hat{x})\} \tag{3.32b}$$

While in the differentiable case (3.32a) or (3.32b) follows directly from (3.31), a similar conclusion does not hold in general in the nondifferentiable case, as can be seen from the following example. Let $x = (x^1, x^2)^T \in \mathbb{R}^2$, let $f(x) = -\frac{1}{2} x^1$,

let

$$h(x) = \begin{cases} (x^1)^2 + (x^2)^2 - 5 & \text{if } x^1 \leq 1 \\ x^1 + (x^2)^2 - 5 & \text{if } x^1 \geq 1 \end{cases}$$

and let $\hat{x} = (1,2)^T$. Then $\hat{x}$ is feasible for P in (3.29) and $\partial h(\hat{x}) = co\{(1,4)^T,(2,4)^T\}$. It is reasily seen that for all $c \geq 1$ (3.31) holds, but neither (3.32a) nor (3.32b).

This example shows that when $\partial h(\hat{x})$ is not contained in a one dimensional subspace of $\mathbb{R}^2$ and $h(\hat{x}) = 0$, then $co\{\partial h(\hat{x}) \cup -\partial h(\hat{x})\}$ can be "blown up" by increasing c so that $\hat{x}$ becomes a stationary point for $f_c(\cdot)$, i.e. arbitrary feasible point of P become stationary points of P_2^c. Hence it seems that an exact penalty function method can be generalized to the nondifferentiable case only when the generalized gradients of all the equality constraints are each contained in a one dimensional subspace of $\mathbb{R}^n$, so that $co\{\partial h^j(\hat{x}) \cup -\partial h^j(\hat{x})\}$ does not have an interior point in any multidimensional space. In the presence of inequality constraints alone, exact penalty methods should work, for the following reason. Suppose that $\hat{x}$ satisfies $\psi(\hat{x}) = 0$ and $0 \in \partial f(\hat{x}) + c\partial\psi(\hat{x})$ for some $c > 0$. Then we have that

$$\xi_f + c\alpha\zeta_\psi = 0 \tag{3.33}$$

for some $\xi_f \in \partial f(\hat{x})$, $\xi_\psi \in \partial\psi(\hat{x})$ and $\alpha \in [0,1]$. Consequently,

$$(1+c\alpha)\frac{1}{1+c\alpha}\xi_f + \frac{c\alpha}{1+c\alpha}\xi_\psi = 0 \tag{3.34}$$

i.e. $0 \in co\{\partial(\hat{x}),\partial\psi(\hat{x})\}$. Hence it should be possible to solve P by exact penalty function methods, provided the following assumption holds:

Assumption 3.2 For all $j \in \underline{\ell}$, the functions $h^j(\cdot)$ are continuously differentiable.

For the differentiable case the approach based on P_c^1 is considerably more attractive, since it permits the use of a broad class of algorithms for solving P. However, this advantage is lost for the nondifferentiable case. We will therefore consider here the more traditional approach based on P_c^2. Although it is not possible to precompute a satisfactory penalty $\hat{c}$ for P_c^2, the theory in [P10] on abstract exact penalty methods shows that such a penalty can be computed adaptively, provided an appropriate test function can be constructed. We shall exhibit such a test function for the problems in question.

We now define

$$\eta(x) \triangleq \max_{j\in\ell} |h^j(x)| \tag{3.35a}$$

$$\psi(x)_+ \triangleq \max_{i\in\underline{m}} g^i(x)_+ \tag{3.35b}$$

and

$$\phi(x) \triangleq \max\{\eta(x),\psi(x)_+\} \tag{3.35c}$$

Next we establish a number of properties of the problem P_c^2. The first one is obvious.

Proposition 3.2: Suppose $\hat{x}$ is a local minimizer for P_c^2 such that $\phi(\hat{x}) = 0$. Then $\hat{x}$ is also a local minimizer for P.

Proposition 3.3: Suppose that Assumption 3.2 holds and that $\hat{x} \in \mathbb{R}^n$ is feasible for P, which for some $c > 0$ satisfies

$$0 \in \partial f(\hat{x}) + c \sum_{i=1}^{m} \partial g^i(\hat{x})_+ + c \sum_{j=1}^{\ell} \partial |h(\hat{x})| \tag{3.36}$$

Then $\hat{x}$ satisfies (1.8).

Proof By assumption, there exist: (i) a $\xi_f \in \partial f(\hat{x})$, (ii) $\xi_i \in \partial g^i(\hat{x})$ and a $t_i \in [0,1]$ for all $i \in I(\hat{x})$, (iii) $t_j \in [-1,1]$ for all $j \in \underline{\ell}$, such that

$$\xi_f + c \sum_{i\in I(\hat{x})} t_i \xi_i + c \sum_{j\in \underline{\ell}} t_j \nabla h^j(\hat{x}) = 0. \tag{3.37}$$

By dividing each element of (3.37) by $1+c(\sum_{i\in I(\hat{x})} t_i + \sum_{j\in\underline{\ell}} |t_j|)$ we get (1.8).

Before we can establish the existence of finite penalities, we must invoke the following, commonly used hypothesis.

