Revue française d'informatique et de recherche opérationnelle. Série rouge

E. Polak

On the convergence of optimization algorithms

Revue française d'informatique et de recherche opérationnelle. Série rouge, tome 3, nº R1 (1969), p. 17-34

<http://www.numdam.org/item?id=M2AN_1969__3_1_17_0>

R.I.R.O.
(3e année, N° 16, 1969, p. 17-34)

ON THE CONVERGENCE OF OPTIMIZATION ALGORITHMS

par E. POLAK

Department of Electrical Engineering and Computer Sciences, University of California, Berkeley and Institut Blaise-Pascal, 23, rue du Maroc, Paris (19e)

Résumé. — *Cet article présente un théorème de convergence pour une classe d'algorithmes de recherche de « point désirable », qui convient particulièrement à la synthèse et à l'obtention de bornes qualitatives de sensibilité des procédures de recherche. Pour illustrer son aptitude à la synthèse de nouveaux algorithmes, il est utilisé pour obtenir de nouvelles variantes des algorithmes de direction réalisable, gradient projeté et décomposition duale.*

INTRODUCTION

One of the greatest frustrations in the study of optimization algorithms is the almost total lack of a general theory. This lack is possibly due to the fact that algorithms are inventions and that their convergence proofs are usually done on an ad hoc basis. In response to this challenge, however, a few papers [1], [2], [3] have appeared in the last two years, in which attempts were made to extract, from available proofs, a number of principles governing the convergence of certain classes of algorithms.

The present paper is less concerned with the process of extracting general principles hidden in published convergence proofs than with the construction of a theory of algorithms which can be used to synthesize new methods or modify old ones. Specifically, it shows that certain forms of necessary conditions of optimality are particularly suitable for utilization in algorithms. Also, it presents a new convergence theorem (somewhat akin to theorems in [2] and [3], a particular case of which first appeared in [4] and which is particularly easy to use in the synthesis of new optimization algorithms. To illustrate its applicability, a few modifications of feasible directions [5] and gradient projection [6] algorithms are presented, as well as a new hybrid type algorithm and a new dual type algorithm. Its applicability to other algorithms is described in [4], [7], [8]. Thus, this convergence theorem opens up a new possibility for a unified study of a broad class of algorithms.

I. PRELIMINARY RESULTS

We shall restrict ourselves in this paper to the following canonical problem.

(1) **Problem :** Given continuously differentiable functions $f^0, f^1, ..., f^m$ from R^n into R^1, find a vector $\hat{x} \in R^n$ satisfying $f^i(\hat{x}) \leqslant 0$ for $i = 1, 2, ..., m$, such that

$$f^0(\hat{x}) = \min \{ f^0(x) \mid f^i(x) \leqslant 0, i = 1, 2, ..., m \}. \tag{2}$$

To be sure that (2) makes sense, we shall assume that either the set $\Omega = \{ x \mid f^i(x) \leqslant 0, i = 1, 2, ..., m \}$ is compact, or else that for every $\alpha \in R^1$ the set $\{ x \mid f^0(x) \leqslant \alpha \}$ is either bounded or empty.

(3) **Definition :** We shall call the elements of the constraint set Ω *feasible*, and we shall say that a vector $\hat{x} \in \Omega$ is *optimal* if it satisfies (2).

We begin by recalling a few characterizations of an optimal vector $\hat{x}$. These characterizations will subsequently be used in algorithm stop rules (see [10], [11]).

(4) **Theorem :** If $\hat{x}$ is optimal for (1) and S is any compact subset of R^n containing the origin in its interior, then

$$\min_{h \in S} \max_{i \in J_0(\hat{x})} \langle \nabla f^i(\hat{x}), h \rangle = 0, \tag{5}$$

where, for $\alpha \geqslant 0$ and any $x \in \Omega$,

$$J_\alpha(x) = \{ 0 \} \cup \{ i \mid f^i(x) + \alpha \geqslant 0, i \in \{ 1, 2, ..., m \} \}. \tag{6}$$

Proof: Suppose that (5) does not hold at an optimal $\hat{x}$, then there is a nonzero $h^* \in S$ such that

$$\min_{h \in S} \max_{i \in J_0(\hat{x})} \langle \nabla f^i(\hat{x}), h \rangle = \max_{i \in J_0(\hat{x})} \langle \nabla f^i(\hat{x}), h^* \rangle = -\delta \tag{7}$$

where $\delta > 0$. Hence there exists a $\lambda^* > 0$ such that

$$\begin{aligned} &f^i(\hat{x} + \lambda h^*) \leqslant 0 \text{ for } i \in \bar{J}_0(\hat{x}) \text{ and } \lambda \in (0, \lambda^*], \\ &f^i(\hat{x} + \lambda h^*) - f^i(\hat{x}) < 0 \text{ for } i \in J_0(\hat{x}) \text{ and } \lambda \in (0, \lambda^*], \end{aligned} \tag{8}$$ (1)

i.e. any $x = \hat{x} + \lambda h^*$, $\lambda \in (0, \lambda^*]$ is feasible and results in a lower cost than $\hat{x}$, which contradicts the optimality of $\hat{x}$. Q.E.D.

(9) **Remark :** If Ω has no interior, then (5) is satisfied at all $x \in \Omega$, which makes (5) a useless condition in this case.

---

(1) $\bar{J}_0(\hat{x})$ denotes the complement of $J_0(\hat{x})$ in $0, 1, 2, ..., m$.

(10) **Corollary :** If $\hat{x}$ is optimal for (1), then there exist multipliers $\xi^0 \leqslant 0$, $\xi^1 \leqslant 0, \ldots, \xi^m \leqslant 0$, not all zero, such that

(11)
$$\sum_{i=0}^{m} \xi^i \nabla f^i(\hat{x}) = \theta$$
$$\xi^i f^i(\hat{x}) = 0 \text{ for } i = 1, 2, \ldots, m.$$

Proof: Let F be a matrix whose rows are $\nabla f^i(\hat{x})$, $i \in J_0(\hat{x})$, and let p be the cardinality of $J_0(\hat{x})$. Then, by (5), the subspace $FR^n = \{ y \mid y = Fx, x \in R^n \}$ must be separated from the convex cone $\{ y \mid y \leqq 0 \} \subset R^p$, i.e. there exists a nonzero vector $\xi \in R^p$ such that

(12)
$$\langle \xi, y \rangle \geqslant 0 \quad \text{for all} \quad y \leqq 0, y \in R^p,$$
$$\langle \xi, Fx \rangle = 0 \quad \text{for all} \quad x \in R^n.$$

Assuming that the components of $y \in R^p$ and $\xi \in R^p$ are numbered with indices from $J_0(\hat{x})$, rather than consecutively, (12) yields

(13)
$$\xi^i \leqslant 0 \text{ for } i \in J_0(\hat{x})$$
$$\sum_{i \in J_0(\hat{x})} \xi^i \nabla f^i(\hat{x}) = 0.$$

Setting $\xi^i = 0$ for all $i \in \bar{J}_0(\hat{x})$, we now get (10) and (11). Q.E.D.

(14) **Theorem :** Suppose, in addition to the assumptions stated in (1) that the functions f^i, $i = 0, 1, 2, \ldots, m$ are convex and that Ω has an interior. Then any vector $\hat{x} \in \Omega$ satisfying (5) is optimal.

