International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

An Algorithm for Constrained Optimization with Semismooth Functions

Mifflin, R.

IIASA Research Report
February 1977

RR-77-3

AN ALGORITHM FOR CONSTRAINED OPTIMIZATION WITH SEMISMOOTH FUNCTIONS

R. MIFFLIN

FEBRUARY 1977

International Institute for Applied Systems Analysis
2361 Laxenburg, Austria

PREFACE

Large-scale optimization models arise in many areas of application at IIASA. For example, such models are useful for estimating the economic value of introducing solar and wind generated electrical energy into an existing power grid and for determining equilibrium prices for agricultural commodities in international trade as a function of national policies. Certain methods of decomposition for solving such optimization problems require the solution of a relatively small problem whose objective function is not everywhere differentiable. This paper gives an implementable algorithm that can be used to solve such nonsmooth optimization problems.

ABSTRACT

We present an implementable algorithm for solving constrained optimization problems defined by functions that are not everywhere differentiable. The method is based on combining, modifying and extending the nonsmooth optimization work of Wolfe, Lemarechal, Feuer, Poljak and Merrill. It can be thought of as a generalized reset conjugate gradient algorithm.

We also introduce the class of weakly upper semismooth functions. These functions are locally Lipschitz and have a semicontinuous relationship between their generalized gradient sets and their directional derivatives. The algorithm is shown to converge to stationary points of the optimization problem if the objective and constraint functions are weakly upper semismooth. Such points are optimal points if the problem functions are also semiconvex and a constraint qualification is satisfied. Under stronger convexity assumptions, bounds on the deviation from optimality of the algorithm iterates are given.

An Algorithm for Constrained Optimization with Semismooth Functions

1. INTRODUCTION

In this paper we present an implementable algorithm for solving very general constrained optimization problems of the following type:

$$\text{minimize} \quad f(x)$$
$$\text{subject to } h(x) \leqq 0$$

where $x \in R^n$ and f and h are real-valued functions that are "locally Lipschitz", i.e. Lipschitz continuous on each bounded subset of R^n. These problems are "nonsmooth" in the sense that the problem functions f and h need not be differentiable everywhere. However, locally Lipschitz functions do have "generalized gradients" (Clarke [2,3]) and a necessary optimality condition [3] is that the zero vector is a certain convex combination of generalized gradients of f and h. This "stationarity" condition is sufficient for optimality if f and h are "semiconvex" [27] and a constraint qualification is satisfied.

Our algorithm combines, extends and modifies ideas contained in Wolfe [39], Feuer [10,11], Poljak [31] and Merrill [36] and, by means of a map defined in [36], deals with "corners" arising from constraints in the same manner as it handles discontinuities of the problem function gradients. It has accumulation points that satisfy the above stationarity condition if f and h are "weakly upper semismooth" as defined in section 2. Such functions have a semicontinuous relationship between their generalized gradients and directional derivatives where this relationship is properly weaker than the corresponding one for "semismooth" functions introduced in [27].

The difficulties in minimizing a nonsmooth function are well discussed in [39] and [10], where implementable descent algorithms are given. Wolfe's method [39] is for a convex function and Feuer [10,11] has extended it for finding a stationary point of a function that is the pointwise maximum or minimum of a family of continuously differentiable functions. In [27] we show that such functions are properly contained in the class of semismooth functions. The algorithm in [39] is closely related to that of Lemarechal [21] and for a quadratic function these both coincide with the method of conjugate gradients [17] and, hence, have finite termination in this case, as does an algorithm of Shor [35,36,37].

The descent approach for convex functions of Bertsekas and Mitter [1] has been made implementable by Lemarechal [19] and has been extended in theory to locally Lipschitz functions by Goldstein [14]. Descent algorithms for min-max objectives, which are also difficult to implement, are given in Demjanov [5] and Goldstein [13].

Lemarechal [20] has also suggested a method for constrained convex programming problems which deals with nonlinear constraint functions by means of an exact penalty function approach [4,9, 28,40].

Shor's [34] nondescent "subgradient algorithm" for unconstrained convex problems was extended to constrained problems by Poljak [31], who developed a method that uses subgradients of the objective function at feasible points and subgradients of the constraint functions at infeasible points. This idea is related to a concept employed by Merrill [36] for solving constrained problems by means of a fixed point algorithm. Similar ideas were also developed by Hansen [15], Hansen and Scarf [16] and Eaves [6] for solving convex programming problems by fixed point-type algorithms [7,33]. These methods are combinatorial in nature and able to solve equilibrium problems that are more general than convex programming problems. Our algorithm differs from these, because it is a feasible point method which depends significantly on the constrained optimization nature of the problem. The method may use information from infeasible points, but the objective function f need not be evaluated at such points. Our algorithm employs a line search procedure along directions that may be infeasible, and, hence, the method is not a feasible direction [41] algorithm. However, it is related to the similar feasible direction methods of Mangasarian [24] (see also [12]) and Pironneau and Polak [29] for continuously differentiable functions. As with ours, these methods have search direction finding subproblems that are quadratic programming problems involving convex combinations of problem function gradients. Our method differs, because there is no linear term in the subproblem objective related to complementary slackness and not all of the subproblem data need be changed from iteration to iteration. Because we do not assume differentiability, our subproblems may include more than one generalized gradient from the same problem function. This can be a good idea even in the case of differentiable functions, because it can bring curvature information about the functions into search direction determination and, thus, have the potential for better than linear convergence. There are tests in our algorithm which attempt to smooth or balance the process of retaining or dropping accumulated gradient information, and hopefully allow the method to behave like a reset conjugate gradient [22,25] algorithm when applied to smooth unconstrained problems. This process is flexible and gives the algorithm the potential for a good rate of convergence.

The algorithm is defined in section 3 where we also discuss how it compares to and differs from the methods in [10], [21] and [39] when applied to unconstrained problems.

In section 4, under the assumption that f and h are weakly upper semismooth, we show that either our line search procedure is finite or f is unbounded from below on the set of feasible points.

In section 5 we show stationarity of the algorithm's accumulation points. Under convexity assumptions, we give bounds on the deviation from optimality of the iterates for a version of the algorithm which uses a gradient deletion rule that is especially designed for convex problems.

Throughout this paper we mostly adhere to the notation in [32] and [39]. For example, conv(S) denotes the convex hull of a set $S \subset R^n$, i.e. $x \in \mathrm{conv}(S)$ if and only if $x = \sum_{i=1}^{p} \lambda_i x^i$ where p is a positive integer, $\lambda_i \geqq 0$ and $x^i \in S$ for $i = 1,2,\ldots,p$ and $\sum_{i=1}^{p} \lambda_i = 1$. The scalar product of $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ in R^n, defined by $\sum_{i=1}^{n} x_i y_i$, is denoted $\langle x,y \rangle$ and the Euclidean norm of x, defined by $|\langle x,x \rangle^{1/2}|$, is denoted $|x|$.

2. DEFINITIONS AND PRELIMINARY RESULTS

2a. Locally Lipschitz and Semismooth Functions

Let B be an open subset of R^n and $F : R^n \to R$ be *Lipschitz* on B, i.e. there exists a positive number K such that

$$|F(y) - F(z)| \leqq K|y-z| \qquad \text{for all } y,z \in B \quad .$$

If F is Lipschitz on each bounded subset of R^n then F is called *locally Lipschitz*.

Let $x \in B$ and $d \in R^n$. As in Clarke [3], let

$$F^0(x;d) = \limsup_{\substack{h \to 0 \\ t \downarrow 0}} [F(x+h+td) - F(x+h)]/t$$

and let $\partial F(x)$ denote the *generalized gradient* of F at x defined by

$$\partial F(x) = \{g \in R^n : \langle g,d \rangle \leqq F^0(x;d) \text{ for all } d \in R^n\} \quad .$$

The following proposition collects together useful properties of F^0 and ∂F.

Proposition 1.

(a) $\partial F(x)$ is a nonempty convex compact subset of R^n [3] .

(b) $F^0(x;d) = \max [\langle g,d \rangle : g \in \partial F(x)]$ [3] .

(c) If $\{x_k\} \subset B$ converges to x and $g_k \in \partial F(x_k)$ for each k then $|g_k| \leqq K$ and each accumulation point g of $\{g_k\}$ satisfies $g \in \partial F(x)$, i.e. ∂F is bounded on bounded subsets of B and is *uppersemicontinuous* on B [3] .