Assumption 3.3: For every $x \in \mathbb{R}^n$ and any $t_1, t_2, \ldots, t_\ell \in \{-1,1\}$

$$0 \in s_g \sum_{i\in I_0(x)} \partial g^i(x) + s_h \sum_{j\in J(x)} t_j \quad \nabla h^j(x) \tag{3.38}$$

where

$$I_0(x) = \{i \in \underline{m} | g^i(x) = \psi(x)\}$$

$$J(x) \triangleq \{j \in \underline{\ell} | h^j(x)| = \phi(x),$$

$$s_g = \begin{cases} 1 & \text{if } \psi_+(x) = \phi(x) \\ 0 & \text{if } \psi_+(x) < \phi(x) \end{cases}$$

and

$$s_h = \begin{cases} 1 & \text{if } \eta(x) = \phi(x) \\ 0 & \text{if } \eta(x) < \phi(x). \end{cases}$$

We are now ready to establish the existence of exact penalities.

Proposition 3.4: Suppose that $\hat{x}$ is a local minimizer for P. Then there exists a $\hat{c} > 0$ such that

$$0 \in \partial f(\hat{x}) + c \sum_{i\in I(\hat{x})} \partial g^i(x)_+ + c \sum_{j\in\underline{\ell}} \partial |h^j(\hat{x})| \tag{3.39}$$

for all $c > \hat{c}$, i.e. $\hat{x}$ is stationary for P_c.

Proof By Theorem 1.1 and assumption 3.3, there exist

$$\xi_f \in \partial f(\hat{x}),\ \lambda^i \geq 0,\ \xi_{\psi,i} \in \partial g^i(\hat{x}) \cap \partial g^i(\hat{x})_+,\quad i \in I(\hat{x}),$$

and $\lambda^j \in \mathbb{R}$, $j \in \underline{\ell}$, such that

$$\xi_f + \sum_{i\in I(\hat{x})} \lambda^i \xi_{\psi,i} + \sum_{j\in\underline{\ell}} \lambda^j \nabla h^j(\hat{x}) = 0 \tag{3.40}$$

Therefore, for all $c > 0$,

$$\xi_f + c \sum_{i \in I(\hat{x})} \frac{\lambda^i}{c} \xi_{\psi,i} + c \sum_{j \in \underline{\ell}} \frac{\lambda^j}{c} \nabla h^j(\hat{x}) = 0 \qquad (3.41)$$

Obviously, there exists a $\hat{c} > 0$ such that, for all $c \geq \hat{c}$, satisfying $\frac{\lambda^i}{c} < 1$ and $\frac{\lambda^i}{c} \xi_{\psi,i} \in \partial g^i(\hat{x})$, $i \in I(\hat{x})$; and $|\frac{\lambda^j}{c}| < 1$, $j \in \underline{\ell}$. Hence, for $c \geq \hat{c}$, (3.39) follows from (3.41) and the fact that $0 \in \partial |h^j(\hat{x})|$, for all $j \in \underline{\ell}$, and $0 \in \partial g^i(\hat{x})$ for all $i \in I(\hat{x})$. ¤

The following proposition is a direct corollary of Assumption 3.3.

Proposition 3.5: Suppose that $\hat{x} \in \mathbb{R}^n$ is such that $\phi(\hat{x}) > 0$. Then there exists a $\hat{c} > 0$ such that

$$0 \in \partial f(\hat{x}) + c\, S_g \sum_{i \in I_0(\hat{x})} \partial g^i(\hat{x}) + c\, S_h \sum_{j \in J(\hat{x})} \nabla |h^j(\hat{x})|$$

where

$$i_0(x) = \{i \in \underline{m} \mid g^i(\hat{x}) = \psi(\hat{x})\},$$

$$J(\hat{x}) = \{j \in \underline{\ell} \mid |h^j(\hat{x})| = \eta(\hat{x})\},$$

$$s_g = \begin{cases} 1 & \text{if } \psi_+(\hat{x}) = \phi(\hat{x}) \\ 0 & \text{if } \psi_+(\hat{x}) < \phi(\hat{x}) \end{cases}$$

$$s_h = \begin{cases} 1 & \text{if } \eta(\hat{x}) = \phi(\hat{x}) \\ 0 & \text{if } \eta(\hat{x}) < \phi(\hat{x}) \end{cases}$$

Proof: By Assumption 3.3 there exists a $\delta > 0$ such that for every $\xi^i \in \partial g^i(\hat{x})$ with $i \in I_0(\hat{x})$ and every $j \in J(\hat{x})$,

$$S_g \sum_{i \in I_0(\hat{x})} \xi^i + S_h \sum_{j \in J(\hat{x})} \nabla |h^j(\hat{x})| \; > \delta$$

Now, it is clear that proposition 3.5 holds with

$$\hat{c} = \frac{1}{\delta} \cdot \max\{\|\xi_f\| \mid \xi_f \in \partial f(\hat{x})\}.$$