Proof: Suppose (5) is satisfied at a non-optimal $\hat{x} \in \Omega$ and let x_0 be any point in the interior of Ω. Then there exists a $x^* \in \Omega$ such that $f^0(x^*) < f^0(\hat{x})$, and, for some $\lambda \in (0, 1)$, the point $x_1 = \lambda x_0 + (1 - \lambda)x^*$ is in the interior of the set $\{ x \mid f^0(x) - f^0(\hat{x}) \leqslant 0, f^i(x) \leqslant 0, i = 1, 2, \ldots, m \}$. Hence, by convexity of the f^i we obtain

(15) $$\langle \nabla f^i(\hat{x}), x_1 - \hat{x} \rangle \leqslant f^i(x_1) - f^i(\hat{x}) < 0 \text{ for } i \in J_0(\hat{x}).$$

But $\alpha(x_1 - \hat{x}) \in S$ for some $\alpha > 0$, and hence (15) contradicts (5). Q.E.D.

(16) **Corollary :** Under the assumptions of theorem (14), any $\hat{x} \in \Omega$ which satisfies (11) for some multipliers $\xi^0 < 0$, $\xi^1 \leqslant 0, \ldots, \xi^m \leqslant 0$ (note $\xi^0 \neq 0$!) is optimal.

The proof of the above is trivial and therefore omitted.

In order to establish the convergence properties of the algorithms we are about to present, we shall need the following new theorem. The reader should note that it belongs to the same family of convergence results as theorems by Topkis and Veinott [3] and Zangwill [2]. However, the theorem below is more direct and more general than the Topkis and Veinott result and is easier to apply, though not quite as general as the Zangwill result.

(17) **Theorem** : Let T be a subset of R^n, let $c\colon T \to R^1$ be a « stop » function, and let a: $T \to T$ be a « search » function. Suppose that: (i) T contains *desirable* points which can be characterized by the fact that $\hat{x} \in T$ is desirable if and only if

(18) $$c(a(\hat{x})) \leqslant c(\hat{x})\,;$$

(ii) Either $c(\cdot)$ is continuous at all non-desirable $x \in T$ or else $c(x)$ is bounded from above for $x \in T$;

(iii) For every non-desirable $x \in T$ there exists a $\varepsilon(x) > 0$ and a $\delta(x) > 0$ such that

(19) $$c(a(x')) - c(x') \geqslant \delta \qquad \text{for all} \qquad x' \in T,\ \|x - x'\| \leqslant \varepsilon.$$

Let $\{ x_i \}$ be a sequence in T constructed according to the rule

(20) $$x_{i+1} = a(x_i),\ i = 0, 1, 2, \ldots$$

and satisfying

(21) $$c(x_{i+1}) > c(x_i).$$

Then, either $\{ x_i \}$ is finite and its last element is desirable, or $\{ x_i \}$ is infinite and every accumulation point of $\{ x_i \}$ is desirable.

Proof: Suppose that $\{ x_i \}$ is finite and that x_s is its last element. Then the construction of new elements must have stopped because $c[a(x_s)] \leqslant c(x_s)$, i.e., because x_s is desirable.

Now suppose that $\{ x_i \}$ is infinite and that $x_i \to x^*$ for $i \in K$, $K \subset \{ 0, 1, 2, \ldots, \}$, with x^* not desirable. Then there exist $\varepsilon^* > 0$ and $\delta^* > 0$ and an integer $k \in K$ such that for all $i \geqslant k$, $i \in K$,

(22) $$\|x_i - x^*\| \leqslant \varepsilon^*$$

and

(23) $$c(x_{i+1}) - c(x_i) \geqslant \delta^*.$$

Hence, for any two successive points x_i, x_{i+j}, i, $i + j \in K$, $i \geqslant k$, of the subsequence, we have

(24)

$$c(x_{i+j}) - c(x_i) = [c(x_{i+j}) - c(x_{i+j-1})] + \ldots + [c(x_{i+1}) - c(x_i)] > \delta^*.$$

But, because of (21) and (ii), $c(x_i) \to c^* < \infty$ for $i \in K$, $i \to \infty$, which is contradicted by (24). Hence each accumulation point of $\{ x_i \}$ must be desirable. Q.E.D.

We shall now show how the above theorem may be used to prove the convergence of some well known algorithms. It will be seen that the nature of theorem (17) is such that not only does it permit us to prove convergence of these algorithms but that it also enables us to establish certain qualitative bounds on deviations from the ideal subprocedures making up these algorithms, which are compatible with convergence. It will also be seen that it provides guidelines for the derivation of new algorithms from old ones.

II. METHODS OF FEASIBLE DIRECTIONS

In this section, we shall consider a class of methods introduced by Zoutendijk [5] together with some new modifications. We shall assume that the set Ω has an interior, since otherwise these methods make no sense as we shall soon see.

(25) *Definition*: For $\alpha \geqslant 0$, let $\varphi_\alpha : \Omega \to R^1$ be defined by

$$\varphi_\alpha(x) = \min_{h \in S} \max_{i \in J_\alpha(x)} \langle \nabla f^i(x), h \rangle$$

where $J_\alpha(x)$ is defined as in (6), and S is any given compact set containing the origin in its interior (note that when Ω has no interior $\varphi_\alpha(x) \equiv 0$).

(26) REMARK: To evaluate $\varphi_\alpha(x)$ we solve

$$\text{minimize } \sigma$$

(27) $$\text{subject to } \sigma - \langle \nabla f^i(x), h \rangle \geqslant 0 \quad \text{for} \quad i \in J_\alpha(x), h \in S.$$

The optimal pair $\sigma_\alpha(x)$, $h_\alpha(x)$ for this problem satisfies

$$\varphi_\alpha(x) = \sigma_\alpha(x), \; \sigma_\alpha(x) = \max_{i \in J_\alpha(x)} \langle \nabla f^i(x), h_\alpha(x) \rangle .$$

In solving (27), we shall always set $h_\alpha(x) = 0$ whenever $\sigma_\alpha(x) = 0$ and $h_\alpha(x)$ is not unique. Note that a *sensible* choice for S would be $S = \{ h \mid |h^i| \leqslant 1 \}$, or $S = \{ h \mid \|h\| \leqslant 1 \}$.

The algorithm we are about to present in the form of an idealized computer program will find points $\hat{x} \in \Omega$ such that $\varphi_0(\hat{x}) = 0$. Note that these algorithms are parametrized by the particular choice for the set S.

(28) **Algorithm** : Suppose that a $x_0 \in \Omega$ (1) and $\varepsilon > \varepsilon' > 0$ are given.

Step 1: Set $\varepsilon(x_0) = \varepsilon$ [We shall use the abbreviated notation $\varepsilon_0 = \varepsilon(x_0)$].

(1) To find a $x_0 \in \Omega$, solve, using the algorithm (28), the problem

$$\min \{ \sigma \mid f^i(x) - \sigma \leqslant 0, i = 1, 2, ..., m \},$$

with initial feasible point x', σ' where x' is arbitrary and

$$\sigma' = \max \{ f^i(x') \mid i = 1, 2, ..., m \}.$$

Since the optimal value $\hat{\sigma}$ for this problem satisfies $\hat{\sigma} < 0$, (28) will construct a $x_0 \in \Omega$ in a finite number of steps, provided Ω has an interior.

Step 2: Compute $\varphi_{\varepsilon_0}(x_0)$ and $h_{\varepsilon_0}(x_0)$ by solving (27) for $\alpha = \varepsilon_0$, $x = x_0$.

Step 3: If $\varphi_{\varepsilon_0}(x_0) < -\varepsilon_0$, set $h(x_0) = h_{\varepsilon_0}(x_0)$ and go to step 4.

If $\varphi_{\varepsilon_0}(x_0) \geqslant -\varepsilon_0$ and $\varepsilon_0 \leqslant \varepsilon'$ compute $\varphi_0(x_0)$.

