Copyright © 1969, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

COMPUTATIONAL METHODS IN DISCRETE OPTIMAL CONTROL AND NONLINEAR PROGRAMMING: A UNIFIED APPROACH

by

E. Polak

Memorandum No. ERL-M261

24 February 1969

E L E C T R O N I C S R E S E A R C H L A B O R A T O R Y

College of Engineering
University of California, Berkeley
94720

ACKNOWLEDGMENT

The author wishes to thank all those who have given some of their time in helping him clarify various concepts and ideas. In particular, the author is grateful to Dr. E. J. Messerli, Mr. G. Meyer and Mr. R. Klessig for their assistance in eliminating errors in these notes and in improving the presentation. Finally, thanks are due to Miss Bonnie Bullivant for typing these notes as class notes.

Research sponsored by the National Aeronautics and Space Administration under Grant NGL-05-003-016 (Sup 6).

The emphasis in these notes is on unification. This unification manifests itself in the use of a specific convergence theory, in the use of specific necessary conditions of optimality, and in the unified treatment of algorithms for discrete optimal control and for mathematical programming. For the sake of making the presentation of a large number of algorithms easier within the time and space available, the author has slightly modified a number of standard algorithms so as to fit them better into a unified framework and has supplied new proofs of convergence. In selecting algorithms to be discussed in these notes, the author has given preference to methods which can be used both for optimal control and for mathematical programming problems. Also preference has been given to methods which can be discussed without introducing a great deal of additional background material. As a result, dynamic programming, set approximation and cutting plane methods, and the reduced gradient and convex simplex methods were omitted.

The author's major contribution lies in the development and exploration of the convergence theory presented, in exhibiting the relation between the type of points different algorithms will compute for a given problem and in showing how large families of feasible directions algorithms can be generated from related and equivalent families of necessary conditions of optimality. The list of references at the end of the notes includes only those papers and books which were used by the author to some degree in the preparation of these notes. They are by no means an exhaustive bibliography. A number of results in these notes are presented without reference to other people's work. Some of these results are part of the oral tradition, while others are claimed to be probably new. The reason for hedging and saying "probably new" is that the author has found on several occasions that results not known in the oral tradition in which he participates were common knowledge elsewhere.

NOTE TO READER

ε-Procedures: Throughout these notes the reader will find algorithms including statements of the form: if..., set $\epsilon = \frac{1}{2}\epsilon$. The factor $\frac{1}{2}$ is used quite often in practice, however, the algorithms remain convergent if any scale factor $\beta \in (0,1)$ is used instead, i.e., the reader may replace $\frac{1}{2}\epsilon$ by $\beta\epsilon$ in all such statements. In fact, the β is easily seen to be a design parameter.

Notation and Numbering:

1) The symbols $\|\cdot\|$ and $\langle .,. \rangle$ are used to denote the euclidean norm and the usual scalar product, respectively.

2) Components of vectors are always superscripted: e.g., $z = (z^1, z^2, \ldots, z^n) \in R^n$, elements of a sequence are always subscripted, e.g., $z_0, z_1, z_2, \ldots$.

3) We denote the interior of a set Ω by $\overset{\circ}{\Omega}$ and its boundary by $\partial\Omega$.

4) When referring to an equation within the same section, only the equation number is used, e.g., (7); when referring to an equation in the same chapter but in a different section, the section number and equation number are used, e.g., (1.25); and when referring to an equation in a different chapter three numbers are used, e.g., (I.3.16), the first giving the chapter, the second the section, and the last the equation.

When a section within the same chapter is referred to, the chapter number is omitted.

TABLE OF CONTENTS

I. PRELIMINARY RESULTS

1. Optimal Control and Nonlinear Programming Problems

We shall present in these notes a number of algorithms for solving discrete optimal control problems of the following kind.