We now construct an exact penalty function method which computes the required penalty parameter c adaptively, making use of the scheme proposed in [P10]. This scheme uses a test function $t_c(\cdot)$ to determine whether c should be increased or not. As in (2.18) and (2.19), we define, for $\varepsilon \geq 0$ and any $x \in \mathbb{R}^n$,

$$h_{c,\varepsilon}(x) \triangleq -Nr(\partial_\varepsilon f_c(x)) \tag{3.42a}$$

and (with $\delta > 0$),

$$\varepsilon_c(x) \triangleq \max\{\varepsilon \in \sum | \ h_{c,\varepsilon}(x)^2 \geq \delta\varepsilon\} \tag{3.42b}$$

Then for any $c > 0$, $x \in \mathbb{R}^n$ we define

$$\theta_c(x) \triangleq -\|h_{c,\varepsilon_c(x)}(x)\|^2 \tag{3.42c}$$

and

$$t_c(x) \triangleq -\varepsilon_c(x) + \frac{1}{c}\phi(x) \tag{3.42d}$$

In accordance with [M3], we therefore propose the following conceptual

Algorithm 3.2:

Parameters: $\alpha, \beta, \nu \in (0,1)$, $\varepsilon_0 > 0$, $\delta > 0$, and a sequence $\{c_j\}_{i=0}^{\infty} \subset \mathbb{R}^+$ $c_j \nearrow \infty$.

Data: $x_0 \in \mathbb{R}^n$

Step 0: Set $i = 0$, $j = 0$.

Step 1: If $t_{c_j}(x_i) > 0$, set $z_j = x_i$ and increase j to the first j^* such that $t_{c_{j^*}}(x_i) \leq 0$. Set $j = j^*$.

Step 2: If $0 \in \partial f_{c_j}(x_i)$, stop. Else compute x_{i+1} by applying Algorithm 2.1 to $f_{c_j}(\cdot)$, from x_i, using the parameters supplied. Set $i = i + 1$ and go to Step 1. ¤

Theorem 3.4:

(i) If $\{z_j\}$ is finite, with last element z_j^*, then either the sequence $\{x_i\}$ is finite and its last element, say x_k satisfies $\phi(x_k) = 0$ and (1.8), or it is infinite and any accumulation point of $\{x_i\}$, say $\hat{x}$, satisfies $\phi(\hat{x}) = 0$ and (1.8).

(ii) If $\{z_j\}$ is infinite, then it has no accumulation points.

Proof (i) Suppose that both $\{x_i\}$ and $\{z_j\}$ are finite, $\{x_i\}$ terminating at x_k. Then for some $j = j^*$, we must have $0 \in \partial f_{c_{j^*}}(x_k)$ and $t_{c_{j^*}}(x_k) \leq 0$. Since $\varepsilon_{c_{j^*}}(x_k) = 0$, it follows that $\phi(x_k) = 0$, and since $0 \in \partial f_{c_{j^*}}(x_k)$ it follows from Proposition 3.3 that

$$0 \in \text{co}\{\partial f(\hat{x}); \partial g^i(\hat{x})_+, i \in I(\hat{x}); \partial |h^j(\hat{x})|, j \in \underline{\ell}\}$$

Next, suppose that $\{x_i\}$ is infinite, with $x_i \xrightarrow{k} \hat{x}$, $k \subset N^+$ and that $\{z_j\}$ is finite, terminating at j^*. Let i_{j^*} be such that $x_{i_{j^*}} = z_{j^*}$. Then for all $i > i_{j^*}$ we have that

$$t_{c_{j^*}}(x_1) = -\varepsilon_{c_{j^*}}(x_i) + \frac{1}{c_{j^*}}\phi(x_i) \leq 0 \tag{3.43}$$

But as in the proof of Theroem 2.1 we have that $\varepsilon_{c_{j^*}}(x_i) \xrightarrow{K} \varepsilon_{c_{j^*}}(\hat{x}) = 0$, and hence from (3.36) and the continuity of ϕ we get that $\phi(\hat{x}) = 0$. Finally, since $\varepsilon_{c_{j^*}}(\hat{x}) = 0$ implies that $0 \in \partial f_{c_{j^*}}(\hat{x})$, it follows that $0 \in \text{co}\{\partial f(x);\ \partial g^i(\hat{x})_+,\ i \in I(\hat{x});\ \partial|h^j(\hat{x})|,\ j \in \underline{\ell}\}$.

(ii) Now, suppose that $\{z_j\}$ is infinite and that $z_j \xrightarrow{K} \hat{x}$ for some subsequence indexed by $K \subset N^+$. Now, $c_j \nearrow \infty$ and $\{z_j\}_{j\in K}$ is compact. Hence $\{\phi(z_j)\}_{j\in K}$ is bounded and therefore $\frac{1}{c_j}\phi(z_j) \xrightarrow[j\to\infty]{K} 0$. Since $t_{c_{j-1}}(z_j) > 0$ for all $j \in K$, we must therefore have that $\varepsilon_{c_{j-1}}(z_j) \xrightarrow[j\to\infty]{K} 0$. Now, because of Assumption 3.1, there exist a $j_0 \in K$ and an $\bar{\varepsilon} > 0$, $\bar{\varepsilon} \in E$ such that for all $j \geq j_0$, $j \in K$

$$\|Nr(\partial_{\varepsilon} f_{c_{j-1}}(z_j))\|^2 > \bar{\varepsilon} \tag{3.44}$$

which shows that $\varepsilon_{c_{j-1}}(z_j) \geq \bar{\varepsilon}$ for all $j \geq j_0$, $j \in K$, which contradicts the fact that $\varepsilon_{c_{j-1}}(z_j) \overset{K}{\to} 0$.