If $\varphi_0(x_0) = 0$, set $x_0 = x_0$ and *Stop.*

If $\varphi_0(x_0) < 0$, set $\varepsilon_0 = \varepsilon_0/2$ and go to step 2.

If $\varphi_{\varepsilon_0}(x_0) \geqslant -\varepsilon_0$ and $\varepsilon_0 > \varepsilon'$, set $\varepsilon_0 = \varepsilon_0/2$ and go to step 2.

Step 4: Compute $\lambda(x_0) \geqslant 0$ such that

$$\lambda(x_0) = \max \{ \lambda \mid f^i(x_0 + \alpha h(x_0)) \leqslant 0 \text{ for all } \alpha \in [0, \lambda] \quad \text{and } i = 1, 2, ..., m \}$$

Step 5: Compute $\mu(x_0) \in [0, \lambda(x_0)]$ to be the smallest value in that interval such that

$$(30) \qquad f^0(x_0 + \mu(x_0)h(x_0)) = \min \{ f^0(x_0 + \mu h(x_0)) \mid \mu \in [0, \lambda(x_0)] \}.$$

Step 6: Set $x_0 = x_0 + \mu(x_0)h(x_0)$ and go to step 1.

(31) **Theorem:** Let $x_0, x_1, x_2, ...$ be a sequence in Ω constructed by the algorithm (28), i.e. $x_1, x_2, ...$ are the consecutive values assigned to x_0 in step 3 or step 1. Then, either the sequence $\{ x_i \}$ is finite and its last element, say x_k, satisfies $\varphi_0(x_k) = 0$ or else $\{ x_i \}$ is infinite and every accumulation point $\hat{x}$ in $\{ x_i \}$ satisfies $\varphi_0(\hat{x}) = 0$.

Proof: Obviously, the algorithm (28) defines a map a: $\Omega \to \Omega$. We shall show that this map together with the map $-f^0(-f^0$ taking the place of c and Ω the place of T) satisfy the assumptions of theorem (17). For the purpose of applying theorem (17) *we shall agree to call a point* $\hat{x} \in \Omega$ *desirable if* $\varphi_0(\hat{x}) = 0$.

First we must show that the characterization (18) is satisfied. Thus, suppose that $x_0 \in \Omega$ satisfies $\varphi_0(x_0) = 0$. Then, since for all $\varepsilon_0 > 0$, $J_{\varepsilon_0}(x_0) \supset J_0(x_0)$, we must have $-\varepsilon_0 < \varphi_0(x_0) \leqslant \varphi_{\varepsilon_0}(x_0)$. Hence, after a finite number of halving of ε_0 in step 3, the algorithm will find that $\varphi_0(x_0) = 0$ and will set $x_0 = x_0$, i.e. $a(x_0) = x_0$. This is in agreement with (18).

Now, given a point $x_0 \in \Omega$, the algorithm can only construct a new point x_1 such that $f^0(x_1) \leqslant f^0(x_0)$. Hence, suppose that the algorithm sets $x_1 = x_0$ (i.e. $x_0 = x_0$ in step 3 or step 6). If x_0 was reset to x_0 in step 3, $\varphi_0(x_0) = 0$. Suppose x_0 was reset to x_0 in step 6, i.e. $\mu(x_0)h(x_0) = 0$. But this implies that $\varphi_{\varepsilon_0}(x_0) = 0$, i.e. $\varphi_{\varepsilon_0}(x_0) \geqslant -\varepsilon_0$: a condition in step 3 which does not permit a continuation to step 6. Thus x_0 can only be reset to the value x_0 in step 3 and then it satisfies $\varphi_0(x_0) = 0$.

We shall now show that condition (19) is satisfied. Let $x_0 \in \Omega$ be any point such that $\varphi_0(x_0) < 0$. Then, from (30)

$$(32) \qquad f^0(x_0 + \mu(x_0)h(x_0)) - f^0(x_0) \triangleq -\delta_0$$

where $\delta_0 > 0$. Now, from (6) it follows that there must exist a $\rho' > 0$ such that

(33) $$J_{\varepsilon_0}(x) \subset J_{\varepsilon_0}(x_0) \text{ for all } x \in A(x_0, \rho'),$$

where $A(x_0, \rho') = \{ x \mid x \in \Omega, \|x - x_0\| \leqslant \rho' \}$ and ε_0 is the value of $\varepsilon(x_0)$ used in computing $h(x_0)$ in step 2. Let $M\colon R^n \to R^1$ be defined by

(34) $$M(x) = \min_{h \in S} \max_{i \in J_{\varepsilon_0}(x_0)} < \nabla f^i(x), h >.$$

Then M is continuous (1) and there is a $\rho'' > 0$ such that

(35) $$|M(x) - \varphi_{\varepsilon_0}(x_0)| \leqslant \varepsilon_0/2 \text{ for all } x \in A(x_0, \rho'').$$

Let $\rho = \min \{ \rho', \rho'' \}$, then, because of (33) and (35) and the fact that $\varphi_{\varepsilon_0}(x_0) \leqslant - \varepsilon_0$, we have, for all $x \in A(x_0, \rho)$, that

(36) $$\varphi_{\varepsilon_0}(x) \leqslant M(x) \leqslant - \varepsilon_0/2.$$

But $J_{\varepsilon_0/2}(x) \subset J_{\varepsilon_0}(x)$, and hence, for all $x \in A(x_0, \rho)$, we have

(37) $$\varphi_{\varepsilon_0/2}(x) \leqslant \varphi_{\varepsilon_0}(x) \leqslant - \varepsilon_0/2.$$

We therefore conclude that for all $x \in A(x_0, \rho)$ the algorithm (28) *will use a value $\varepsilon(x) \geqslant \varepsilon_0/2$ in computing $h(x)$ in step* 2, i.e. for all $x \in A(x_0, \rho)$ and for all $i \in J_{\varepsilon(x)}(x)$, $\langle \nabla f^i(x), h(x) \rangle \leqslant - \varepsilon_0/2$.

Now, for any $x \in A(x_0, \rho)$ and $i = 0, 1, 2, \ldots, m$, we have, by the mean value theorem, that

(38) $$f^i(x + \lambda h(x)) = f^i(x) + \lambda \langle \nabla f^i(x + \zeta h(x)), h(x) \rangle,$$

where $\zeta \in [0, \lambda]$. Since the functions $\langle \nabla f^i(\cdot), \cdot \rangle$, $i = 0, 1, 2, \ldots, m$, are uniformly continuous on $A(x_0, \rho) \times S$, for each $i \in \{ 0, 1, 2, \ldots, m \}$, there exists a $\lambda^i > 0$ such that

(39) $$|\langle \nabla f^i(x + \zeta h(x)), h(x) \rangle - \langle \nabla f^i(x), h(x) \rangle| \leqslant \varepsilon_0/4$$

for all $\zeta \in [0, \lambda^i]$. Similarly, since the functions $f^i(\cdot)$ are uniformly continuous on $A(x_0, \rho)$ and since S is compact, there exist $\bar{\lambda}^i > 0$, $i = 1, 2, \ldots, m$, such that

(40) $$|f^i(x + \zeta h(x)) - f^i(x)| \leqslant \varepsilon_0/2.$$

Now, for each $x \in A(x_0, \rho)$ and for each $i \in J_{\varepsilon(x)}(x)$, $\langle \nabla f^i(x), h(x) \rangle \leqslant - \varepsilon_0/2$, and for each $x \in A(x_0, \rho)$ and for each $i \in \bar{J}_{\varepsilon(x)}(x)$, $f^i(x) \leqslant - \varepsilon_0/2$. Hence, setting $\bar{\mu} = \min \{ \lambda^0, \lambda^1, \ldots, \lambda^m, \bar{\lambda}^1, \bar{\lambda}^2, \ldots, \bar{\lambda}^m \}$, we have, for any $x \in A(x_0, \rho)$

(41) $$\begin{aligned} &f^i(x + \bar{\mu} h(x)) - f^i(x) \leqslant - \bar{\mu}\varepsilon_0/4 \quad \text{for all} \quad i \in J_{\varepsilon(x)}(x); \\ &f^i(x + \bar{\mu} h(x)) \leqslant 0 \quad \text{for all} \quad i \in \bar{J}_{\varepsilon(x)}(x). \end{aligned}$$

---

(1) *Proposition:* Let $M(x) = \min_{y \in Y} g(x, y)$ where $g\colon R^n \times R^n \to R^1$ is continuous and $Y \subset R^n$ is compact. Then $M(\cdot)$ is continuous.