(d) Let y and z be in a convex subset of B. Then there exists $\lambda \in (0,1)$ and $g \in \partial F(y+\lambda(z-y))$ such that

$$F(z) - F(y) = \langle g, z-y\rangle \quad ,$$

i.e. a mean value result holds [18] .

(e) Let $\{t_k\} \downarrow 0$, $\{h_k\} \to 0 \in R^n$ and F^* be any accumulation of

$$\{[F(x+h_k+t_k d) - F(x+h_k)]/t_k\} \quad .$$

Then there exists $g \in \partial F(x)$ such that

$$F^* = \langle g,d\rangle \text{ [27]} \quad .$$

If $\lim_{t \downarrow 0} [F(x+td) - F(x)]/t$ exists it is denoted by $F'(x;d)$ and called the *directional derivative* of F at x in the direction d. Note that if $F'(x;d)$ exists then, by (e) above, there exists $g \in \partial F(x)$ such that

$$F'(x;d) = \langle g,d\rangle \quad .$$

Definition 1 and Proposition 2 to follow are given in [27] along with other properties and examples of semismooth functions.

Definition 1. $F : R^n \to R$ is *semismooth* at $x \in R^n$ if

(a) F is Lipschitz on a ball about x

and

(b) for each $d \in R^n$ and for any sequences $\{t_k\} \subset R_+$, $\{\theta_k\} \subset R^n$ and $\{g_k\} \subset R^n$ such that

$$\{t_k\} \downarrow 0,\ \{\theta_k/t_k\} \to 0 \in R^n \text{ and } g_k \in \partial F(x+t_k d+\theta_k) \quad ,$$

the sequence $\{<g_k,d>\}$ has exactly one accumulation point.

Proposition 2.

If F is semismooth at x then for each $d \in R^n$, $F'(x;d)$ exists and equals $\lim_{k\to\infty} <g_k,d>$ where $\{g_k\}$ is any sequence as in Definition 1.

Definition 2. $F : R^n \to R$ is *weakly upper semismooth* at $x \in R^n$ if

(a) F is Lipschitz on a ball about x

and

(b) for each $d \in R^n$ and for any sequences $\{t_k\} \subset R_+$ and $\{g_k\} \subset R^n$ such that $\{t_k\} \downarrow 0$ and $g_k \in \partial F(x+t_k d)$ it follows that

$$\liminf_{k\to\infty} <g_k,d> \geq \limsup_{t\downarrow 0} [F(x+td) - F(x)]/t \quad .$$

Proposition 3.

If F is *weakly upper semismooth* at x then for each $d \in R^n$, $F'(x;d)$ exists and there exist sequences $\{\tau_k\} \subset R_+$ and $\{g_k\} \subset R^n$ such that $\{\tau_k\} \downarrow 0$, $g_k \in \partial F(x+\tau_k d)$ and

$$\lim_{k\to\infty} <g_k,d> = F'(x;d) \quad .$$

Proof: Suppose $\{\tau_k\} \downarrow 0$ is a sequence such that

$$\lim_{k\to\infty} [F(x+\tau_k d) - F(x)]/\tau_k = \liminf_{t\downarrow 0} [F(x+td) - F(x)]/t \quad .$$

By (d) of Proposition 1, there exists $t_k \in (0,\tau_k)$ and $g_k \in \partial F(x+t_k d)$ such that

$$F(x+\tau_k d) - F(x) = \tau_k <g_k,d> \quad .$$

Then, by Definition 2, since $\{t_k\} \downarrow 0$, we have

$$\lim_{k\to\infty} [F(x+\tau_k d) - F(x)]/\tau_k = \lim_{k\to\infty} \langle g_k, d\rangle \geqq \limsup_{t\downarrow 0} [F(x+td) - F(x)]/t \quad .$$

So,

$$\liminf_{t\downarrow 0} [F(x+td) - F(x)]/t = \lim_{k\to\infty} \langle g_k, d\rangle \geqq \limsup_{t\downarrow 0} [F(x+td) - F(x)]/t$$

and the desired results follow immediately.□

It is clear from the above definitions and propositions that the following holds:

Proposition 4.

If F is semismooth at x then F and $-F$ are weakly upper semismooth at x.

We say that F is weakly upper semismooth (semismooth) on $X \subset R^n$ if F is weakly upper semismooth (semismooth) at each $x \in X$.

An example of a locally Lipschitz function $F(x)$ for $x \in R$ that is weakly upper semismooth on R but not semismooth at $x = 0$ is the following:

$$F(x) = x^2 \qquad \text{for } x \leqq 0 \text{ or } x \geqq 1 \quad ,$$

and for each integer $n = 1, 2, \ldots$

$$F(x) = \begin{cases} (1 + \frac{1}{n})(x - \frac{1}{n+1}) & \text{for } \frac{1}{n}[1 - (\frac{1}{n+1})^2] \leqq x \leqq \frac{1}{n} \\ \frac{1}{n}(x - (\frac{1}{n+1})^2) & \text{for } \frac{1}{n+1} \leqq x \leqq \frac{1}{n}[1 - (\frac{1}{n+1})^2] \end{cases} \quad .$$

It can be verified that $F'(0;1) = 0$ and $\partial F(0) = \text{conv } \{0, 1\}$ is the set of possible accumulation points of $\{g_k\}$ where $g_k \in \partial F(x_k)$ and $\{x_k\} \downarrow 0$. Note also that the locally Lipschitz function $-F(x)$ is not weakly upper semismooth at $x = 0$.

From [27, Proposition 3] and Proposition 4 we have the following:

Proposition 5.

If $F : R^n \to R$ is convex, then F is locally Lipschitz,

$$\partial F(x) = \{g \in R^n : F(y) \geq F(x) + \langle g, y-x\rangle \text{ for all } y \in R^n\} \text{ for each } x \in R^n \quad ,$$

F is semismooth on R^n and, hence, F is weakly upper semismooth on R^n.

Remark: ∂F in Proposition 5 is called the *subdifferenital* [32] of the convex function F. We refer to the inequality in the expression for ∂F as the *subgradient inequality*.

2b. Stationarity

Corresponding to the locally Lipschitz optimization problem functions f and h, define $M : R^n \to 2^{R^n}$ by

$$M(x) = \begin{Bmatrix} \partial f(x) & \text{if } h(x) < 0 \\ \text{conv } \{\partial f(x) \cup \partial h(x)\} & \text{if } h(x) = 0 \\ \partial h(x) & \text{if } h(x) > 0 \end{Bmatrix} \quad \text{for } x \in R^n \quad .$$

This map was introduced and used by Merrill [36, Chapter 12] for problems with differentiable and/or convex functions.

We say that $x \in R^n$ is *feasible* if $h(x) \leq 0$ and that $\bar{x} \in R^n$ is *optimal* if $\bar{x}$ is feasible and $f(\bar{x}) \leq f(x)$ for all feasible x. We call $\bar{x} \in R^n$ *stationary* if $\bar{x}$ is feasible and $0 \in M(\bar{x})$. The following necessary optimality result is proved directly in [27] and follows from a more general result in [3]:

Proposition 6.

If $\bar{x}$ is optimal then $\bar{x}$ is stationary.

From parts (a) and (c) of Proposition 1, the definition of M and Caratheodory's theorem [32, Theorem 17.1] one can derive the following result useful for establishing convergence of our algorithm:

Proposition 7.

M is bounded on bounded subsets of R^n, M is uppersemicontinuous on R^n, and for each $x \in R^n$ $M(x)$ is convex.

3. THE ALGORITHM

For $x \in R^n$, $d \in R^n$ and parameters m_1 and m_2 satisfying $0 < m_2 < m_1 < 1$ we define

$$LT = \{t \geqq 0 : f(x+td) - f(x) \leqq -m_2 t|d|^2,\ h(x+td) \leqq 0\}$$

and

$$RT = \{t \geqq 0 : \langle g(t), d\rangle \geqq -m_1|d|^2\} \quad ,$$

where $g(t)$ for $t \geqq 0$ is an element of $M(x+td)$ returned by a user-supplied subroutine. For ease of exposition, we assume that $g(t) \in \partial f(x+td)$ if $h(x+td) = 0$ and we denote $g(0)$ by g_x.