To conclude this discussion, we must point that one could also construct a similar exact penalty function method in which each constraint is penalized individually, by setting

$$f_c(x) \triangleq f(x) + \sum_{i=1}^{\ell+m} c^i g^i(x)_+ \tag{3.45}$$

with $g^{m+j} \triangleq |h^j|$ for $j = 1,2,\ldots\ell$. The penalties c^i must then be increased individually when $t^k_{c^i_j}(x) > 0$, with

$$t^i_c(x) \triangleq \theta_c(x) + \frac{1}{c} g^i(x)_+ \tag{3.46}$$

4. Constrained Optimization: Implementable Algorithms

We shall consider only the problem (3.1) and the implementation of phase I - phase II methods, since the implementation of exact penalty function methods is essentially the same as in Algorithm 2.2.

We shall consider problem 3.1 in the compact form

$$\min\{f(x) | \psi(x) \leq 0\} \tag{4.1}$$

with $f, \psi : \mathbb{R}^n \to \mathbb{R}^1$ locally Lipschitz and semi-smooth. Furthermore, we shall assume that $0 \notin \partial\psi(x)$ for all x such that $\psi(x) \geq 0$. We shall make repeated use of the bisection method described in Section 2 (equs.(2.26)-(2.30)) which can be used (for semi-smooth functions) to find a $\xi \in \partial_{\varepsilon} f(x)$ (or $\xi \in \partial_{\varepsilon}\psi(x)$) such that $\langle \xi, h \rangle \leq \bar{\alpha}\|h\|^2$ whenever $h \in \mathbb{R}^n$ is such that

$$f(x-\lambda h) - f(x) > -\alpha\lambda\|h\|^2 \tag{4.2a}$$

or

$$\psi(x-\lambda h) - \psi(x) > -\alpha\lambda\|h\|^2, \tag{4.2b}$$

with $\lambda\|h\| \leq \varepsilon$ and $0 < \alpha < \bar{\alpha} < 1$.

We now present an implementation of Algorithm 3.1.

Algorithm 4.1 (Implementable)

Data: $\varepsilon_0 > 0$, $\delta > 0$, α, β, $\nu \in (0,1)$, $\bar{\alpha} \in (\alpha,1)$, $x_0 \in \mathbb{R}^n$

Step 0: Set $i = 0$.

Step 1: Set $\varepsilon = \varepsilon_0$.

Step 2: If $\psi(x_i) \geq -\varepsilon$, go to step 7.

CASE 1: $\psi(x_i) < -\varepsilon$.

Step 3: Set $j = 0$ and compute an $h_0^f \in \partial_\varepsilon f(x_i)$.

Step 4: If $\|h_j^f\|^2 < \delta\varepsilon$, set $\varepsilon = \nu\varepsilon$ and go to step 3.

Else, proceed.

Step 5: Set $s_j = \arg\max\{\beta^k | \beta^k \leq (\varepsilon/\|h_j^f\|),\ k \in \mathbb{N}^+\}$.

Step 6: If

$$f(x_i - s_j h_j^f) - f(x_i) \leq -s_j\|h_j^f\|^2, \tag{4.3a}$$

set $h_i = h_j^f$ and go to step 13.

Else, (i) use the bisection method to compute a $\xi_j^f \in \partial_\varepsilon f(x_i)$ such that

$$\langle \xi_j^f, h_j^f \rangle \leq \bar{\alpha}\|h_j^f\|^2, \tag{4.3b}$$

(ii) compute $h_{j+1}^f = \mathrm{Nr}\ \mathrm{co}\{\xi_j^f, h_j^f\}$, set $j = j + 1$ and go to step 4.

Step 7: If $\psi(x_i) > 0$ go to step 14.

CASE 2: $\psi(x_i) \in [-\varepsilon, 0]$.

Step 8: Set $j = 0$. Compute $\xi_0^f \in \partial_\varepsilon f(x_i)$, $\xi_0^\psi \in \partial_\varepsilon \psi(x_i)$ and $h_0^f = \mathrm{Nr}(\mathrm{co}\{\xi_0^f, \xi_0^\psi\})$.

Step 9: If $\|h_j^f\|^2 < \delta\varepsilon$, set $\varepsilon = \nu\varepsilon$ and to to step 2.

Step 10: Set $s_j = \arg\max\{\beta^k \mid \beta^k \leq (\epsilon/\|h_j^f\|),\ k \in \mathbb{N}^+\}$.

Step 11: If

$$\psi(x_i - s_j h_j^f) - \psi(x_i) \leq -s_j \alpha \|h_j^f\|^2, \tag{4.4a}$$

set $h_{j+1}^\psi = h_j^\psi$ and go to step 12.