Since for all $x \in A(x_0, \rho)$ we must have $\mu(x) \geqslant \bar{\mu}$, we are led to the conclusion that

$$(42) \quad -f^0(x + \mu(x)h(x)) - (-f^0(x)) \geqslant \bar{\mu}\varepsilon_0/4, \qquad \text{for all} \quad x \in A(x_0, \rho),$$

i.e. that condition (19) is satisfied. This completes our proof.

We have already observed that by setting $S = \{ x \in R^n \mid |x^i| \leqslant 1 \}$, we can compute $\varphi_{\varepsilon(x)}(x)$ and $h(x)$ by solving a linear programming problem, i.e. these quantities are obtainable by finite step procedures. Thus, the weak link in the algorithm seems to be the requirement of solving equations of the form $f^i(x + \lambda h) = 0$ and of minimizing the function $f^0(\cdot)$ along the linear segment $\{ x \mid x = x_0 + \mu h(x_0), \mu \in [0, \lambda(x_0)] \}$. The following propositions which are obvious in the light of theorem (17), shows to what extent these operations may be approximated without affecting the convergence properties of the algorithm.

(43) **Proposition :** Suppose that in step 6 of the algorithm (28) x_0 is reset to $x_0 + \mu_0 h(x_0)$, where, for a fixed $\mu \in (0, 1)$, μ_0 satisfies

$$(44) \quad (f^0(x_0) - f^0(x_0 + \mu_0 h(x_0))) \geqslant \mu(f^0(x_0) - f^0(x_0 + \mu(x_0)h(x_0)).$$

Then theorem (31) remains valid.

(45) **Proposition :** Suppose that the $f^i(\cdot)$ are convex, that the sets $\{ x \mid f^i(x) \leqslant 0 \}$ are bounded for $i = 1, 2, ..., m$ and that steps 4 and 5 of the algorithm (28) are replaced by the steps 4′, 5′ below. Then theorem (31) still remains valid.

Step 4′: Compute $\lambda^0 > 0$, $\lambda^1 > 0$, $\lambda^2 > 0, ..., \lambda^m > 0$ to satisfy, for any $0 < \delta < 1/2$,

$$(1 - \delta)\lambda^0 \langle \nabla f^0(x_0), h(x_0) \rangle \leqslant f^0(x_0 + \lambda^0 h(x_0)) - f^0(x_0) \leqslant \delta\lambda^0 \langle \nabla f^0(x_0), h(x_0) \rangle;$$

$$\lambda^i\delta \langle \nabla f^i(x_0), h(x_0) \rangle \leqslant f^i(x_0 + \lambda^i h(x_0)) - f^i(x_0) \leqslant -f^i(x_0) \qquad \text{for} \quad i \neq 0, i \in J_{\varepsilon_0}(x_0);$$

$$-\lambda_i\delta \leqslant f^i(x_0 + \lambda^i h(x_0)) \leqslant 0, \qquad \text{for} \quad i \in \bar{J}_{\varepsilon_0}(x_0).$$

Step 5′: Set $\mu(x_0) = \min \{ \lambda^i \mid i \in \{ 0, 1, 2, ..., m \}\}$.

The introduction of ε_0 into the algorithm (28) ensures that for each non-optimal $x_0 \in \Omega$, there exists a $\rho > 0$ and a $\lambda_m > 0$ such that for all $x \in \Omega$, $\|x - x_0\| \leqslant \rho$, we have $x + \lambda h(x) \in \Omega$ for $\lambda \in [0, \lambda_m]$, i.e. it ensures a minimal step size about each non-optimal $x_0 \in \Omega$. This effect is obvious from the proof of theorem (31).

A second important, but not entirely independent, effect of using ε_0 in (28) is to ensure that we do not solve systems of simultaneous equations of the form $f^i(x) = 0$, $i \in I$, for points on the intersection of surfaces when these points are not optimal. The solution of such a system of nonlinear equations by gradient methods requires an infinite number of operations and hence solution points would become convergence points of a sequence $x_0, x_1, x_2, \ldots$ constructed by an algorithm not using an ε procedure. Thus, an algorithm would jam (or zigzag) without « the antijamming precautions » defined by the use of ε_0 in the algorithm (28).

III. GRADIENT PROJECTION METHODS

We shall now consider two variants of Rosen's gradient projection method [6]. These methods are particularly attractive when the constraint set Ω is a convex polytope with interior and $f^0(\cdot)$ is convex. When Ω has no interior, one simply restricts oneself to the linear manifold containing Ω.

(46) **Assumption :** We shall suppose that the cost function $f^0(\cdot)$ is convex and that the constraint functions $f^i(\cdot)$, $i = 1, 2, \ldots, m$ are of the form

$$f^i(x) = \langle f_i, x \rangle - b^i, \tag{47}$$

where $f_i \in R^n$ and $b^i \in R^1$. We also assume that the set

$$\Omega = \{ x \mid f^i(x) \leqslant 0, \, i = 1, 2, \ldots, m \}$$

has an interior.

(48) *Definition:* For every $x \in \Omega$ and $\alpha \geqslant 0$ let

$$I_\alpha(x) = \{ i \mid \langle f_i, x \rangle - b^i + \alpha \geqslant 0, i \in \{ 1, 2, \ldots, m \}\}.$$

(49) **Assumption :** We shall suppose that there exists a $\alpha^* > 0$ such that for every $x \in \Omega$ and $\alpha \in [0, \alpha^*]$ the vectors f_i, $i \in I_\alpha(x)$ are linearly independent.

(50) *Definition:* For every $\alpha \in [0, \alpha^*]$ and $x \in \Omega$ let

$$F_{I\alpha(x)} = (f_i)_{i \in I\alpha(x)} \tag{51}$$

be a matrix whose columns are f_i, $i \in I_\alpha(x)$ (ordered linearly on i). Let $P_{I\alpha(x)}$ be the matrix which projects R^n into the subspace spanned by the vectors f_i, $i \in I_\alpha(x)$, and let $P^\perp_{I\alpha(x)}$ (1) be the matrix which projects R^n into the subspace orthogonal to all the f_i, $i \in I_\alpha(x)$, i.e.

$$P_{I\alpha(x)} = F_{I\alpha(x)}(F^{\mathrm{T}}_{I\alpha(x)}F_{I\alpha(x)})^{-1}F^{\mathrm{T}}_{I\alpha(x)} \tag{52}$$

$$P^\perp_{I\alpha(x)} = I - P_{I\alpha(x)}. \tag{53}$$

(Note that matrices $P_{I\alpha(x)}$, $P^\perp_{I\alpha(x)}$ are symmetric and positive semidefinite.)