1. Problem: Given a dynamical system described by the difference equation

2. $$x_{i+1} - x_i = f_i(x_i, u_i),\ x_i \in \mathbb{R}^\nu,\ u_i \in \mathbb{R}^\mu,$$

$$i = 0, 1, 2, \ldots, k-1,$$

find a control sequence $\mathcal{u} = (u_0, u_1, \ldots, u_{k-1})$ and a corresponding trajectory $\mathcal{x} = (x_0, x_1, \ldots, x_k)$ which minimize the cost functional

3. $$\sum_{i=0}^{k-1} f_i^0(x_i, u_i)$$

subject to the constraints

$$s_i(u_i) \leq 0, \quad i = 0, 1, \ldots, k-1$$

4. $$g_i(x_i) = 0;\ h_i(x_i) \leq 0, \quad i = 0, 1, \ldots, k,$$

where $s_i: \mathbb{R}^\mu \to \mathbb{R}^{\mu_i}$, $g_i: \mathbb{R}^\nu \to \mathbb{R}^{\ell_i}$, $h_i: \mathbb{R}^\nu \to \mathbb{R}^{\nu_i}$. We refrain at this point from loading down the problem statement with assumptions. The required differentiability assumptions are usually clear from the context of an algorithm. Other assumptions, such as linearity or convexity, will be stated when necessary.

3044

We shall now indicate the origin of discrete optimal control problems. Suppose that we have a differential dynamical system of the form

$$5. \qquad \frac{d}{dt}\, x(t) = f(x(t), u(t), t), \quad t \in [0, t_f]$$

where $x(t) \in R^{\nu}$ is the state of the system at time t and $u(t) \in R^{\mu}$ is the input to the system at time t. Also suppose that we are given a performance functional

$$6. \qquad \int_0^{t_f} f^0(x(t), u(t), t)dt$$

which we wish to minimize subject to $x(0) \in X_0 \subset R^{\nu}$, $x(t_f) \in X_f \subset R^{\nu}$ and $u(t) \in U \subset R^{\mu}$. In the form stated, this problem may be computationally intractable, at least as far as available computers or meaningful execution times are concerned. Thus one is forced to impose additional restrictions on the problem to make is solvable. A fairly simple device is to restrict $u(\cdot)$ to the class of piece-wise constant functions with at most k discontinuities. For example,

$$7. \qquad u(t) = u_i \quad \text{for} \quad t \in \left[i\ \frac{t_f}{k},\ (i+1)\ \frac{t_f}{k} \right),$$

$$i = 0, 1, \ldots, k-1, \ u_i \in U .$$

Then, if we let $x_i(t)$, $i = 0, 1, \ldots, k-1$, be the solution of (5) for $t \in [it_f/k, (i+1)t_f/k]$, satisfying $x_i(it_f/k) = x_i$ and corresponding to $u(t) = u_i$ for $t \in [it_f/k, (i+1)t_f/k)$, and set $x_{i+1} = x_i((i+1)t_f/k)$, we find that

$$8. \qquad x_{i+1} = x_i + \int_{it_f/k}^{(i+1)t_f/k} f(x_i(t), u_i, t)dt ,$$

$$i = 0, 1, \ldots, k-1 ,$$

which defines the functions $f_i(x_i, u_i)$ in (2). Similarly, with u(t) restricted

as in (7), (6) becomes

9. $$\sum_{i=0}^{k-1} f_i^0(x_i,u_i)$$

where $f_i^0(x_i,u_i) \equiv \int_{it_f/k}^{(i+1)t_f/k} f^0(x_i(t), u_i, t)dt$. Thus, the additional restriction from the form (5), (6) to the form (1). of the input u(t) by (7), results in a transformation of the optimal control problem/ It should be noted that (7) does not represent the only restriction on u(t) which results in a discrete optimal control problem. Other possibilities exist, with

10. $$u(t) = \sum_{\ell=0}^{\beta} u_i^{\ell} t^{\ell}, \quad t \in \left[it_f/k,\ (i+1)t_f/k \right), \qquad i = 0, 1, \ldots, k-1$$

possibly being the most common class.

The discrete optimal control problem (1) can be viewed as a nonlinear programming problem of the form

11. $$\min \{ f^0(z) \mid f(z) \leq 0,\ r(z) = 0 \},$$

where f^0: $\mathbb{R}^n \to \mathbb{R}^1$, f: $\mathbb{R}^n \to \mathbb{R}^m$, and r: $\mathbb{R}^n \to \mathbb{R}^{\ell}$.