Else, (i) use the bisection method to compute a $\xi_{j+1}^\psi \in \partial_\varepsilon \psi(x_i)$ such that

$$\langle \xi_{j+1}^\psi, h_j^f \rangle \leq \bar{\alpha} \|h_j^f\|^2, \tag{4.4b}$$

(ii) compute $h_{j+1}^f = \mathrm{Nr}(\mathrm{co}\{\xi_j^f, \xi_i^\psi, \xi_{i+1}^\psi\})$, set $j = j + 1$ and go to step 9.

Step 12: If

$$f(x_i - s_j h_j^f) - f(x_i) \leq -s_j \alpha \|h_j^f\|^2, \tag{4.5a}$$

set $h_i = -h_j^f$ and go to step 13.

Else, (i) use the bisection method to compute a $\xi_{j+1}^f \in \partial_\varepsilon f(x_i)$ such that

$$\langle \xi_{j+1}, h_j^f \rangle \leq \bar{\alpha} \|h_j^f\|^2, \tag{4.5b}$$

(ii) compute $h_{j+1}^f = \mathrm{Nr}(\mathrm{co}\{\xi_j^f, \xi_{j+1}^f, \xi_j^\psi\})$, set $j = j + 1$ and go to step 9.

Step 13: Compute

$$\lambda_i = \arg\max\{\beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \leq \alpha\beta^k \|h_i\|^2;\ \psi(x_i + \beta^k h_i) \leq 0,\ k \in \mathbb{N}^+\} \tag{4.6}$$

set $x_{i+1} = x_i + \lambda_i h_i$, set $i = i + 1$ and go to step 1.

CASE 3: $\psi(x_i) > 0$.

Step 14: Set $j = 0$. Compute $\xi_0^f \in \partial_\varepsilon f(x_i)$, $\xi_0^\psi \in \partial_\varepsilon \psi(x_i)$, $\Gamma(x_i) = e^{-\psi(x_i)}$,

$h_0^f = \mathrm{Nr}\ \mathrm{co}\{\xi_0^f, \xi_0^\psi\}$. Set $h_0^\psi = \xi_0^\psi$, $h_0^\Gamma = \Gamma(x_i)h_0^f + (1-\Gamma(x_i))h_0^\psi$.

Step 15: If $\max\{\|\Gamma(x_i)h_j^f\|^2, \|(1-\Gamma(x_i))h_j^f\|^2\} < \delta\varepsilon$ set $\varepsilon = \nu\varepsilon$ and go to step 14.

Step 16: Set $s_j = \arg\max\{\beta^k | \beta^k \leq \{\varepsilon/\|h_j^\Gamma\|, k \in \mathbb{N}^+\}$.

Step 17: If

$$\psi(x_i - s_j h_j^\Gamma) - \psi(x_i) \leq -s_j \alpha \|h_j^\Gamma\|^2, \tag{4.7a}$$

set $h_i^\Gamma = -h_j^\Gamma$ and go to step 20.

Else, use the bisection method to compute

a $\xi_{j+1}^\psi \in \partial_\varepsilon \psi(x_i)$ such that

$$\langle \xi_{j+1}^\psi, h_j^\Gamma \rangle \leq \bar{\alpha} \|h_j^\Gamma\|^2 \tag{4.7b}$$

and proceed.

Step 18: If $j \leq [\Gamma(x_i)^{-1}]$ (the integer part of) and $f(x_i - s_j h_j^\Gamma) - f(x_i) > s_j \alpha \|h_j^\Gamma\|^2$, use the bisection method to compute a $\xi_{j+1}^f \in \partial_\varepsilon f(x_i)$ such that

$$\langle \xi_{j+1}^f, h_j^f \rangle \leq \bar{\alpha} \|h_g^f\|^2 . \tag{4.8}$$

Else set $\xi_{j+1}^f = \xi_j^f$.

Step 19: Compute

$h_{j+1}^\psi = \mathrm{Nr}(\mathrm{co}\{h_j^\psi, \xi_{j+1}^\psi\})$, $h_{j+1}^f = \mathrm{Nr}(\mathrm{co}\{\xi_{j+1}^f, \xi_j^f, h_j^\psi, \varepsilon_{j+1}^\psi\})$

$h_{j+1}^\Gamma = \Gamma(x_i)h_{j+1}^f + (1-\Gamma(x_i))h_{j+1}^\psi$.

Set $j = j + 1$ and go to step 15.

Step 20: Compute

$$\lambda_i = \arg\max\{\beta^k | \psi(x_i + \beta^k h_i) - \psi(x_i) \leq -\beta^k \alpha \|h_i\|^2;\ k \in \mathbb{N}^+\} \tag{4.9}$$

set $x_{i+1} = x_i + \lambda_i h_i$.

set $i = i + 1$ and go to step 1. ¤

Theorem 4.1: Suppose that Algorithm 4.1 constructs a sequence $\{x_i\}$. If $\{x_i\}$ is finite, with last element x_N (i.e. the algorithm jams at x_N) then $\psi(x_N) \leq 0$ and $0 \in co\{\partial f(x_N) \cup \partial_0^+\psi(x_N)\}$. If $\{x_i\}$ is infinite, then any accumulation point $\hat{x}$ of $\{x_i\}$ satisfies $\psi(\hat{x}) \leq 0$, $0 \in co\{\partial f(\hat{x}) \cup \partial_0^+\psi(\hat{x})\}$.