Consequently, for every $x \in \Omega$ and $\alpha \in [0, \alpha^*]$ we have

(54) $\nabla f^0(x) = P_{I\alpha(x)} \nabla f^0(x) + P^\perp_{I\alpha(x)} \nabla f^0(x) = F_{I\alpha(x)}\xi_\alpha(x) + P^\perp_{I\alpha(x)} \nabla f^0(x)$

(1) When $I_\alpha(x)$ is empty, we shall assume that $P_{I\alpha(x)}$ is the zero matrix and that $P^\perp_{I\alpha(x)}$ is the identity matrix.

where

(55) $$\xi_\alpha(x) = (F^{\mathrm{T}}_{I\alpha(x)}F_{I\alpha(x)})^{-1}F^{\mathrm{T}}_{I\alpha(x)} \nabla f^0(x).$$

It now follows directly from corollaries (10) and (16) that $\hat{x}$ is optimal if and only if

(56) $$P^{\perp}_{I0(\hat{x})} \nabla f^0(\hat{x}) = 0, \qquad \xi_0(\hat{x}) \leqq 0.$$

We make one more observation before stating an algorithm. Consider the expansion (54) and let $j \in I_\alpha(x)$. Then, from (54) (since

$$P^{\perp}_{I\alpha(x)-j}P^{\perp}_{I\alpha(x)} = P^{\perp}_{I\alpha(x)}),$$

(57) $$P^{\perp}_{I\alpha(x)-j} \nabla f^0(x) = \xi^j_\alpha(x)P^{\perp}_{I\alpha(x)-j}f_j + P^{\perp}_{I\alpha(x)} \nabla f^0(x),$$

and, since (57) is a decomposition into orthogonal components,

(58) $$\|P^{\perp}_{I\alpha(x)-j} \nabla f^0(x)\|^2 = (\xi^j_\alpha(x))^2 \|P^{\perp}_{I\alpha(x)-j}f_j\|^2 + \|P^{\perp}_{I\alpha(x)} \nabla f^0(x)\|^2.$$

Finally note that

(59) $$\langle f_j, P_{I\alpha(x)-j} \nabla f^0(x) \rangle = \xi^i_\alpha(x) \langle f_i, P^{\perp}_{I\alpha(x)-j}f_j \rangle.$$

(60) **Algorithm :** Suppose we are given a $\varepsilon \in [0, \alpha^*]$, with α^* as in (49), a $\varepsilon' \in (0, \varepsilon)$ and a $x_0 \in \Omega$.

Step 1: Set $\varepsilon(x_0) = \varepsilon$ [We shall use the abbreviated notation $\varepsilon_0 = \varepsilon(x_0)$].

Step 2: Compute

(61) $$h_{\varepsilon_0}(x_0) = P^{\perp}_{I\varepsilon_0(x_0)} \nabla f^0(x_0).$$

Step 3: If $\|h_{\varepsilon_0}(x_0)\|^2 > \varepsilon_0$, set $h(x_0) = -h_{\varepsilon_0}(x_0)$ and go to step 6.

If $\|h_{\varepsilon_0}(x_0)\|^2 \leqslant \varepsilon_0$ and $\varepsilon_0 \leqslant \varepsilon'$, compute $h_0(x_0)$ [as in (61)] and $\xi_0(x_0)$ [as in (55)].

If $\|h_0(x_0)\|^2 = 0$ and $\xi_0(x_0) \leqq 0$, set $x_0 = x_0$ and stop (x_0 is optimal). Otherwise, set $h(x_0) = -h_{\varepsilon_0}(x_0)$ and go to step 4.

If $\|h_{\varepsilon_0}(x_0)\|^2 \leqslant \varepsilon_0$ and $\varepsilon_0 > \varepsilon'$ go to step 4.

Step 4: Compute $\xi_{\varepsilon_0}(x_0)$ [as in (55)].

If $\xi_{\varepsilon_0}(x_0) \leqq 0$ set $h(x_0) = -h_{\varepsilon_0}(x_0)$ and go to step 5.

If $\xi_{\varepsilon_0}(x) > 0$, compute

(62) $$\bar{h}_{\varepsilon_0}(x_0) = P^{\perp}_{I\varepsilon_0(x_0)-j} \nabla f^0(x_0)$$

such that

(63) $$\|\bar{h}_{\varepsilon_0}(x_0)\| = \max_{i \in I\varepsilon_0(x_0)} \|P^{\perp}_{I\varepsilon_0(x_0)-i} \nabla f^0(x_0)\|.$$

$$\xi^i_{\varepsilon_0}(x_0) > 0$$

Set $h(x_0) = -\bar{h}_{\varepsilon_0}(x_0)$ and go to step 5.

Step 5: If $\|h(x_0)\|^2 \leqslant \varepsilon_0$, set $\varepsilon_0 = \varepsilon_0/2$ and go to step 2.

If $\|h(x_0)\|^2 > \varepsilon_0$ go to step 6.

Step 6: Compute $\mu(x_0) > 0$ to be the smallest value satisfying

(64) $f^0(x_0 + \mu(x_0)h(x_0)) = \min \{ f^0(x_0 + \mu h(x_0)) \mid (x_0 + \mu h(x_0)) \in \Omega \}$

Step 7: Set $x_0 = x_0 + \mu(x_0)h(x_0)$ and go to step 1.

(65) **Theorem :** Let $x_0, x_1, x_2, \ldots$ be a sequence in Ω constructed by the algorithm (60), i.e. $x_1, x_2, \ldots$ are the consecutive values assigned to x_0 in step 7. Then, either $\{ x_i \}$ is finite and its last element is optimal, or else $\{ x_i \}$ is infinite and every accumulation point of $\{ x_i \}$ is optimal. (When f^0 is strictly convex, the problem has a unique optimal solution $\hat{x}$ and then $x_i \rightarrow \hat{x}$.)

Proof: We shall again make use of theorem (17) under the assumption that $T = \Omega$, a: $\Omega \rightarrow \Omega$ is defined by the algorithm (60), and $c = -f^0$. We begin by showing that the characterization (18) is satisfied. Suppose x_0 is optimal. Then $h_0(x_0) = 0$ and $\xi_0(x_0) \lesseqgtr 0$. Now, for any $\varepsilon_0 > 0$, $I_{\varepsilon_0}(x_0) \supset I_0(x_0)$ and hence

$$\xi_{\varepsilon_0}(x_0) \lesseqgtr 0 \tag{66}$$

and

$$\|h_{\varepsilon_0}(x_0)\| = \|h_0(x_0)\| = 0. \tag{67}$$

Consequently, after a finite number of halvings of ε_0 in step 5, the algorithm will stop in step 3, ressetting x_0 to its original value. This satisfies (18).

By construction, the algorithm stops setting $x_0 = x_0$ in step 3 if and only if x_0 is optimal. This is the only possible condition for setting $x_0 = x_0$, since it is not possible to have $\mu(x_0)h(x_0) = 0$ in step 7 for the following reasons. First, $h(x_0) = 0$ is not allowed in step 6 and hence in step 7. Second, if $h(x_0) \neq 0$ then $\mu(x_0) \neq 0$, since for all

$i \in I_{\varepsilon_0}(x_0)$, $\langle h(x_0), f_i \rangle \leqslant 0$ and $\langle \nabla f^0(x_0), h(x_0) \rangle = -\|h(x_0)\|^2 < 0$.