To transcribe the discrete optimal control problem into the form (11), we may proceed in one of two possible ways. The first is to set $z = (x_0, x_1, \ldots, x_k, u_0, u_1, \ldots, u_{k-1})$ and to define

12. $$f^0(z) \equiv \sum_{i=0}^{k-1} f_i^0(x_i,u_i)$$

13.
$$r(z) \triangleq \begin{pmatrix} x_1 - x_0 - f_0(x_0,u_0) \\ \vdots \\ x_k - x_{k-1} - f_{k-1}(x_{k-1},u_{k-1}) \\ g_0(x_0) \\ g_1(x_1) \\ \vdots \\ g_k(x_k) \end{pmatrix}$$

14.
$$f(z) \triangleq \begin{pmatrix} h_0(x_0) \\ h_1(x_1) \\ \vdots \\ h_k(x_k) \\ s_0(u_0) \\ \vdots \\ s_{k-1}(u_{k-1}) \end{pmatrix}$$

For the second transcription we set $z = (x_0, u_0, u_1, \ldots, u_{k-1})$, and define $x_i(x_0,\mathcal{u})$ to be the solution of (2) at time i corresponding to the initial state x_0 and the control sequence $\mathcal{u} = (u_0, u_1, \ldots, u_{k-1})$. Then, we define

15.
$$f^0(z) \triangleq \sum_{i=0}^{k-1} f_i^0(x_i(x_0,\mathcal{u}), u_i)$$

16.
$$r(z) \triangleq \begin{pmatrix} g_0(x_0) \\ g_1(x_1(x_0,u)) \\ \vdots \\ g_k(x_k(x_0,u)) \end{pmatrix}$$

17.
$$f(z) \triangleq \begin{pmatrix} h_0(x_0) \\ h_1(x_1(x_0,u)) \\ \vdots \\ h_k(x_k(x_0,u)) \\ s_0(u_0) \\ \vdots \\ s_{k-1}(u_{k-1}) \end{pmatrix}$$

There are two reasons for transcribing the optimal control problem into the form (11). The first is that the form (11) is considerably simpler than the form (1) and is conceptually simpler to handle. The second reason is that our awareness of the equivalence between the problems (1) and (11) makes it possible for us to utilize a large number of very sophisticated nonlinear programming algorithms in solving (1). Hence, whenever possible, we shall first explain an algorithm in terms of the problem (11) and then particularize it for the form (1). This will avoid the possibility of having simple ideas obscured by the very cumbersome structure of (1).

We shall later see that for some algorithms, (15), (16), (17) give the only usable transcription, whereas to apply other algorithms we may prefer to use (12), (13), (14).

2. Optimality Conditions

The algorithms for solving the problem (1.11) which we shall describe in the chapters to follow, are incapable of distinguishing between a local or a global optimum. In fact, they can only be used to construct a convergent sequence of points whose limit satisfies a particular optimality condition. We therefore pause to state the most frequently used optimality conditions and to identify the cases for which these conditions can be trivially satisfied. Obviously, whenever an optimality condition is trivially satisfied, any algorithm which depends on it becomes useless. The proofs of the theorems stated below can be found in [C1].

1. Theorem: If $\hat{z}$ is optimal for (1.11), i.e., $f^0(\hat{z}) = \min \{f^0(z) \mid f(z) \leq 0, r(z) = 0\}$, then there exist multipliers $\mu^0 \leq 0$, $\mu^1 \leq 0, \ldots, \mu^m \leq 0$ and $\psi^1, \psi^2, \ldots, \psi^\ell$, not all zero, such that

$$2. \qquad \sum_{i=0}^{m} \mu^i \nabla f^i(\hat{z}) + \sum_{i=1}^{\ell} \psi^i \nabla r^i(\hat{z}) = 0$$

and

$$3. \qquad \mu^i f^i(\hat{z}) = 0 \quad \text{for} \quad i = 1, 2, \ldots, m .$$

This theorem is due to F. John [J1].

4. Corollary: If there exists a vector $h \in \mathbb{R}^n$ such that $\langle \nabla f^i(\hat{z}), h \rangle > 0$ for all $i \in \{1, 2, \ldots, m\}$ satisfying $f^i(\hat{z}) = 0$, then $(\mu^0, \psi^1, \psi^2, \ldots, \psi^\ell) \neq C$.