G is a set of generalized gradients. A typical element of G is denoted g_j and associated with each $g_j \in G$ there is a $y_j \in R^n$ such that $g_j \in M(y_j)$. The algorithm requires the solution of the problem of minimizing $|z|^2 = \sum_{i=1}^{n} z_i^2$ subject to $z = \sum_{g_j \in G} \lambda_j g_j$, $\sum_j \lambda_j = 1$, $\lambda_j \geqq 0$ for all j. The minimizing z is denoted by Nr(G), i.e. Nr(G) is the n-vector in conv (G) nearest to the origin with respect to Euclidean distance. Since this problem is a quadratic programming problem having a very special structure, especially efficient finite algorithms such as in [38] can be designed for its solution.

The algorithm requires a starting feasible point, i.e. an $x_0 \in R^n$ such that $h(x_0) \leqq 0$. If such a point is not immediately available, we may apply the algorithm to the unconstrained problem of minimizing h over R^n. Under certain assumptions (see Theorem 5.2, Corollary 5.3 and Theorem 5.5 below) this algorithm will find a feasible point.

In addition to assuming $h(x_0) \leqq 0$, we assume that $g_0 \neq 0$ where $g_0 \in \partial f(x_0)$. Besides m_1 and m_2, the algorithm requires positive parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$ and q satisfying $\alpha_2 < \alpha_1$, $q \geqq 1$ and $\beta_2 \leqq \beta_1 < 1/|g_0|^{q-1}$. Given the above data and definitions the algorithm is as follows:

Step 0 (Initialization). Set $x = x_0$, $G = \{g_0\}$, $d = -g_0$ and $\delta = |g_0|$.

<u>Step 1</u> (Line Search). Set $t_L = 0$, $t_N = +\infty$ and $t_R = +\infty$ and choose $t > 0$.

Loop: If $t \in LT$ set $t_L = t$. Otherwise set $t_N = t$.
If $t \in RT$ set $t_R = t$.
If $t_R - t_L \leqq \alpha_2\delta/|d|$ go to End. Otherwise replace t by $2t$ if $t_N = +\infty$ or by $\frac{1}{2}(t_L + t_N)$ if t_N is finite and go to Loop.

End: Set $y_L = x + t_L d$, $g_L = g(t_L)$, $y_R = x + t_R d$ and $g_R = g(t_R)$.

<u>Step 2</u> (Update x, G, δ and d).

a. Replace x by y_L.
b. Replace G by $G \cup \{g_L, g_R\}$.
c. Delete all possible g_j from G according to deletion rules I or II given below so that if $g_j \in M(y_j)$ is deleted then $|x - y_j| > \alpha_1\delta$.
d. Compute $Nr(G)$.
e. If $|Nr(G)| < \beta_2\delta^q$ replace δ by $\beta_1\delta^q$ and go to Step 2c. Otherwise set $d = -Nr(G)$, replace δ by $\min[\delta, |d|]$ and go to Step 1.

<u>Deletion Rules</u>. Delete $g_j \in M(y_j)$ from G if

$$\text{I.}\qquad |x - y_j| > \alpha_1\delta \tag{3.1}$$

$$\text{IIa.}\qquad h(y_j) > 0$$

and

$$\langle g_j, x - y_j\rangle < -\alpha_1\delta|g_j| \tag{3.2}$$

$$\text{b.}\qquad h(y_j) \leqq 0 \quad ,$$

$$f(x) + \langle g_x, y_j - x\rangle \leqq f(y_j) \tag{3.3}$$

and

$$f(y_j) - f(x) + \langle g_j, x - y_j\rangle < -\alpha_1\delta|g_x - g_j| \tag{3.4}$$

where

$$g_x \in \partial f(x) \cap G \quad .$$

Using the Cauchy-Schwartz inequality it is not difficult to establish the following result that shows that the deletion requirement of Step 2c is satisfied:

Lemma 3.1. If (3.2) holds, or if (3.3) and (3.4) hold, then (3.1) holds.

Remarks: Some inspiration for rule IIa came from Elzinga and Moore's [8] central cutting plane method.

It is clear that (3.3) is satisfied if f is convex on a convex set containing x and y_j. Thus, (3.3) need not be checked if it is known that f is convex. The advantage of rule II over rule I, when applied to convex problems, is that the former requires storage of two scalars, $h(y_j)$ and $\langle g_j, y_j \rangle$ if $h(y_j) > 0$ or $[\langle g_j, y_j \rangle - f(y_j)]$ if $h(y_j) \leq 0$, instead of the n-vector y_j. Rule IIb also has a good feature for the case when f is *polyhedral*, i.e., the maximum of a finite number of affine functions. In this case if x and y_j are on the same polyhedral piece, i.e., $f(x) = f(y_j) + \langle g_j, x-y_j \rangle$, then rule IIb will not drop g_j no matter how far y_j is away from x. Use of this rule causes the polyhedral example due to M.J.D. Powell in [39] to be solved in a finite number of steps, if the line search procedure is modified to find the exact minimum of $f(x+td)$ over $t > 0$, which is possible in the polyhedral case.

These deletion tests which are applied before each Nr(G) calculation cause selective dropping of old generalized gradients. When applied to unconstrained problems, this makes our method significantly different from the methods in [10,21,39], because these latter algorithms accumulate gradient information until certain distances are too large and then drop all but the most recently generated gradient. Our method also differs from those in [10,21,39] because of the way it incorporates a convergence variable δ that is automatically generated and forced to zero by tests involving user-supplied parameters.

For the case of quadratic f and no constraint h the finitely terminating conjugate gradient property in [39, Section 6] is retained if our line search is modified to be exact and α_1 happens to be so large that no deletion at Step 2c occurs.

Our line search subroutine is a modification of the bisection-type procedure in [39] which was modelled on the differentiable case. The idea of using two points from the line search rather than one appears to be new and is crucial in dealing with constraints. Our procedure has a stopping criterion depending on the convergence variable δ and different decision rules from those in [39] due to the fact we work on nonconvex and/or constrained problems and $LT \cap RT$ may have an empty interior.

4. LINE SEARCH CONVERGENCE AND ASSOCIATED RESULTS

Throughout the remainder of this paper we assume that f and h are weakly upper semismooth functions on $S \subset R^n$ where S is the set of all points in R^n lying within a Euclidean distance of $\alpha_2|g_0|$ of

$$S_0 = \{z \in R^n : f(z) \leqq f(x_0), h(z) \leqq 0\} \quad .$$

In this section we discuss convergence of the line search procedure in Step 1 of the algorithm and give some implications of this procedure's termination conditions. This discussion depends on our parameter choices satisfying $0 < m_2 < m_1 < 1$.

Theorem 4.1. *Suppose* $x \in S_0$, $|d| \neq 0$ *and* $\delta > 0$. *Then the line search procedure of Step 1 either*

(a) *terminates with* t_L, y_L, y_R *and* g_R *satisfying*

$$h(y_L) \leqq 0 \tag{4.1}$$

$$f(y_L) - f(x) \leqq -m_2 t_L |d|^2 = -m_2 |y_L - x|\,|d| \tag{4.2}$$

$$|y_L - y_R| \leqq \alpha_2 \delta \tag{4.3}$$

and

$$\langle g_R, d\rangle \geqq -m_1 |d|^2 \quad , \tag{4.4}$$

or

(b) *generates a sequence* $\{t_k\} \to +\infty$ *such that* $\{f(x+t_k d)\} \to -\infty$ *and* $h(x+t_k d) \leqq 0$ *for all* k.

Proof: If every t generated by the search satisfies $t \in LT$ and $t \notin RT$ then t_N and t_R remain $+\infty$, the procedure does not terminate and doubling causes $t \to +\infty$. In this case the definition of LT shows that $h(x+td) \leqq 0$ for all t and $f(x+td) \to -\infty$, since $-m_2|d|^2 < 0$, so (b) holds.

Suppose (b) does not hold. Then some t either satisfies $t \notin LT$ or $t \in RT$. In the former case, t_N becomes finite, doubling ceases and bisection begins, unless the procedure terminates, because $t - t_L = t_R - t_L \leqq a_2\delta/|d|$. If the former case does not hold, i.e. $t \in LT$, then $t \in LT \cap RT$ and the search terminates. If the search does not terminate, then bisection causes $t_N - t_L$ to approach zero, because either t_L or t_N is replaced by $\frac{1}{2}(t_L + t_N)$ in each loop.