We must now show that (19) is satisfied, i.e. that if $x_0 \in \Omega$ is not optimal. then there exists a $\rho > 0$ and $\delta > 0$ such that

$$-(f^0(x + \mu(x)h(x)) - f^0(x)) \geqslant \delta \quad \text{for all} \quad x \in \Omega, \|x - x_0\| \leqslant \rho. \tag{68}$$

Let ε_0 be the last value of $\varepsilon(x_0)$ (i.e. just before being reset again in step 1). Then, either

$$\|h_{\varepsilon_0}(x_0)\|^2 > \varepsilon_0, \tag{69}$$

or else

$$\|\bar{h}_{\varepsilon_0}(x_0)\|^2 > \varepsilon_0. \tag{70}$$

Suppose (69) took place, i.e. that $h(x_0) = -h_{\varepsilon_0}(x_0)$. Then there exists a $\rho' > 0$ such that

(71) $$\|P^{\perp}_{I_{\varepsilon_0}(x_0)} \nabla f^0(x)\|^2 \geqslant \varepsilon_0/2 \quad \text{for all} \quad x \in A(x_0, \rho')$$

(where, as before, $A(x_0, \rho') = \{ x \mid x \in \Omega \ \|x - x_0\| \leqslant \rho' \}$),

Let $\rho'' > 0$ be such that $I_{\varepsilon_0}(x) \subset I_{\varepsilon_0}(x_0)$ for all $x \in A(x_0, \rho'')$ and let $\rho = \min \{ \rho', \rho'' \}$. Then, for all $x \in A(x_0, \rho)$ and $\alpha \in [0, \varepsilon_0]$,

(72) $$\|P^{\perp}_{I\alpha(x)} \nabla f^0(x)\| \geqslant \|P^{\perp}_{I\varepsilon_0(x)} \nabla f^0(x)\| \geqslant \|P^{\perp}_{I_{\varepsilon_0}(x_0)} \nabla f^0(x)\| \geqslant \varepsilon_0/2.$$

We therefore conclude that if (69) took place, then for all $x \in A(x_0, \rho)$ the algorithm will use a final value of $\varepsilon(x) \geqslant \varepsilon_0/2$.

Now suppose that (70) took place, i.e. that $\bar{h}(x_0) = -h_{\varepsilon_0}(x_0)$. Then, either $\|h_{\varepsilon_0}(x_0)\| > 0$ or $\|h_{\varepsilon_0}(x_0)\| = 0$.

Suppose $\|h_{\varepsilon_0}(x_0)\| = \delta' > 0$. Let $\rho'' > 0$ be such that $I_{\varepsilon_0}(x) \subset I_{\varepsilon_0}(x_0)$ for all $x \in A(x_0, \rho'')$. Then there exists a $\rho \in (0, \rho'')$, such that for all $x \in A(x_0, \rho)$ and for all $\alpha \in [0, \varepsilon_0]$,

(73) $$\|h_\alpha(x)\| > \|P^{\perp}_{I\alpha(x)} \nabla f^0(x)\|^2 \geqslant \|P^{\perp}_{I\varepsilon_0(x)} \nabla f^0(x)\|^2 \geqslant \|P^{\perp}_{I_{\varepsilon_0}(x_0)} \nabla f(x)\|^2 \geqslant \delta'/2,$$

and hence for all $x \in A(x_0, \rho)$, the algorithm will set $\varepsilon(x) \geqslant [\delta'/2] > 0$ (1).

Now suppose that $\|h_{\varepsilon 0}(x_0)\| = 0$. Then $\nabla f^0(x_0) = \sum_{i \in I_{\varepsilon_0}(x_0)} \xi^i_{\varepsilon 0}(x_0) f_i$ and in this representation the coefficients are unique. Now let

$$\delta_1 = \min \{ \|P_I^{\perp} \nabla f^0(x_0)\| \mid I \subset I_{\varepsilon 0}(x_0), \quad \|P_I^{\perp} \nabla f^0(x_0)\| > 0 \}$$

(74) and

$$\delta_2 = \min \{ \max_{\substack{\xi^i_{\varepsilon_0}(x_0) > 0 \\ i \in I}} \|P^{\perp}_{I-i} \nabla f^0(x_0)\| \mid I \subset I_{\varepsilon 0}(x_0), \|P_I^{\perp} \nabla f^0(x_0)\| = 0 \} .$$

Obviously, $\delta_1 > 0$ and $\delta_2 > 0$. Let $\delta' = \min \{ \varepsilon_0, \delta_1, \delta_2 \}$, and, again, let $\rho'' > 0$ be such that $I_{\varepsilon 0}(x) \subset I_{\varepsilon 0}(x_0)$ for all $x \in A(x_0, \rho'')$. Then there exists a $\rho \in (0, \rho'')$ such that for all $x \in A(x_0, \rho)$ and all $x \in [0, \varepsilon_0]$, either

$$\|P^{\perp}_{I\alpha(x)} \nabla f(x)\|^2 \geqslant \delta'/2$$

(75) or

$$\max_{\substack{i \in I\alpha(x) \\ \xi^i_\alpha(x) > 0}} \|P_{I\alpha(x)-i} \nabla f(x)\|^2 \geqslant \delta'/2.$$

We therefore conclude that if (70) took place, then for all $x \in A(x_0, \rho)$, the algorithm will use a final value of $\varepsilon(x) \geqslant [\delta'/2] > 0$.

(1) Let k be an integer such that $\varepsilon/2^{k+1} \geqslant \delta'/2 \geqslant \varepsilon/2^k$. Then we define $[\delta'/2] = 2^{k+1}$.

Now, for all $x \in A(x_0, \rho)$, and for all $i \in I_{\varepsilon(x)}(x)$, $\langle f_i, h(x) \rangle \leqslant 0$ [see (59), (61)], and so, as far as these constraints are concerned, one can displace oneself an arbitrary amount in the direction $h(x)$ from x without violation. We now conclude (as in the case of the feasible directions algorithm) that there exists a $\lambda_m > 0$ such that $x + \lambda h(x)/\|h(x)\| \in \Omega$ for all $\lambda \in [0, \lambda_m]$ and $x \in \Omega$, $\|x - x_0\| \leqslant \rho$.

Next, we note that $\langle \nabla f^0(x), h(x) \rangle \leqslant -\varepsilon_0/2$ (or $-[\delta'/2]$) for all $x \in A(x_0, \rho)$ and that there exists a γ such that $\|h(x)\| \leqslant \gamma$ for all $x \in A(x_0, \rho)$. It now follows, by means of an argument essentially identical to the one following (31), in the proof of the feasible directions algorithm, that (68) is satisfied for some $\delta > 0$. This completes our proof.

Since $\langle \nabla f(x), h(x) \rangle = -\|h(x)\|^2$, one may wish to accelerate the algorithm (60) by increasing $\|h(x)\|$ as much as possible at each step. The following acceleration procedure is very easily seen as not affecting the convergence properties of the algorithm (60). (To account for it we need to modify the proof of theorem (65) only very slightly.)

Step 1′: [Acceleration procedure, to be inserted between step 1 and step 2 of (60)]:

Compute $\xi_{\varepsilon_0}(x_0)$, $\bar{h}_{\varepsilon_0}(x_0)$ [as in (55), (62)].

If $\xi_{\varepsilon_0}(x_0) \leqslant 0$ go to step 3.

If

$$\xi_{\varepsilon_0}(x_0) > 0 \text{ and } \|\bar{h}_{\varepsilon_0}(x_0)\| \geqslant 2\, \|h_{\varepsilon_0}(x_0)\| \text{ set } h(x_0) = \bar{h}_{\varepsilon_0}(x_0)$$

and go to step 5.

If $\xi_{\varepsilon_0}(x_0) > 0$ and $\|h_{\varepsilon_0}(x_0)\| < 2\, \|h_{\varepsilon_0}(x_0)\|$, go to step 2.

This concludes our discussion of the convergence of gradient projection methods. We shall next discuss methods which are a cross between gradient methods and methods of feasible directions.