Proof: a) Suppose that $\{x_i\}$ is finite, terminating at x_N. Suppose that either $\psi(x_N) > 0$ or that $\psi(x_N) \leq 0$ and $0 \notin co\{\partial f(x_N) \cup \partial_0^+\psi(x_N)\}$.

Case 1: Suppose that $\psi(x_N) \leq 0$ and $0 \notin co\{\partial f(x_N) \cup \partial_0^+\psi(x_N)\}$. Then, referring to (3.13f), $\varepsilon'(x_N) > 0$ and we can consider two subcases:

Subcase 1a: The algorithm is cycling between steps 3 and 6. In this case, because of Proposition 2.3, we must have $\|h_j^f\| \to 0$ as $j \to \infty$ and hence $\varepsilon \searrow 0$ as $j \to \infty$. Consequently, there exists a j_0 such that $\varepsilon \leq \varepsilon'(x_N)$ for all $j \geq j_0$ and hence (see 3.13b) we must have that $\|h_j^f\|^2 \geq \|h^f_{\varepsilon(x_N)}(x_N)\|^2 > \delta\varepsilon(x_N)$ for all $j \geq j_0$, which is clearly a contradiction.

Subcase 1b: The algorithm is cycling between steps 2, 7 and 12. Since by Proposition 2.3, $h_j^f \to 0$ as $j \to \infty$, $\varepsilon \searrow 0$. If $\psi(x_N) < 0$, then there exists a j such that $\psi(x_N) < -\varepsilon$ and hence the algorithm transfers permanently into the loop defined by steps 3 to 6. But we have already shown that the algorithm cannot jam up in this loop. Hence, suppose that $\psi(x_N) = 0$. In this case, there exists a j_0 such that $\varepsilon \geq \varepsilon'(x_N)$ for all $j \geq j_0$ and hence, $\|h_j^f\|^2 \geq \|h^f_{\varepsilon'(x_N)}(x_N)\|^2 \geq \delta\varepsilon'(x_N) > 0$ and, again, we have a contradiction.

Case 2: Suppose that $\psi(x_N) > 0$. Then, by Assumption 3.2, $0 \notin \partial\psi(x_N)$ and $\varepsilon'(x_N) > 0$. Now, if $j \to \infty$, then, by Proposition 2.3, we must have $h_j^{\psi} \to 0$ as $j \to \infty$ and hence, by construction in step 19, $h_j^{f} \to 0$ as $j \to \infty$. Consequently, $\varepsilon \searrow 0$ as $j \to \infty$, so that there exists j_0 such $\varepsilon \leq \varepsilon^1(x_N)$ for all $j \geq j_0$. But then, for all $j \geq j_0$ we must have that $\|h_j^{\psi}\|^2 \geq \|h^{\psi}_{\varepsilon(x_N)}(x_N)\|^2$ and $\|h_j^{f}\|^2 \geq \|h^{f}_{\varepsilon(x_N)}(x_N)\|^2$. Consequently, $\max\{\|\Gamma(x_N)h_j^{f}\|^2, \|(1-\Gamma(x_N))h_j^{\psi}\|^2\} \geq -\theta^1_{\varepsilon(x_N)}(x_N) \geq \delta\varepsilon(x_N) \geq \delta\varepsilon$, which contradicts the conclusion that $\varepsilon \searrow 0$.

We have thus shown that the algorithm cannot jam up at a point x_N such that $\psi(x_N) > 0$ or $\psi(x_N) \leq 0$ and $0 \notin \text{co}\{\partial f(x_N) \cup \partial_0^+\psi(x_N)\}$.

b) Suppose that the sequence $\{x_i\}$ is infinite and that $x_i \overset{K}{\to} \hat{x}$, with $K \subset \{0,1,2,\ldots\}$ and either $\psi(\hat{x}) > 0$ or $\psi(\hat{x}) \leq 0$ and $0 \notin \text{co}\{\partial f(\hat{x}) \cup \partial_0^+\psi(x)\}$.

Case 1: $\psi(x_i) > 0$ for i. In this case $\psi(\hat{x}) \geq 0$ and $\varepsilon^1(\hat{x}) > 0$. By Lemma 3.2, there exists an i_0 such that $\varepsilon^1(x_i) \geq \nu\varepsilon'(\hat{x}) > 0$ for all $i \geq i_0$, $i \in K$. Consequently, since the test value of ε in the implementable algorithm is always greater than or equal to that in the conceptual algorithm, we must have, by Lemma 3.1, that $\|h_i\|^2 \geq \delta\nu\varepsilon^1(\hat{x}) > 0$ for all $i \geq i_0$, $i \in K$. Also, there exists a $b < \infty$ such that $\|h_i\| \leq b$ for all $i \in K$. Consequently, in (4.7a), for all $i \in K$, $i \geq i_0$ and $j = 0,1,2,\ldots$, we must have $s_j \geq \beta\varepsilon^1(x_i) \geq \beta\nu^2\varepsilon^1(\hat{x})/b$. Hence, by (4.9)

$$\psi(x_{i+1}) - \psi(x_i) \leq -[\beta\nu\varepsilon^1(\hat{x})/b]\,\delta\alpha\nu\varepsilon^1(\hat{x}) = -\delta\alpha\beta\nu^2\varepsilon^1(\hat{x})^2/b < 0 \tag{4.10}$$

for all $i \in K$, $i \geq i_0$. However, by continuity, $\psi(x_i) \overset{K}{\to} \psi(\hat{x})$ and hence, since $\psi(x_i) \searrow$, we must have that $\psi(x_i) \to \psi(\hat{x})$. But this is contradicted by (4.10) and hence the theorem is proved for the case where $\psi(x_i) > 0$ for all i.