Let us suppose bisection has begun, i.e., $f(x+td) \not\to -\infty$, and assume, for contradiction purposes, that the search does not terminate. In this case the interval $[t_L, t_N]$ converges to some $\hat{t} \geqq 0$. Since $t_L \uparrow \hat{t}$ and f and h are continuous on S, the definition of LT shows that $\hat{t} \in LT$, i.e.

$$f(x+\hat{t}d) - f(x) \leqq -m_2\hat{t}|d|^2 \tag{4.5}$$

and

$$h(x+\hat{t}d) \leqq 0 \quad . \tag{4.6}$$

Since $t_N \notin LT$, $\hat{t} \in LT$ and $t_N \downarrow \hat{t}$, t_N must take on an infinite number of distinct values greater than $\hat{t}$. If $t_N \in RT$ infinitely often then $(t_R - t_L) = (t_N - t_L) \to 0$ for these t_N and the search must stop, because $a_2\delta/|d|$ is positive. So, suppose $t_N \in RT$ for only finitely many bisections. Then for infinitely many bisections we have

$$\langle g(t_N), d\rangle < -m_1|d|^2 \quad ,$$

so

$$\liminf_{t_N \downarrow \hat{t}} \langle g(t_N), d\rangle \leqq -m_1|d|^2 \quad . \tag{4.7}$$

There are two cases to consider depending on whether or not $x + t_N d$ is feasible infinitely often.

<u>Case I</u>. Suppose for infinitely many t_N we have

$$h(x+t_N d) > 0 \quad . \tag{4.8}$$

Then $g(t_N) \in \partial h(x+t_N d)$ and combining (4.6) and (4.8) with the fact that $t_N > \hat{t}$ gives

$$\frac{h(x+t_N d)-h(x+\hat{t}d)}{t_N-\hat{t}} > 0 \quad .$$

Thus, since h is weakly upper semismooth and $g(t_N) \in \partial h(x+\hat{t}d+(t_N-\hat{t})d)$,

$$\liminf_{t_N \downarrow \hat{t}} \langle g(t_N),d\rangle \geqq \limsup_{t_N \downarrow \hat{t}} \frac{h(x+t_N d)-h(x+\hat{t}d)}{t_N-\hat{t}} \geqq 0 \quad .$$

But this contradicts (4.7), because $-m_1|d|^2 < 0$.

<u>Case II</u>. Suppose for infinitely many t_N (4.8) does not hold. Then $g(t_N) \in \partial f(x+t_N d)$ and, since $t_N \notin LT$,

$$f(x+t_N d) - f(x) > -m_2 t_N |d|^2 \quad ,$$

which combined with (4.5) gives

$$f(x+t_N d) - f(x+\hat{t}d) > -m_2(t_N-\hat{t})|d|^2 \quad .$$

Thus, since f is weakly upper semismooth and $g(t_N) \in \partial f(x+\hat{t}d+(t_N-\hat{t})d)$,

$$\liminf_{t_N \downarrow \hat{t}} \langle g(t_N),d\rangle \geqq \limsup_{t_N \downarrow \hat{t}} \frac{f(x+t_N d)-f(x+\hat{t}d)}{t_N-\hat{t}} \geqq -m_2|d|^2 \quad .$$

But this also contradicts (4.7), because $m_2 < m_1$ and $|d| \neq 0$. Therefore neither case occurs and the search terminates. From various definitions and rules of the algorithm it is easy to show that (4.1) through (4.4) hold at termination.□

From the assumptions that $h(x_0) \leq 0$, $|g_0| \neq 0$ and $0 < \beta_2 \leq \beta_1 < 1/|g_0|^{q-1}$, Theorem 4.1 and the rules of the algorithm it is easy to establish inductively that the following holds:

Lemma 4.2. All values assigned to x, d, δ, y_L and y_R by the algorithm satisfy $x \in S_0$, $|d| \neq 0$, $0 < \delta \leq |g_0|$, $y_L \in S_0$ and $y_R \in S$.

The next result shows that in the case of a convex problem we do not need the variable t_N in the line search procedure, because it may be replaced by t_R wherever it appears, since if $t \notin LT$ then $t \in RT$.

Theorem 4.3. If f *and* h *are convex functions on* R^n *then every value of* t *generated by the line search procedure satisfies* $t \in LT \cup RT$.

Proof: If $t \in LT$ we are done. So, suppose $t \notin LT$. Then either

$$h(x+td) > 0 \tag{4.9}$$

or

$$f(x+td) - f(x) > -m_2 t|d|^2 \quad . \tag{4.10}$$

If (4.9) holds then $g(t) \in \partial f(x+td)$ and, by the convexity of h, the subgradient inequality and the feasibility of x, we have

$$h(x+td) - t\langle g(t),d\rangle \leq h(x) \leq 0 \quad . \tag{4.11}$$

Combining (4.9) and (4.11) yields

$$\langle g(t),d\rangle \geq 0 \quad . \tag{4.12}$$

If (4.9) does not hold then (4.10) holds and $g(t) \in \partial f(x+td)$. By the convexity of f and the subgradient inequality we have

$$f(x+td) - t\langle g(t),d\rangle \leq f(x) \quad . \tag{4.13}$$

Combining (4.10) and (4.13) gives

$$\langle g(t),d\rangle \geqq -m_2|d|^2 \quad . \tag{4.14}$$

Either by (4.12) or by (4.14) and the fact that $m_2 < m_1$ we have

$$\langle g(t),d\rangle \geqq -m_2|d|^2 > -m_1|d|^2 \quad ,$$

so $t \in RT$. □

In order to derive convergence results for the algorithm in the next section we need the following lemma, which does not depend on the convergence assumptions of section 5. It gives the reason for augmenting G with a g_R satisfying (4.4) where $m_1 < 1$. A similar result for $m_1 \leqq 1/2$ is given in [39].

Lemma 4.4. Let $d = -Nr(G)$ be a search direction used at Step 1 to generate a g_R that is added to G at Step 2b to form $G_+ = G \cup \{g_L, g_R\}$ and suppose no g_j is deleted from G_+ at Step 2c. Let $d_+ = -Nr(G_+)$ be computed at Step 2d and suppose $c \geqq \max\{|g_j| : g_j \in G_+\}$. Then

$$|d_+|^2 \leqq |d|^2 \max\{m_1, 1-[(1-m_1)^2|d|^2/4c^2]\} \quad .$$

Proof: By assumption

$$|d_+| = |Nr(G_+)| \leqq |Nr(G \cup \{g_R\})| \leqq |Nr(\{-d, g_R\})| \quad .$$

So,

$$|d_+|^2 \leqq \min_{0 \leqq \mu \leqq 1} |\mu(-d) + (1-\mu)g_R|^2 \quad . \tag{4.15}$$

Let

$$a = \langle g_R, g_R+d\rangle = |g_R|^2 + \langle g_R, d\rangle \tag{4.16}$$

and

$$b = \langle d, g_R + d \rangle = |d|^2 + \langle g_R, d\rangle \quad . \tag{4.17}$$

Recall that $0 < m_1 < 1$ and $|d| \neq 0$, so by (4.4), (4.16) and (4.17) we have

$$b \geq (1 - m_1)|d|^2 > 0 \tag{4.18}$$

and

$$a + b = |g_R + d|^2 > 0 \quad . \tag{4.19}$$

So, for $\mu \in R$,

$$|\mu(-d) + (1-\mu)g_R|^2 = |g_R|^2 - 2a\mu + (a+b)\mu^2$$

is a strictly convex function of μ with a global minimum at

$$\mu = a/(a+b)$$

and, therefore, by (4.18) and (4.19), with a constrained minimum for $\mu \in [0,1]$ at

$$\mu = \begin{cases} a/(a+b) & \text{if } a \geq 0 \\ 0 & \text{if } a \leq 0 \end{cases} \quad .$$

So, if $a \leq 0$, then, by (4.16) and (4.4),

$$\min_{0 \leq \mu \leq 1} |\mu(-d) + (1-\mu)g_R|^2 = |g_R|^2 \leq -\langle g_R, d\rangle \leq m_1|d|^2 \quad . \tag{4.20}$$

Suppose $a > 0$. Then

$$\min_{0 \leq \mu \leq 1} |\mu(-d) + (1-\mu)g_R|^2 = |g_R|^2 - a^2/(a+b) \quad . \tag{4.21}$$