IV. METHODS OF FEASIBLE DIRECTIONS WITH PROJECTION OPERATORS

In the algorithm (28), to obtain a « feasible direction » $h(x_0)$ we had to solve a minimization problem. In the algorithm (60) this process was replaced by the computation of a projection operator which, generally, is easier to calculate. However, algorithm (60) is only applicable to problems with linear inequality constraints. We shall now present a modification of (60) which applies to more general stituations. This modification was inspired by a closely related heuristic algorithm described in [9].

(76) **Assumption :** We shall suppose in this section that all the functions f^i, $i = 0, 1, 2, ..., m$ in (1) are convex and that the set

$$\Omega = \{ x \mid f^i(x) \leqslant 0, i = 0, 1, 2, ..., m \}$$

has an interior.

(77) **Assumption :** We shall suppose that there exists a $\alpha^* > 0$ such that for every $\alpha \in [0, \alpha^*]$ and $x \in \Omega$, the vectors $\nabla f^i(x)$, $i \in I_\alpha(x)$ are linearly independent [where $I_\alpha(x)$ was defined in (48)].

We shall retain in this section the notation introduced in the preceding one with the following, rather obvious modification. For every $\alpha \in [0, \alpha^*]$ and $x \in \Omega$ we shall let

(78) $$F_{I\alpha(x)} = (\nabla f^i(x))_{i \in I\alpha(x)}$$

be a matrix whose columns are the $\nabla f^i(x)$, $i \in I_\alpha(x)$ (ordered linearly on i). The projection matrices $P_{I\alpha(x)}$, $P^\perp_{I\alpha(x)}$ will still be defined by (52) and (53), respectively, with the matrix $F_{I\alpha(x)}$ now defined by (78), etc.

(79) **Algorithm :** Suppose we are given a $\varepsilon \in [0, \alpha^*]$ with α^* as in (77), an $\varepsilon' \in (0, \varepsilon)$ and a $x_0 \in \Omega$.

Step 1: Set $\varepsilon(x_0) = \varepsilon$. [We shall use the abbreviated notation $\varepsilon_0 = \varepsilon(x_0)$.]

Step 2: Compute

(80) $$h_{\varepsilon_0}(x_0) = P^\perp_{I\varepsilon_0(x_0)} \nabla f^0(x_0).$$

Step 3: If $\|h_{\varepsilon_0}(x_0)\|^2 > \varepsilon_0$, set $h(x_0) = - h_{\varepsilon_0}(x_0)$ and go to step 6.

If $\|h_{\varepsilon_0}(x_0)\|^2 \leqslant \varepsilon_0$ and $\varepsilon_0 < \varepsilon'$ compute $h_0(x_0)$ [with $\varepsilon_0 = 0$ in (80)] and $\xi_0(x_0)$ [as in (55)].

If $h_0(x_0) = 0$ and $\xi_0(x_0) \lesseqgtr 0$, set $x_0 = x_0$ (x_0 is optimal).

Otherwise set $h(x_0) = - h_{\varepsilon_0}(x_0)$ and go to step 4.

If $\|h_{\varepsilon_0}(x_0)\|^2 \leqslant \varepsilon_0$ and $\varepsilon_0 > \varepsilon'$ go to step 4.

Step 4: Compute $\xi_{\varepsilon_0}(x_0)$.

If $\xi_{\varepsilon_0}(x_0) \lesseqgtr 0$ set $h(x_0) = - h_{\varepsilon_0}(x_0)$ and go to step 5.

If $\xi_{\varepsilon_0}(x_0) > 0$, compute

(81) $$\bar{h}_{\varepsilon_0}(x_0) = P^\perp_{I\varepsilon_0(x_0)-j} \nabla f^0(x_0)$$

such that

(82) $$\|\bar{h}_{\varepsilon_0}(x_0)\| = \max_{\substack{i \in I\varepsilon_0(x_0) \\ \xi^i_{\varepsilon_0}(x_0) > 0}} \|P^\perp_{I\varepsilon_0(x_0)-i} \nabla f^0(x_0)\| .$$

Set $h(x_0) = - \bar{h}_{\varepsilon_0}(x_0)$ and go to step 5.

Step 5: If $\|h(x_0)\|^2 \leqslant \varepsilon_0$ set $\varepsilon_0 = \varepsilon_0/2$ and go to step 2.

If $\|h(x_0)\|^2 > \varepsilon_0$ go to step 6.

Step 6: Set $K_{\varepsilon_0}(x_0) = I_{\varepsilon_0}(x_0)$ when $h(x_0) = -h_{\varepsilon_0}(x_0)$ and set

$$K_{\varepsilon_0}(x_0) = I_{\varepsilon_0}(x_0) - j$$

when $h(x_0) = -\bar{h}_{\varepsilon_0}(x_0)$. Compute

(83) $$v(x_0) \times \beta(x_0)h(x_0) + F_{K_{\varepsilon_0}(x_0)}(F^T_{K_{\varepsilon_0}(x_0)}F_{K_{\varepsilon_0}(x_0)})^{-1}t$$

where $t = -\varepsilon_0(1, 1, ..., 1)$ and $\beta(x_0) \geqslant 1$ is the smallest positive scalar such that

(84) $$\langle \nabla f^k(x_0), v(x_0) \rangle \leqslant -\varepsilon_0$$

for $k = 0$ when $h(x_0) = -h_{\varepsilon_0}(x_0)$ and for $k = 0, j$ when $h(x_0) = -\bar{h}_{\varepsilon_0}(x_0)$.

Step 7: Compute $\lambda(x_0) > 0$ such that

(85) $\lambda(x_0) = \max \{ \lambda \mid f^i[x_0 + \zeta v(x_0)] \leqslant 0, \zeta \in [0, \lambda], i = 1, 2, ..., m \}$.

Step 8: Compute $\mu(x_0)$ to be the smallest value satisfying

(86) $f^0[x_0 + \mu(x_0)v(x_0)] = \min \{ f^0[x_0 + \mu v(x_0)] \mid \mu \in [0, \lambda(x_0)] \}$.

Step 9: Set $x_0 = x_0 + \mu(x_0)v(x_0)$ and go to step 1.

(87) REMARK : Note that the above algorithm differs from the algorithm (60) only in the operations defined in step 6.

(88) **Theorem :** Let $x_0, x_1, x_2, ...$ be a sequence in Ω constructed by the algorithm (79), i.e. $x_1, x_2, ...$ are the consecutive values assigned to x_0 in step 7. Then either $\{ x_i \}$ is finite and its last element is optimal, or else $\{ x_i \}$ is infinite and every accumulation point of $\{ x_i \}$ is optimal. (When either f^0 is strictly convex or Ω is strictly convex, or both, there is a unique optimal solution for the problem (1), and hence a unique accumulation point for the sequence $\{ x_i \}$, when infinite.)

Proof : Again, we shall simply show that the assumptions of theorem (17) are satisfied. We omit a demonstration that condition (18) is satisfied since in this case it is identical to the one given for algorithm (60) in the proof of theorem (65).