5. Corollary: Suppose that the vectors $\nabla r^i(\hat{z})$, $i = 1, 2, \ldots, \ell$ are linearly independent. If there exists a vector $h \in \mathbb{R}^n$ such that $\langle \nabla f^i(\hat{z}), h \rangle > 0$ for all $i \in \{1, 2, \ldots, m\}$ satisfying $f^i(\hat{z}) = 0$, and $\langle \nabla r^i(\hat{z}), h \rangle = 0$ for $i = 1, 2, \ldots, \ell$, then $\mu^0 < 0$.

The two corollaries are special cases of the Kuhn-Tucker conditions [K4].

6. Corollary: If $\hat{z}$ is optimal for (11) and $r(\cdot) = 0$, then

$$7. \qquad \min_{h \in S} \max_{i \in J_0(\hat{z})} \langle \nabla f^i(\hat{z}), h \rangle = 0$$

where S is any subset of R^n containing the origin in its interior, and for any $\alpha \geq 0$ and any $z \in \{z' \mid f^i(z') \leq 0, i = 1, 2, \ldots, m\}$.

8. $J_\alpha(z) = \{0\} \cup \{i \mid f^i(z) + \alpha \geq 0,\ i \in \{1, 2, \ldots, m\}\}$.

(A form of this condition was probably first published by Zoutendijk [Z3].)

Proof: Suppose (7) is false. Then there exists a $\delta^* > 0$ and a vector $h^* \in S$ such that

9. $$\langle \nabla f^i(\hat{z}), h^* \rangle \leq -\delta^* \quad \text{for all} \quad i \in J_0(\hat{z})$$

Taking the scalar product of h^* with both sides of (2) (in which we must set $\nabla r^i(\hat{z}) = 0$, $i = 1, 2, \ldots, m$, since $r(\cdot) = 0$), we now get a contradiction.

10. Remark: For the case $r(\cdot) = 0$, Theorem (1) can be deduced from Corollary (6), i.e., when $r(\cdot) = 0$, these two conditions are equivalent.

11. Proposition: Suppose that the set $\Omega = \{z \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$ has no interior and that $r(\cdot) = 0$. Then (2), (3), and hence also (6) can be satisfied for every $z \in \Omega$ (i.e., this condition is trivial when Ω has no interior).

Proof: Suppose that for some $i \in \{1, 2, \ldots, m\}$, $\{z \mid f^i(z) \leq 0\}$ consists of only one point, z^*. Then $\Omega = \{z^*\}$ and $\nabla f^i(z^*) = 0$. We may therefore set $\mu^i = -1$, $\mu^j = 0$, $j \neq i$, $j = 1, 2, \ldots, m$ and satisfy (2) and (3). Now, suppose that $\{z \mid f^i(z) \leq 0,\ i \in I\}$, $I \subset \{1, 2, \ldots, m\}$ has an interior, but that $\{z \mid f^i(z) \leq 0,\ i \in I \cup \{j\}\}$ where $j \in \{1, 2, \ldots, m\}$, does not have an interior. Then, any point z^* in Ω is optimal for the problem

11a. $$\min \{f^j(z) \mid f^i(z) \leq 0,\ i \in I\}$$

and satisfies $f^j(z^*) = 0$. Furthermore, by (1), we get for (11a), that there exist multipliers $\mu^j \leq 0$ and $\mu^i \leq 0$, $i \in I$, not all zero, such that

(11b) $$\mu^j \nabla f^j(z^*) + \sum_{i \in I} \mu^i \nabla f^i(z^*) = 0 ,$$

(11c) $$\mu^i f^i(z^*) = 0 , \quad i \in I .$$

Setting all other $\mu^i = 0$, we find that we satisfy (2) and (3) by means of the above multipliers.

12. Theorem: Suppose that $r(\cdot)$ is affine and that the functions $f^i(\cdot)$, $i = 0, 1, \ldots, m$, are convex. If $\hat{z}$ satisfies $r(\hat{z}) = 0$, $f^i(\hat{z}) \leq 0$ for $i = 1, 2, \ldots, m$, and there exist multipliers $\mu^1 \leq 0$, $\mu^2 \leq 0, \ldots, \mu^m \leq 0$ and $\psi^1, \psi^2, \ldots, \psi^\ell$, such that

13. $$-\nabla f^0(\hat{z}) + \sum_{i=1}^{m} \mu^i \nabla f^i(\hat{z}) + \sum_{i=1}^{\ell} \psi^i \nabla r^i(\hat{z}) = 0 ,$$

$$\mu^i f^i(\hat{z}) = 0 \quad \text{for} \quad i = 1, 2, \ldots, m ,$$

then $\hat{z}$ is optimal for (1.11) (see [K4]).