From (4.16) and (4.17) we have

$$a - b = |g_R|^2 - |d|^2 \quad ,$$

so

$$(a^2 - b^2)/(a+b) = |g_R|^2 - |d|^2 \quad ,$$

or

$$|g_R|^2 - a^2/(a+b) = |d|^2 - b^2/(a+b) \quad .$$

Thus, from (4.18) and (4.19),

$$|g_R|^2 - a^2/(a+b) \leqq |d|^2 - (1-m_1)^2|d|^4/(a+b) \quad . \tag{4.22}$$

By assumption $c \geqq \max\ [|d|, |g_R|]$, so, by (4.16), (4.17) and the Cauchy-Schwartz inequality,

$$a + b \leqq 4c^2 \quad . \tag{4.23}$$

Combining (4.21), (4.22) and (4.23) gives

$$\min_{0 \leqq \mu \leqq 1} |\mu(-d) + (1-\mu)g_R|^2 \leqq |d|^2\{1 - [(1-m_1)^2|d|^2/4c^2]\}. \tag{4.24}$$

The desired result then follows from (4.15), (4.20) and (4.24). □

Remarks: Lemma 4.4 also holds if any g_j is deleted from G_+ for which $\lambda_j = 0$ where $-d = Nr(G) = \sum_{g_i \in G} \lambda_i g_i$, $\sum_i \lambda_i = 1$ and $\lambda_i \geqq 0$ for all i. Thus, such g_j may also be deleted at Step 2c and this device can be used to keep the number of elements in G bounded, because, by Caratheodory's Theorem, Nr(G) can be expressed as a convex combination of $n+1$ or less elements of G.

Lemma 4.4 also holds if $G_+ = G \cup \{g_R\}$, so g_L need not be added to G at Step 2b, but in order to implement deletion rule IIb g_L must be saved, because it replaces g_x when x_L replaces x.

We conclude from Lemma 4.4 that $|d_+|$ is less than a fraction of $|d|$ and that if there is an infinite number of consecutive iterations where each -Nr(G) computed at Step 2d is a search direction

d, no significant g_j is deleted from G and all $|g_j|$ are uniformly bounded then $|d| \to 0$. This idea is used in the next section to show that $\delta \downarrow 0$ when $f(x)$ and $g \in M(y)$ are uniformly bounded for all x and y generated by the algorithm.

5. CONVERGENCE OF THE ALGORITHM

Throughout this section we assume that each execution of the line search procedure of Step 1 terminates and that the following boundedness assumption holds:

$$\text{There exists a positive number } C \text{ such that}$$
$$|g| \leq C \text{ for all } y \in S \text{ and } g \in M(y) \quad . \tag{5.1}$$

Note that if S is bounded then a value for C is $\sup \{|g| : g \in M(y), y \in S\}$ which is finite, because, by Proposition 7, M is bounded on bounded subsets of R^n. Under this assumption Lemma 4.2 implies that all g_j generated by the algorithm satisfy $|g_j| \leq C$.

The next result is the principal lemma from which the various convergence theorems dealing with stationarity and optimality follow. It is the only result in this section that does not depend on which deletion rule is used by the algorithm.

Lemma 5.1. Suppose (5.1) holds. Then either $\delta \downarrow 0$ or $f(x) \downarrow -\infty$.

Proof: There exists a number $\bar{\delta} \geq 0$ such that $\delta \downarrow \bar{\delta}$, because the successive values of δ are positive and form a monotone non-increasing sequence.

Suppose $\bar{\delta} > 0$. We must show that $f(x) \downarrow -\infty$. Define sequences $\{x_k\}$ and $\{\delta_k\}$ by setting $k = -1$ at Step 0 and, at entry to Step 1 replacing k by $k+1$ and then setting $x_k = x$ and $\delta_k = \delta$. Note that the loop consisting of Steps 2c-2d-2e-2c cannot be executed infinitely often, because, since $\beta_1 \delta^{q-1} \leq \beta_1 |g_0|^{q-1} < 1$, the δ-change at Step 2e would imply that $\delta \downarrow 0$, a contradiction.

Thus, the sequences $\{x_k\}$ and $\{\delta_k\}$ are infinite, $\{\delta_k\} \downarrow \bar{\delta}$ and we may assume without loss of generality that all exits from Step 2e are to Step 1. Now we show, by contradiction, that $\{f(x_k)\} \downarrow -\infty$.

Suppose $\{f(x_k)\}$ is bounded from below. From (4.2) with $x_{k+1} = x_L$ and $x_k = x$ we have that

$$f(x_{k+1}) - f(x_k) \leq -m_2 |x_{k+1} - x_k||d| \tag{5.2}$$

where, by Step 2e and the monotonicity of $\{\delta\}$,

$$|d| \geq \delta_k \geq \bar{\delta} > 0 \quad . \tag{5.3}$$

Thus, $\{f(x_k)\}$ is monotone nonincreasing. So, there exists a real number $\bar{f}$ such that $\{f(x_k)\} \downarrow \bar{f}$. By (5.2) and (5.3), for $i < \ell$ we have

$$f(x_\ell) - f(x_i) = \sum_{k=i}^{\ell-1} (f(x_{k+1}) - f(x_k)) \leq -m_2 \bar{\delta} \sum_{k=i}^{\ell-1} |x_{k+1} - x_k| .$$

Therefore, by the definition of $\bar{f}$ and the triangle inequality we have for $i \leq \ell$

$$\bar{f} - f(x_i) \leq f(x_\ell) - f(x_i) \leq -m_2\bar{\delta}|x_\ell - x_i| \quad . \tag{5.4}$$

Since $\alpha_2 < \alpha_1$, we may choose η such that $(\alpha_2/\alpha_1) < \eta < 1$. Then, since $\{\delta_k\} \downarrow \bar{\delta} > 0$ and $\{f(x_k)\} \downarrow \bar{f}$, there exists an integer I such that for all $i > I$

$$\alpha_2 \delta_{i-1} \leq \eta\alpha_1\bar{\delta} \tag{5.5}$$

and

$$f(x_i) - \bar{f} \leq f(x_I) - \bar{f} \leq m_2(1 - \eta)\alpha_1\bar{\delta}^2 \quad . \tag{5.6}$$

So, by (5.4) and (5.6), for $\ell \geq i > I$

$$|x_\ell - x_i| \leq (1 - \eta)\alpha_1\bar{\delta} \quad . \tag{5.7}$$

Consider any g_j that enters G after the definition of x_I, i.e. there is an $i \geq I + 1$ such that $x_{i-1} = x$, $\delta_{i-1} = \delta$, $x_i = y_L$ and the y_j associated with g_j equals y_L or y_R. By (4.3) and (5.5), we have

$$|x_i - y_j| \leq \alpha_2\delta_{i-1} \leq \eta\alpha_1\bar{\delta} \quad . \tag{5.8}$$

If such a g_j is deleted from G then, by Step 2c, there exists an $\ell \geq i$ such that

$$|x_\ell - y_j| > \alpha_1\delta \geq \alpha_1\bar{\delta} \quad .$$

But, by the triangle inequality, (5.7) and (5.8), we have

$$|x_\ell - y_j| \leq |x_\ell - x_i| + |x_i - y_j| \leq (1-\eta)\alpha_1\bar{\delta} + \eta\alpha_1\bar{\delta} = \alpha_1\bar{\delta} \quad ,$$

which is a contradiction. Thus, no such g_j is deleted from G, so the only candidates for deletion from G are the finite number of g_j's that entered G at or before the definition of x_I. Therefore, there are an infinite number of consecutive iterations where G is replaced by $G \cup \{g_L, g_R\}$, no g_j is deleted from G and, hence, by Lemma 4.4, since $|g_j| \leq C$ for all j,

$$|Nr(G)| = |d| \to 0 \quad .$$

But this contradicts (5.3). So, $\{f(x_k)\} \downarrow -\infty$ when $\bar{\delta} > 0$. □

From here on we assume $f(x) \not\downarrow -\infty$, so, by Lemma 5.1, $\delta \to 0$ and, thus, for infinitely many algorithm variable triples (x, G, δ) at Step 2e we have $|Nr(G)| < \delta$. Each time $|Nr(G)| < \delta$ occurs let an integer sequence index k be increased by 1 and define sequence quantities $x^k = x$, $G^k = G$ and $\delta^k = \delta$. Note that $\{|Nr(G^k)|\} \to 0$, since $\{\delta^k\} \to 0$. Also, note that these sequences do not necessarily correspond to the ones defined in the previous proof.