Case 2: There exists an i_0 such that $\psi(x_{i_0}) \leq 0$. Then, by construction, we must have $\psi(x_i) \leq 0$ for all $i \geq i_0$. Now suppose that $x_i \overset{K}{\to} \hat{x}$, $K \subset \{0,1.2,\ldots\}$, with $\psi(\hat{x}) \leq 0$ and $0 \notin co\{\partial f(\hat{x}) \cup \partial_0^+\psi(\hat{x})\}$. Then $\varepsilon^1(\hat{x}) > 0$ and there exists an $i_1 \geq i_0$ such that $\varepsilon'(x_i) \geq \nu\varepsilon^1(\hat{x}) > 0$ for all $i \in K$, $i \geq i_1$. Consequently, with $b = \sup\{\|h_i\| \,|\, i \in K\} < \infty$, we have once more that $s_j \geq \nu\varepsilon'(x_i)/b \geq \beta\nu\varepsilon'(\hat{x})/b$ for all $i \in K$, $i \geq i_0$, and $\|h_i\|^2 \geq \|h^f_{\varepsilon(x_i)}(x_i)\|^2 \geq \delta\varepsilon'(x_i) \geq \delta\nu\varepsilon(\hat{x})$ for all $i \in K$, $i \geq i_0$. It now follows from (4.3a) and (4.5a) that

$$f(x_{i+1}) - f(x_i) \leq -\alpha\beta\nu^2\varepsilon^1(\hat{x})^2/b \tag{4.11}$$

for all $i \quad K$, $i \geq i$. Now, $f(x_i) \overset{K}{\to} f(\hat{x})$ by continuity and $f(x_i)$, hence $f(x_i) \nrightarrow f(\hat{x})$. But this is contradicted by (4.11) and hence we are done.

References

A1. L. Armijo, "Minimization of functions having Lipschitz continuous first partial derivatives," Pacific Journal of Mathematics 16 (1966) 1-3.

B1. C. Berge, Topological Spaces, Macmilan Co., N.Y., 1963.

B2. D.P. Bertsekas and S.K. Mitter, "A descent numerical method for optimization problems with nondifferentiable cost functionals," Journal of Control, 11, no. 4, (1973) pp. 637-652.

C1. F. Clarke, "Generalized gradients and applications," Trans. Amer. Math. Soc., Vol. 205, pp. 247-262, 1975.

C2. F. Clarke, "A new approach to Lagrange Multipliers," Math. of Oper. Res., 1 (1976) pp. 165-174.

C3. F. Clarke, "Optimal Control and the true Hamiltonian," SIAM Review, Vol. 21, No. 2.

C4. F. Clarke, Private communication, April 1979.

C5. A. R. Conn, "Constraint optimization using a nondifferentiable penalty function," SIAM J. Numer. Anal., Vol. 10, No. 4, pp. 760-784, 1973.

C6. J. Cullum, W. E. Donath and P. Wolfe, "The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices," Mathematical Programming Study, No. 3, November 1975, pp. 35-55.

D1. J. M. Danskin, "The theory of maxmin with applications," SIAM Journal of Applied Mathematics, 14, No. 4 (1966) pp. 641-655.

D2. V. F. Demjanov , "Algorithms for some minmax problems," JCSS 2 (1968).

D3. V. F. Demjanov and V. N. Malozemov," Introduction to minimax, John Wiley & Sons, New York (1974).

D4. V.F. Demjanov, "Differentiability of a maxmin function," I. U.S.S.R. Comp. Math and Math. Physics, Vol. 8, No. 6, pp. 1-15, 1968.

F1. A. V. Fiacco and G. P. McCormick, Nonlinear programming: sequential unconstrained minimization techniques John Wiley & Sons, New York, 1968.

F2. R. Fletcher, "An exact penalty function method for nonlinear programming with inequalities," Math. Prog. 5, No. 2 (1973) pp. 129-150.

G1. A. A. Goldstein, "Optimization of Lipschitz continuous functions," Math. Prog. 13, No. 1 (1977) pp. 14-22.

G2. C. Gonzaga and E. Polak, "On constraints dropping schemes and optimality functions for a class of outer approximations algorithms," SIAM J. Contr. and Optimization, Vol. 17, pp. 477-493, 1979.

L1. G. Lebourg, "Valeur Moyenne pour gradient généralisé, "C.R. Acad. Sci., Paris, Vol. 281, pp. 795-797, 1975.

L2. C. Lemarechal, "Extensions Diverses des Méthodes de Gradient et Applications," Thesis, University of Paris VIII, 1980.