Proof: Let $\Omega' = \{z \mid f^i(z) \leq 0, i = 1, 2, \ldots, m, r^i(z) = 0, i = 1, 2, \ldots, \ell\}$. Then, since the $r^i(\cdot)$ are affine, for any $z \in \Omega'$, $\langle \nabla r^i(\hat{z}), z - \hat{z} \rangle = 0$ and hence, by (13), for any $z \in \Omega'$

13a. $$\langle \nabla f^0(\hat{z}), z - \hat{z} \rangle = \sum_{i=1}^{m} \mu^i \langle \nabla f^i(\hat{z}), z - \hat{z} \rangle$$

Now, since the $f^i(\cdot)$ are convex, we have, for any $z \in \Omega'$ and $i \in \{i \mid f^i(\hat{z}) = 0, i \in \{1, 2, \ldots, m\}\}$,

13b. $$\langle \nabla f^i(\hat{z}), z - \hat{z} \rangle \leq f^i(z) \leq 0 .$$

Making use of (13), (13a), and (13b), we obtain

13c. $$\begin{aligned} f^0(\hat{z}) &\leq f^0(z) - \langle \nabla f^0(\hat{z}), z - \hat{z} \rangle \\ &= f^0(z) - \sum_{i=1}^{m} \mu^i \langle \nabla f^i(\hat{z}), z - \hat{z} \rangle \\ &\leq f^0(z) . \end{aligned}$$

Thus, $\hat{z}$ is optimal.

When applied to the discrete optimal control problem (1.1), the above conditions may assume a highly structured form. We now illustrate this by

considering a special case.

14. Theorem: Suppose that $g_1(\cdot) = g_2(\cdot) = \ldots = g_{k-1}(\cdot) = h_1(\cdot) = h_2(\cdot) = \ldots = h_{k-1}(\cdot) = 0$ for problem (1.1). If the control sequence $\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{k-1}$ and the corresponding trajectory $\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_k$ (i.e., the $\hat{x}_i, \hat{u}_i$ satisfy (1.2)) are optimal for the problem (1.1), then there exist a scalar multiplier $p^0 \leq 0$ and vector multipliers $p_0, p_1, p_2, \ldots, p_k, \pi_0, \pi_k, \mu_0 \leq 0, \ldots, \mu_{k-1} \leqq 0, \xi_0 \leqq 0, \xi_k \leqq 0$, not all zero, such that

15. $$p_i - p_{i+1} = \left(\frac{\partial f_i(\hat{x}_i, \hat{u}_i)}{\partial x_i}\right)^T p_{i+1} - p^0 \left(\frac{\partial f_i^0(\hat{x}_i, \hat{u}_i)}{\partial x_i}\right)^T ,$$

$$i = 0, 1, \ldots, k-1$$

16. $$p_0 = \left(\frac{\partial g_0(\hat{x}_0)}{\partial x_0}\right)^T \pi_0 + \left(\frac{\partial h_0(\hat{x}_0)}{\partial x_0}\right)^T \xi_0$$

17. $$p_k = \left(\frac{\partial g_k(\hat{x}_k)}{\partial x_k}\right)^T \pi_k + \left(\frac{\partial h_k(\hat{x}_k)}{\partial x_k}\right)^T \xi_k$$

18. $$p^0 \left(\frac{\partial f_i^0(\hat{x}_i, \hat{u}_i)}{\partial u_i}\right)^T + \left(\frac{\partial f_i(\hat{x}_i, \hat{u}_i)}{\partial u_i}\right)^T p_{i+1} + \left(\frac{\partial s_i(\hat{u}_i)}{\partial u_i}\right)^T \mu_i = 0$$

19. $$\langle \xi_0, h_0(x_0) \rangle = \langle \xi_k, h_k(x_k) \rangle = 0$$

20. $$\langle s_i(\hat{u}_i), \mu_i \rangle = 0 \quad \text{for} \quad i = 0, 1, 2, \ldots, k-1 .$$