Our first convergence result shows stationarity of accumulation points of $\{x^k\}$, when deletion rule I is used. Consider the following condition:

> f is bounded from below on S_0 and there exists an $\bar{x} \in S_0$ and an infinite set $K \subseteq \{1,2,\ldots,\}$ such that $\{x^k\}_{k\in K} \to \bar{x}$. (5.9)

Remark: By the continuity of f and h, (5.9) holds if S_0 is bounded, for then S_0 is also closed and, hence, compact. Also note that the continuity of h implies $h(\bar{x}) \leq 0$.

Theorem 5.2. Suppose that (5.1) *and* (5.9) *hold and that the algorithm uses deletion rule* I. *Then* $h(\bar{x}) \leq 0$ *and* $0 \in M(\bar{x})$, *i.e.,* $\bar{x}$ *is stationary.*

Proof: For each $k \in K$, by Caratheodory's theorem, there exists a positive integer $p^k \leq n+1$ such that

$$Nr(G^k) \in \text{conv}\,(\bigcup_{\ell=1}^{p^k} \{g_\ell^k\}) \subset \text{conv}\,(\bigcup_{\ell=1}^{p^k} M(y_\ell^k))$$

where for each $\ell \in \{1,2,\ldots,p^k\}$, there is a j depending on ℓ such that $g_\ell^k = g_j$, $y_\ell^k = y_j$ and $g_j \in M(y_j) \cap G^k$. Then there exists an infinite set $K_1 \subset K$ and an integer $p \in \{1,2,\ldots,n+1\}$ such that $p^k = p$ for all $k \in K_1$, and, thus,

$$Nr(G^k) \in \text{conv}\,(\bigcup_{\ell=1}^{p} M(y_\ell^k)) \text{ for all } k \in K_1 \quad . \tag{5.10}$$

By assumption (5.1) and Proposition 7, M is bounded and uppersemicontinuous on S, so, the map $T : S^p \to 2^{R^n}$ defined by

$$T(z_1,z_2,\ldots,z_p) = \text{conv}\,(\bigcup_{\ell=1}^{p} M(z_\ell)) \text{ for } (z_1,z_2,\ldots,z_p) \in S^p \tag{5.11}$$

is uppersemicontinuous on S^p. By deletion rule I

$$|x^k - y_\ell^k| \leq \alpha_1 \delta^k \text{ for each } \ell \in \{1,2,\ldots,p\} \text{ and } k \in K_1 \quad .$$

Thus, since $\{x^k\}_{k\in K} \to \bar{x} \in S$, $\{\delta^k\} \to 0$ and $K_1 \subset K$,

$$\{y_\ell^k\}_{k\in K_1} \to \bar{x} \text{ for each } \ell \in \{1,2,\ldots,p\} \quad . \tag{5.12}$$

Combining (5.10), (5.11) and (5.12) with the facts that T is uppersemicontinuous on S^p and $\{|Nr(G^k)|\} \to 0$ gives

$$0 \in \text{conv}\,(\bigcup_{\ell=1}^{p} M(\bar{x})) = \text{conv}\,(M(\bar{x})) \quad .$$

By definition , $M(\bar{x})$ is convex, so $0 \in M(\bar{x})$.□

Combining Theorem 5.2 with Theorem 9 of [27] gives the following:

Corollary 5.3. Suppose, in addition to the assumptions of Theorem 5.2, that f and h are semiconvex [27] on R^n. Then at least one of the following holds:

(a) $\bar{x}$ is optimal.

(b) $\{z \in R^n : h(z) < 0\}$ is empty.

The remaining convergence results are for convex problems, and, hence, assume the following condition:

$$f \text{ and } h \text{ are convex functions on } R^n \quad . \tag{5.13}$$

The first such result shows how an x generated by the algorithm approximates satisfaction of saddle point optimality conditions in terms of $Nr(G)$ and δ. This result parallels Theorem 1 in [39] for unconstrained problems and depends on our deletion rule II.

Theorem 5.4. Suppose (5.1) *and* (5.13) *hold, the algorithm uses deletion rule* II *and* x, G *and* δ *are algorithm variables at the end of Step* 2d. *Let* $J = \{j : g_j \in G \cap M(y_j),\ h(y_j) \leqq 0\}$, $\bar{J} = \{j : g_j \in G \cap M(y_j),\ h(y_j) > 0\}$, *and* $\lambda_j \geqq 0$ *for* $j \in J \cup \bar{J}$ *satisfy* $Nr(G) = \sum_{j \in J \cup \bar{J}} \lambda_j g_j$ *and* $\sum_{j \in J \cup \bar{J}} \lambda_j = 1$. *Define* $\lambda \in [0,1]$ *by* $\lambda = \sum_{j \in J} \lambda_j$. *Then for all* $z \in R^n$

(a) $$\lambda(f(x) - f(z)) + (1-\lambda)(h(x) - h(z)) \leqq \langle Nr(G), x-z \rangle + (1+\lambda)C\alpha_1\delta \quad ,$$

(b) $$\lambda(f(x) - f(z)) \leqq |Nr(G)||z-x| + 2C\alpha_1\delta \quad \text{if } h(z) \leqq 0 \quad ,$$

and

(c) $$\lambda = 1 \text{ if } h(x) \leqq -C\alpha_1\delta \quad .$$

Proof: Note that $\bar{J}$ may be empty, but J is nonempty, because x is feasible and $g_x \in G \cap M(x)$. Since $g_j \in G$ for $j \in J$ was not deleted at Step 2c by rule IIb and (3.3) was satisfied, because f is convex, we conclude that (3.4) was not satisfied. Therefore, since $\lambda_j \geqq 0$, we have

$$\lambda_j(f(y_j) - f(x)) + \lambda_j\langle g_j, x-y_j \rangle \geqq -\lambda_j\alpha_1\delta|g_x - g_j| \quad \text{for } j \in J \quad . \tag{5.14}$$

Similarly from (3.2) of rule IIa we have

$$\lambda_j\langle g_j, x-y_j\rangle \geqq -\lambda_j\alpha_1\delta|g_j| \text{ for } j\in\bar{J} \quad . \tag{5.15}$$

Also, since $h(y_j) > 0$ for $j\in\bar{J}$ and $h(x) \leqq 0$, we have

$$\lambda_j(h(y_j) - h(x)) \geqq 0 \text{ for } j\in\bar{J} \quad . \tag{5.16}$$

Adding (5.14) summed over $j\in J$ to (5.15) and (5.16) summed over $j\in\bar{J}$ and using the fact that $|g_j| \leqq C$ for all j gives

$$\sum_{j\varepsilon J}\lambda_j(f(y_j) - f(x)) + \sum_{j\varepsilon\bar{J}}\lambda_j(h(y_j) - h(x)) + \sum_{j\varepsilon J\cup\bar{J}}\lambda_j\langle g_j, x-y_j\rangle$$

$$\geqq -(2\sum_{j\varepsilon J}\lambda_j + \sum_{j\varepsilon\bar{J}}\lambda_j)C\alpha_1\delta \quad . \tag{5.17}$$

Since f and h are convex, $g_j\in\partial f(y_j)$ for $j\in J$ and $g_j\in\partial h(y_j)$ for $j\in\bar{J}$, the subgradient inequality implies that for any $z\in R^n$

$$\lambda_j(f(z) - f(y_j)) \geqq \lambda_j[\langle g_j, z-x\rangle + \langle g_j, x-y_j\rangle] \text{ for } j\in J \tag{5.18}$$

and

$$\lambda_j(h(z) - h(y_j)) \geqq \lambda_j[\langle g_j, z-x\rangle + \langle g_j, x-y_j\rangle] \text{ for } j\in\bar{J} \quad . \tag{5.19}$$

Adding (5.18) and (5.19) over $j\in J\cup\bar{J}$ gives

$$\sum_{j\varepsilon J}\lambda_j(f(z) - f(y_j)) + \sum_{j\varepsilon\bar{J}}\lambda_j(h(z) - h(y_j)) \geqq \langle\sum_{j\varepsilon J\cup\bar{J}}\lambda_j g_j, z-x\rangle$$

$$+ \sum_{j\varepsilon J\cup\bar{J}}\lambda_j\langle g_j, x-y_j\rangle \quad . \tag{5.20}$$

Adding (5.17) and (5.20), and noting that $\lambda = \sum_{j\varepsilon J}\lambda_j = 1 - \sum_{j\varepsilon\bar{J}}\lambda_j$ and $Nr(G) = \sum_{j\varepsilon J\cup\bar{J}}\lambda_j g_j$ gives for all $z\in R^n$,

$$\lambda(f(z)-f(x)) + (1-\lambda)(h(z)-h(x)) \geqq \langle Nr(G), z-x\rangle - (1+\lambda)C\alpha_1\delta$$

which is equivalent to the first desired result (a).