L3. C. Lemarechal, "Non smooth optimization: toward a synthesis," Actes Journées de l'optimisation, Montréal (1978) pp. 69-70.

L4. C. Lemarechal, "Minimization of nondifferentiable functions with constraints," Actes 12th Allerton Conference on Circuit Theory, Univ. of Illinois (1974) pp. 16-24.

L5. C. Lemarechal, "Nondifferentiable optimization,subgradient and ε subgradient methods," Lecture notes, No. 117, Optimization and Operations Research, Springer Verlag (1976).

L6. C. Lemarechal, "An extension of Davidon methods to nondifferential problems," Mathematical Programming Studies 3, M. L. Balinski and R. Wolfe Eds., North Holland, Amsterdam, pp. 95-109, 1975.

M1. R. Mifflin, "Semismooth and semiconvex functions in constrained optimization," SIAM Journal of Control, 15, No. 6 (1977) pp. 959-972.

M2. D.Q. Mayne, E. Polak and R. Trahan, "On outer approximation algorithms for computer aided design problems," JOTA, 28, July 1979, pp. 331-352.

M3. D.Q. Mayne and E. Polak, "Feasible directions algorithms for optimization problems with equality and inequality constraints," Math. Prog. 11 (1976) pp. 67-80.

P1. O. Pironneau and E. Polak, "On the rate of convergence of certain methods of centers," Math. Prog. 2, No. 2 (1972) pp. 230-257.

P2. E. Polak and D. Q. Mayne, "on the solution of singular value inequalities over a continuum of frequencies," Memorandum No. UCB/ERL M80/8, January 1980; 20th I.E.E.E. Conf. on Dec. and Control, Albuquerque, N.M., Dec. 12-14, 1980.

P3. E. Polak, R. Trahan and D. Q. Mayne, "Combined phase I - phase II methods of feasible directions," Mathematical Programming, Vol. 17, pp. 61-73, July 1979.

P4. E. Polak and D. Q. Mayne, "An algorithm for optimization problems with functional inequality constraints," I.E.E. Trans. on Auto Control, Vol. AC-21, No. 2, April 1976.

P5. E. Polak, "on a class of computer aided design problems," A link between Science and Applications of Automatic Control, edited by A. Niemi, Pergammon Press, Oxford and New York, 1979.

P6. E. Polak and R. Trahan, "An algorithm for computer aided design of control problems," Proc. I.E.E.E. Conf. on Decision and Control, 1976.

P7. E. Polak and A. Sangiovanni-Vincentelli, Theoretical and computational aspects of optimal design conetering, tolerances and tuning problem," I.E.E.E. Trans. on Auto. Control, Vol. CAS-26, No. 9, pp. 295-318, 1979

P8. E. Polak, Computational Methods in Optimization: A unified approach, Academic Press, New York, 1971.

P9. E. Polak, R. W. H. Sargent and D. J. Sebastian, "On the convergence of Sequential Minimization Algorithms," JOTA, Vol. 14, No. 4, 1974, pp. 439-442.

P10. E. Polak, "On the Global Stabilization of Locally Convergent Algorithms," Automatica, Vol. 12, pp. 337-342, 1976.

P11. B. T. Poljak, "A general method for solving extremum problems," Soviet Math., No. 8, (1966), pp. 593-597.

P12. B. T. Poljak, "Minimization of nonsmooth functionals," U.S.S.R. Computational Mathematics and Mathematical Physics, No. 9 (1969), pp. 14-29.

P13. B. N. Pshenichnyi, Necessary Conditions for the Extremum, Translation edited by L. W. Neustadt, translated by K. Makovski, Marcel Dekker Inc., N.Y. 1971.

P14. T. Pietrzykowski, "The Potential Method for Conditional Maxima in the Locally Compact Metric Spares," Numer. Math., 14 (1970), pp. 325-329.

R1. R. T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, 1970.

R2. W. Rudin, Principles of Mathematical Analysis, McGraw-Hill, N.Y., 1964.

S1. N. Z. Shor, "Utilization of the operation of space dilitation in the minimization of convex functions," Cybernetics 1 (1970) pp. 7-15.

S2. N. Z. Shor and L. P. Shabashova, "Solution of minmax problems by the method of generalized gradient descent with dilitation of space," Cybernetics 1 (1972), pp. 88-94.

T1. S. Tishyadhigama, E. Polak and R. Klessig, "A comparative study of several general convergence conditions for algorithms modeled by point-to-set maps," Math. Prog. Study 10, pp. 172-190, North Holland, 1979.

W1. P. Wolfe, "Finding the nearest point in a polytope," Math. Prog. 11, No. 2 (1976) pp. 128-149.

W2. P. Wolfe, "A method of conjugate subgradients for minimizing nondifferentiable functions," in Nondifferentiable Optimization, M. L. Balinski and P. Wolfe, Eds., Math. Prog. Study 3, pp. 145-173, North Holland, Amsterdam, 1975.

Z1. W. I. Zangwill, "Nonlinear programming via penalty functions," Man. Sc. 13, No. 5 (1967) pp. 344-358.

Z2. G. Zoutendijk, Methods of Feasible Directions, Elsevier, Amsterdam.