We shall now show that for every non-optimal $x_0 \in \Omega$, there exist a $\bar{\rho} > 0$ and a $\delta > 0$ such that

(89) $$-(f^0(x + \mu(x)v(x)) - f^0(x)) \geqslant \delta \text{ for all } x \in A(x_0, \bar{\rho}).$$

First, proceeding as in the proof of theorem (60), and, in addition, using the fact that the f^i are continuously differentiable, we can show that if $x_0 \in \Omega$ is not optimal, then there exists a $\rho > 0$ and a $\delta' > 0$ such that for all $x \in A(x_0, \rho)$

(90) $$\|h(x)\| \geqslant \left(\frac{\delta'}{2}\right) > 0,$$

i.e. $\varepsilon(x) \geqslant [\delta'/2]$, for all $x \in A(x_0, \rho)$. Next, we find that, by (84), for all $x \in A(x_0, \rho)$

(91) $$\langle \nabla f^0(x), v(x) \rangle \leqslant -\varepsilon(x) \leqslant -[\delta'/2]$$

and, if $K_{\varepsilon(x)}(x) \neq I_{\varepsilon(x)}(x)$ (say $K_{\varepsilon(x)}(x) = I_{\varepsilon(x)}(x) - j$),

(92) $$\langle \nabla f^j(x), v(x) \rangle \leqslant -\varepsilon(x) \leqslant -[\delta'/2].$$

Furthermore, for all $i \in K_{\varepsilon(x)}(x)$, $x \in \Omega$, $\|x - x_0\| \leqslant \rho$,

(93) $$\langle \nabla f^i(x), v(x) \rangle = -\varepsilon(x) \leqslant -[\delta'/2]$$

Finally, an inspection of (83), (84), (61) and (62) indicates that there exists $\bar{\rho} \in (0, \rho]$ and a $M \in (0, \infty)$ such that $\|v(x)\| \leqslant M$ for all $x \in A(x_0, \rho)$. The proof may now be completed by following the steps after (37) in the proof of theorem (31).

(94) Remark : The acceleration step 1′ proposed for algorithm (60) can also be utilized in the present algorithm.

We now turn to an entirely different type algorithm whose convergence can also be proved by means of theorem (17).

V. A DECOMPOSITION ALGORITHM

So far, we have presented a number of algorithms whose convergence was proved by setting $c = -f^0$ in theorem (17). In order to show that c may have to be chosen differently, we present a simple decomposition algorithm which is in the class discussed extensively in [7].

Consider the particular problem

(95) $$\text{minimize } \|x\|^2 \text{ subject to } Ax \in \Omega,$$

where $x \in R^N$, $\|x\|^2 = \sum_{i=1}^{N} (x^i)^2$, A is a $n \times N$ matrix with $N \geqslant n$, of rank n, and $\Omega \subset R^n$ is defined by

(96) $$\Omega = \{ z \in R^n \mid f^i(z) \leqslant 0, i = 1, 2, ..., m \}$$

and is assumed to be strictly convex and compact.

(97) *Definition* : Let $S = \{ z \in R^n \mid \|z\| = 1 \}$ and let $v : S \to \Omega$ be defined by $\langle z - v(s), s \rangle \leqslant 0$ for all $z \in \Omega$.

(98) *Definition* : Let $T = \{ s \in S \mid \langle s, v(s) \rangle \leqslant 0 \}$. Let $c : T \to R^1$ be defined by

(99) $$c(s) = \min \{ \|x\|^2 \mid \langle s, Ax - v(s) \rangle = 0 \},$$

and let $w : T \to R^n$ be defined by

(100) $$w(s) = Ax(s),$$

where $x(s) \in R^n$ is such that $c(s) = \|x(s)\|^2$, i,e. $x(s) = c(s)A^Ts / \|A^Ts\|$.

(101) Remark : It is shown in [7] that v, w, and c are continuous maps.

(102) **Algorithm :** Suppose that a $s_0 \in T$ is given.

Step 1: Compute $v(s_0)$, $c(s_0)$, and $w(s_0)$.

Step 2: If $v(s_0) = w(s_0)$, stop. [$c(s_0)$ is the minimum cost for (95) and $x(s_0)$ is the desired solution, i.e. $w(s_0) = Ax(s_0)$].

If $v(s_0) \neq w(s_0)$, compute $a(s_0)$, where a : $T \to T$ is defined by

$$(103) \begin{cases} a(s) \in \sigma(s) \triangleq \{ s' \in T \mid s' = \lambda s + \mu(w(s) - v(s)), \lambda, \mu \in (-\infty, +\infty) \}, \\ c(a(s)) = \max \{ c(s') \mid s' \in \sigma(s) \} \\ \text{and } \|s - a(s)\| \text{ is minimized (to make } a(s) \text{ unique).} \end{cases}$$

Set $s_0 = a(s_0)$ and go to step 1.

(104) **Theorem :** Let $s_0, s_1, s_2, \ldots,$ be a sequence of points in T generated by the algorithm (102), then either $\{ s_i \}$ is finite and its last element, s_k, is such that $c(s_k) = \min \{ \|x\|^2 \mid Ax \in \Omega \}$ and $x(s_k)$ is optimal for (95), or else $\{ s_i \}$ is infinite and $s_i \to \hat{s}$, where $c(\hat{s}) = \min \{ \|x\|^2 \mid Ax \in \Omega \}$ and $x(\hat{s})$ is optimal for (95).

It is shown in [7] that the map $(ca) : T \to R^1$ is continuous and hence that the maps c and a as defined by (99) and (103) respectively, satisfy the conditions of theorem (17).

For practical aspects of algorithms such as (102), i.e. methods of computing $v(s)$ and the effect on convergence of finitely calculable approximations to $v(s)$ and $a(s)$, the reader should consult [4] and [7].

CONCLUSION

In presenting a unified approach to optimization algorithms, we have mostly used as examples variations of well known nonlinear programming algorithms. However, this approach is also fruitful in application to optimal control algorithms such as those in [7], to unconstrained optimization algorithms [8] (modified Newton methods, conjugate gradient methods), and to penalty function algorithms such as [12]. Thus, the scope of the approach presented in this paper is quite large, and it is hoped that it will lead to new developments.

The research reported herein was supported by the National Aeronautics and Space Administration under Grant NsG-354, suppl. 4 and 5.

REFERENCES

[1]. P. WOLFE, « On the Convergence of Gradient Methods under Constraints », IBM Research Report RZ-204, March 1, 1966, IBM Research Laboratory, Zurich, Switzerland.

[2]. W. I. ZANGWILL, « Convergence Conditions for Nonlinear Programming Algorithms », Working Paper No 197, Center for Research in Management Science, University of California, Berkeley, California, November 1966.

[3]. D. M. TOPKIS, A. VEINOTT Jr., On the convergence of some feasible direction algorithms for nonlinear programming, *J. SIAM Control*, vol. 5, n° 2, May 1967, p. 268.

[4]. E. POLAK and M. DEPARIS, « An algorithm for minimum energy control », University of California, Electronics Research Laboratory, Berkeley, California, ERL Memorandum M225, November 1, 1967.

[5]. G. ZOUTENDIJK, « Methods of feasible directions: A study in linear and nonlinear programming », Elsevier, Amsterdam, 1960.

[6]. J. B. ROSEN, « The gradient projection method for nonlinear programming. Part I. Linear constraints », *J. SIAM*, vol. 8, n° 1, March 1960, pp. 181-217.

[7]. E. POLAK, « On primal and dual methods for solving discrete optimal control problems », Proc. of the 2nd International Conference on Computing Methods in Optimization Problems, San Remo, Italy, September 9-13, 1968.

[8]. E. POLAK, G. RIBIERE, « Note sur la convergence de méthodes de directions conjuguées ».

[9]. P. KALFON, G. RIBIERE, J. C. SOGNO, « A method of feasible directions using projection operators », Proc. IFIP Congress 68, Edinburgh, August 1968.

[10]. M. CANNON, C. CULLUM and E. POLAK, « Constrained minimization problems in finite dimensional spaces », *J. SIAM Control*, vol. 4, pp. 528-547, 1966.

[11]. H. W. KUHN and A. W. TUCKER, « Nonlinear programming », Proc. of the Second Berkeley Symposium on Mathematic Statistic and Probability, University of California Press, Berkeley, California, 1951, pp. 481-492.

[12]. J. FREHEL, « Une méthode de programmation non linéaire », IBM France, Research Laboratory, Paris, étude n° FF2-0061-0, July 1968.