Now suppose $h(x) \leqq -C\alpha_1\delta$. We show (c) by showing that $\bar{J}$ is empty. Suppose $\bar{J}$ is nonempty, i.e., there is a y_j corresponding to $g_j \in G$ such that $h(y_j) > 0$. Then, by deletion rule IIa,

$$\langle g_j, x-y_j\rangle \geqq -\alpha_1\delta|g_j| \geqq -C\alpha_1\delta \geqq h(x) \quad . \tag{5.21}$$

Since $g_j \in \partial h(y_j)$, the convexity of h and (5.21) implies

$$h(y_j) + \langle g_j, x-y\rangle \leqq h(x) \leqq \langle g_j, x-y_j\rangle \quad .$$

Hence, $h(y_j) \leqq 0$, but this contradicts the supposition that $h(y_j) > 0$. Thus, $\bar{J}$ is empty, $\lambda = 1$, and (c) holds.

To establish (b), we note that if $h(z) \leqq 0$ then, by (a) and the Cauchy-Schwarz inequality

$$\lambda(f(x)-f(z)) \leqq |Nr(G)||z-x| + (1+\lambda)C\alpha_1\delta + (1-\lambda)(-h(x)) \quad . \tag{5.22}$$

If $\lambda = 1$, then (b) follows immediately from (5.22). If $\lambda < 1$ then, by (c), $-h(x) < C\alpha_1\delta$, which combined with (5.22) gives (b).□

Returning to the sequence $\{x^k\}$, we next show that any accumulation point $\bar{x}$ satisfies saddle-point conditions if the problem functions are convex and the algorithm uses deletion rule II. Define the sequence $\{\lambda^k\} \subset [0,1]$ corresponding to $\{(x^k, G^k, \delta^k)\}$ by letting $\lambda^k = \lambda$ where λ is the multiplier as in Theorem 5.4 corresponding to (x, G, δ) when the latter quantity equals (x^k, G^k, δ^k).

<u>*Theorem 5.5*</u>. *Suppose* (5.1), (5.9) *and* (5.13) *hold and the algorithm uses deletion rule* II. *Let* $\bar{\lambda} \in [0,1]$ *be any accumulation point of* $\{\lambda^k\}_{k\in K}$. *Then*

(a) $h(\bar{x}) \leqq 0$,

(b) $\bar{\lambda}(f(\bar{x}) - f(z)) + (1-\bar{\lambda})(h(\bar{x}) - h(z)) \leqq 0$ for all $z \in R^n$,

(c) $\bar{\lambda} = 1$ if $h(\bar{x}) < 0$,

(d) $\{z \in R^n : h(z) < 0\}$ is empty if $\bar{\lambda} = 0$,

and

(e) $\bar{x}$ is optimal if $\bar{\lambda} > 0$.

Proof: Part (a) follows from the remark following assumption (5.9).

Since $\{x_k\}_{k\in K} \to \bar{x}$, $\{|Nr(G^k)|\} \to 0$, $\{\delta^k\} \to 0$ and f and h are continuous, (a) of Theorem 5.4 with $(x,G,\delta,\lambda) = (x^k,G^k,\delta^k,\lambda^k)$ implies (b).

By (c) of Theorem 5.4, if $h(x^k) \leqq -C\alpha_1\delta^k$ then $\lambda^k = 1$. Thus, if $h(\bar{x}) < 0$, since $\{x^k\}_{k\in K} \to \bar{x}$, $\{\delta^k\} \to 0$ and h is continuous, we have $\lambda^k = 1$ for all k sufficiently large and, hence, $\bar{\lambda} = 1$. Thus, (c) holds.

Parts (d) and (e) are well-known [23] consequences of (a), (b) and (c).□

Theorem 5.4 shows that if x* is optimal and the multiplier λ is positive then

$$f(x) - f(x^*) \leqq (|Nr(G)||x-x^*| + 2C\alpha_1\delta)/\lambda \quad .$$

Under the stronger assumptions given below we can obtain upper bounds on the quantities $|x-x^*|$ and $1/\lambda$ in terms of $|Nr(G)|$ and δ.

Theorem 5.6. In addition to the assumptions of Theorem 5.4, suppose that x* *is optimal and that* f *is* ***strongly convex*** [30] *on* S_0 *i.e., there exists a number* $\mu > 0$ *such that*

$$f(\tfrac{1}{2}(y+z)) \leqq \tfrac{1}{2}f(y) + \tfrac{1}{2}f(z) - \tfrac{\mu}{2}|y-z|^2 \text{ for all } y,z \in S_0 \quad . \tag{5.23}$$

Then

(a) x^* *is the only optimal point*

and

(b) $$\lambda|x-x^*| \leqq \tfrac{1}{2}[|Nr(G)| + (|Nr(G)|^2 + 8C\alpha_1\mu\delta)^{1/2}]/\mu \quad .$$

Furthermore, if there exists $\hat{x} \in R^n$ *such that* $h(\hat{x}) < 0$ *then*

(c) $$\lambda \geqq (-h(\hat{x}) - |\hat{x}-x||Nr(G)| - 2C\alpha_1\delta)/(f(\hat{x}) - f(x^*) - h(\hat{x}))$$

where

(d) $$|\hat{x}-x| \leqq |\hat{x}-x^*| + [(f(x_0) - f(x^*))/\mu]^{1/2} \quad .$$

Proof: Note that, by the convexity of f and h, S_0 is a convex set so if y, z $\in S_0$ then $\frac{1}{2}(y+z) \in S_0$. Part (a) follows immediately from (5.23), by contradiction, if we suppose y and z to be two distinct optimal points.

Since x^* is optimal, (5.23) with $y = x$ and $z = x^*$ implies that

$$f(x^*) \leqq f(\tfrac{1}{2}(x+x^*)) \leqq \tfrac{1}{2}f(x) + \tfrac{1}{2}f(x^*) - \tfrac{\mu}{2}|x-x^*|^2 \quad .$$

Thus,

$$f(x) - f(x^*) \geqq \mu|x-x^*|^2 \quad . \tag{5.24}$$

Combining (5.24) and (b) of Theorem 5.4 with $z = x^*$ gives

$$|Nr(G)||x-x^*| + 2C\alpha_1\delta \geqq \lambda\mu|x-x^*|^2$$

which, when multiplied by $(\lambda/\mu) \geqq 0$, yields

$$0 \geqq t^2 - ut - v = \{t - \tfrac{1}{2}[u + (u^2+4v)^{1/2}]\}\{t - \tfrac{1}{2}[u - (u^2+4v)^{1/2}]\} \tag{5.25}$$

where $t = \lambda|x-x^*|$, $u = |Nr(G)|/\mu$ and $v = 2\lambda C\alpha_1\delta/\mu$. Considered as a function of t the right hand side of (5.25) is a strictly convex quadratic, so an upper bound on all t satisfying (5.25) is the root $\frac{1}{2}[u + (u^2+4v)^{\frac{1}{2}}]$. Thus, $t \leqq \frac{1}{2}[u + (u^2+4v)^{\frac{1}{2}}]$, which, by the definitions of t, u and v, implies (b), since $\lambda \leqq 1$ implies $v \leqq 2C\alpha_1\delta/u$.

Now suppose $h(\hat{x}) < 0$ and note that (c) holds if $\lambda = 1$, because $f(\hat{x}) - f(x^*) - h(\hat{x}) \geqq -h(\hat{x}) > 0$ implies that the right hand side of (c) is bounded above by one. So, suppose $\lambda < 1$, which by (c) of Theorem 5.4 implies

$$h(x) > -C\alpha_1\delta \quad . \tag{5.26}$$

From (a) of Theorem 5.4 with $z = \hat{x}$ and the Cauchy-Schwartz inequality we have

$$\lambda(f(x) - f(\hat{x})) + (1-\lambda)(h(x) - h(\hat{x})) \leqq |Nr(G)||\hat{x}-x| + (1+\lambda)C\alpha_1\delta \quad . \tag{5.27}$$

Combining (5.26) and (5.27) with the fact that $f(x^*) \leqq f(x)$ gives

$$\lambda(f(x^*) - f(\hat{x})) + (1-\lambda)(-C\alpha_1\delta - h(\hat{x})) \leqq |Nr(G)||\hat{x}-x| + (1+\lambda)C\alpha_1\delta \quad ,$$

which is equivalent to (c).

In order to have a lower bound on λ that does not depend on x we need an upper bound on

$$|\hat{x}-x| \leqq |\hat{x}-x^*| + |x^*-x| \quad . \tag{5.28}$$

Combining (5.28) and (5.24) with the fact that $f(x) \leqq f(x_0)$ gives the last desired result (d).□

Our final result shows that under the strong assumptions of Theorem 5.6 we have that the accumulation point existence condition (5.9) for $\{x^k\}$ holds with $K = \{1,2,\ldots\}$ and $\bar{x} = x^*$ and that all the accumulation points of $\{\lambda^k\}$ are bounded below by a positive number.

Corollary 5.7. If all the assumptions of Theorem 5.6 hold then

$$\liminf_{k\to\infty} \lambda^k \geqq (-h(\hat{x}))/(f(\hat{x}) - f(x^*) - h(\hat{x})) > 0$$

and $\{x^k\} \to x^*$.

Proof: The results follow immediately from (b),(c) and (d) of Theorem 5.6 with $(x,G,\delta,\lambda) = (x^k,G^k,\delta^k,\lambda^k)$, since $\{\delta^k\} \to 0$ and $\{|Nr(G^k)|\} \to 0$.□

ACKNOWLEDGMENT

I would like to thank Albert Feuer for his careful reading of an earlier version of this paper and Claude Lemarechal for his many helpful suggestions.

This research was partially done at the University of Oslo, while the author was on leave from Yale University.

The research was sponsored, in part, by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR-74-2695. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] Bertsekas, D.P. and S.K. Mitter, A Descent Numerical Method for Optimization Problems with Nondifferentiable Cost Functionals, *SIAM Journal on Control*, 11 (1973), 637-652.

[2] Clarke, F.H., Generalized Gradients and Applications, *Trans. Amer. Math. Soc.*, 205 (1975), 247-262.

[3] ___________, A New Approach to Lagrange Multipliers, *Mathematics of Operations Research*, 1 (1976), 165-174.

[4] Conn, A.R., Constrained Optimization Using a Nondifferentiable Penalty Function, *SIAM J. Numer. Anal.*, 10 (1973), 760-784.

[5] Demjanov, V.F., Algorithms for Some Minimax Problems, *J. Computer and System Sciences*, 2 (1968), 342-380.

[6] Eaves, B.C., *Nonlinear Programming Via Kakutani Fixed Points*, Working Paper No. 294, Center for Research Management Science, University of California, Berkeley, 1970.

[7] ___________ and R. Saigal, Homotopies for Computation of Fixed Points on Unbounded Regions, *Mathematical Programming*, 3 (1972), 225-237.

[8] Elzinga, J. and T.G. Moore, A Central Cutting Plane Algorithm for the Convex Programming Problem, *Mathematical Programming*, 8 (1975), 134-145.

[9] Evans, J.P., Gould, F.J., and J.W. Tolle, Exact Penalty Functions in Nonlinear Programming, *Mathematical Programming*, 4 (1973), 72-97.

[10] Feuer, A., *An Implementable Mathematical Programming Algorithm for Admissible Fundamental Functions*, Ph.D. Dissertation, Department of Mathematics, Columbia University, New York, April, 1974.

[11] _________, Minimizing Well-Behaved Functions, *Proceedings of Twelfth Annual Allerton Conference on Circuit and System Theory*, Illinois, October 1974, 25-34.

[12] Garcia-Palomares, U., *Superlinearly Convergent Algorithms for Nonlinear Programming*, Ph.D. Dissertation, Computer Sciences Department, University of Wisconsin, Madison, 1972.

[13] Goldstein, A.A., Optimization with Corners, in *Nonlinear Programming 2*, Academic Press, New York, 1975, 215-230.

[14] _______________, Optimization of Lipschitz Continuous Functions, *Mathematical Programming*, to appear.

[15] Hansen, T., *A Fixed Point Algorithm for Approximating the Optimal Solution of a Concave Programming Problem*, Cowles Foundation Discussion Paper No. 277, Yale University, New Haven, 1969.

[16] ________ and H. Scarf, *On the Applications of a Recent Combinatorial Algorithm*, Cowles Foundation Discussion Paper No. 272, Yale University, New Haven, 1969.

[17] Hestenes, M.R. and E. Stiefel, Methods of Conjugate Gradients for Solving Linear Systems, *J. Research National Bureau of Standards*, 49 (1952), 409-436.

[18] Lebourg, G., Valeur moyenne pour gradient généralisé, *C.R. Acad. Sc. Paris*, 281 (1975), 795-797.

[19] Lemarechal, C., An Algorithm for Minimizing Convex Functions, in J.L. Rosenfeld, ed., *Information Processing*, North-Holland, Amsterdam, 1974, 552-556.

[20] ____________, Minimization of Nondifferentiable Functions with Constraints, *Proceedings of Twelfth Annual Allerton Conference on Circuit and System Theory*, Illinois, October 1974, 16-24.

[21] ____________, An Extension of Davidon Methods to Nondifferentiable Problems, in M.L. Balinski and P. Wolfe, eds., *Nondifferentiable Optimization*, Mathematical Programming Study 3, North-Holland, Amsterdam, 1975, 95-109.

[22] Lenard, M., Convergence Conditions for Restarted Conjugate Gradient Methods with Inaccurate Line Searches, *Mathematical Programming*, 10 (1976), 32-51.

[23] Mangasarian, O.L., *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[24] ______________, Dual, Feasible Direction Algorithms, in A.V. Balakrishnan, ed., *Techniques of Optimization*, Academic Press, New York, 1972, 67-88.

[25] McCormick, G.P. and K. Ritter, Alternate Proofs of the Convergence Properties of the Conjugate Gradient Method, *Journal of Optimization Theory and Applications*, 13 (1974), 497-518.

[26] Merrill, O.H., *Applications and Extensions of an Algorithm that Computes Fixed Points of Certain Upper Semicontinuous Point to Set Mappings*, Ph.D. Dissertation, University of Michigan, Ann Arbor, 1972.

[27] Mifflin, R., *Semismooth and Semiconvex Functions in Constrained Optimization*, RR-76-21, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1976; to appear in *SIAM Journal on Control and Optimization*, 1977.

[28] Pietrzykowski, T., An Exact Potential Method for Constrained Maxima, *SIAM J. Numer. Anal.*, 6 (1969), 217-238.

[29] Pironneau, O. and E. Polak, *Rate of Convergence of a Class of Methods of Feasible Directions*, Memo. No. ERL-M301, University of California, Berkeley, July 1971.

[30] Poljak, B.T., Existence Theorems and Convergence of Minimizing Sequences in Extremum Problems with Restrictions, *Doklady Academii Nauk SSR*, 166 (1966), 287-290, [in Russian; English transl.: *Soviet Mathematics Doklady*, 7 (1966), 72-75].

[31] ____________, A General Method of Solving Extremum Problems, *Doklady Akademii Nauk SSR*, 174 (1967), [in Russian; English transl.: *Soviet Mathematics Doklady*, 8 (1967), 593-597].

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

[33] Scarf, H., The Approximation of Fixed Points of a Continuous Mapping, *SIAM Journal on Applied Mathematics*, 15 (1967), 1328-1343.

[34] Shor, N.Z., *On The Structure of Algorithms for the Numerical Solution of Optimal Planning and Design Problems*, Dissertation, Cybernetics Institute AN, Kiev, 1964.

[35] __________, Utilization of the Method of Space Dilation in the Minimization of Convex Functions, *Cybernetics* (1970), 7-15.

[36] __________, Convergence Rate of the Gradient Descent Method with Dilation of the Space, *Cybernetics* (1970), 102-108.

[37] __________, A Class of Almost-Differentiable Functions and a Minimization Method for Functions of this Class, *Cybernetics* (1974), 599-606.

[38] Wolfe, P., *An Algorithm for the Nearest Point in a Polytope*, IBM Research Center Report, Yorktown Heights, New York, August, 1973.

[39] __________, A Method of Conjugate Subgradients for Minimizing Nondifferentiable Functions, in M.L. Balinski and P. Wolfe, eds., *Nondifferentiable Optimization*, Mathematical Programming Study 3, North-Holland, Amsterdam, 1975, 145-173.

[40] Zangwill, W.I., Nonlinear Programming Via Penalty Functions, *Management Science*, 13 (1967), 344-358.

[41] Zoutendijk, G., *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.