To obtain Theorem (14) from Theorem (1), we proceed as follows. First, we note that (2) is equivalent to the statement that $\nabla L(\hat{z}) = 0$, where $L(z) = \mu^0 f^0(z) + \langle \mu, f(z) \rangle + \langle \psi, r(z) \rangle$. Then we transcribe the problem (1.1) into the form (1.11) using the formulas (1.12), (1.13) and (1.14), and set $\mu^0 = p^0$, $\mu = (\xi_0, \xi_k, \mu_0, \mu_1, \ldots, \mu_{k-1})$, $\psi = (p_1, p_2, \ldots, p_k, \pi_0, \pi_k)$. Hence,

$$L(z) = p^0 \sum_{i=0}^{k-1} f_i^0(x_i,u_i) +$$

$$+ \sum_{i=0}^{k-1} \langle p_{i+1},\ x_{i+1} - x_i - f_i(x_i,u_i)\rangle$$

$$+ \langle \pi_0,\ g_0(x_0)\rangle + \langle \pi_k,\ g_k(x_k)\rangle$$

$$+ \langle \xi_0,\ h_0(x_0)\rangle + \langle \xi_k,\ h_k(x_k)\rangle$$

$$+ \sum_{i=0}^{k-1} \langle \mu_i,\ s_i(u_i)\rangle .$$

Now computing $\frac{\partial L(z)}{\partial x_i}$, $i = 0, 1, \ldots, k$ and $\frac{\partial L(z)}{\partial u_i}$, $i = 0, 1, \ldots, k-1$, and setting these equal to zero we obtain (15), (16), (17) and (18).

21. Remark: One may sometimes wish to eliminate the μ_i in (18) and (20). This can be done by substituting for (18) the condition

23. $$\left\langle p^0 \left(\frac{\partial f_i^0(\hat{x}_i,\hat{u}_i)}{\partial x}\right)^T + \left(\frac{\partial f_i(\hat{x}_i,\hat{u}_i)}{\partial x}\right)^T p_{i+1},\ \delta u \right\rangle \leq 0$$

for all δu such that

24. $$\left(\frac{\partial s_i(\hat{u}_i)}{\partial u}\right) \delta u \leqq 0 \quad .$$

3. Convergence Conditions

The convergence theorems to be presented in this text can be thought of as being extensions of Lyapunov Second Method for dynamical systems described by difference equations. Prototypes of convergence results stated in the particular form used in the text can be found in Polyak [P5] and in Zangwill

[Z3], [Z1]; related ideas appear in the works of Varaiya [V3], Levitin and Polyak [L2], Topkis and Veinott [T1], Hurt [H4], and in less explicit form in Arrow, Hurwiz, Uzawa [A1], Zantendijk [Z4], and Zukhovitokii, Polyak and Princk [Z5]. (The author was unawar of Polyak's work, which is closest to the authors' at the time [P1], [P2] and [P3] were written.)

To establish that an algorithm converges and to explore the extent to which it can be perturbed without affecting its convergence, we shall mostly use the following results, which were first stated by Polak in [P3], [P2], [P1].

Suppose that we have a closed set $T \subset \mathbb{R}^n$ which contains desirable points. These points may be desirable because they are optimal for some optimization problem, or

because some optimality condition is satisfied at these points for some optimization problem, or because some function vanishes at these points, etc. Now, suppose that we propose to find desirable points in T by means of a search function $a: T \to T$ and of a stop function $c: T \to \mathbb{R}^1$.

1. Theorem: Suppose that,

(i) $\hat{z} \in T$ is desirable if and only if

2. $$c(a(\hat{z})) \leq c(\hat{z}) \ ;$$

(ii) $c(\cdot)$ is either continuous at all nondesirable $z \in T$ or else $c(z)$ is bounded from above for $z \in T$;

(iii) For every $z \in T$ which is not desirable, there exist an $\varepsilon(z) > 0$ and a $\delta(z) > 0$ such that

3. $$c(a(z')) - c(z') \geq \delta(z)$$

for all $z' \in T$ such that $\|z - z'\| \leq \varepsilon(z)$.

If the sequence $z_i \in T$, $i = 0, 1, 2, \ldots$, is constructed according to

4. $$z_{i+1} = a(z_i), \quad i = 0, 1, 2, \ldots,$$

and construction stops only if for some integer j, $c(z_{j+1}) \leq c(z_j)$, (i.e., $c(z_{i+1}) > c(z_i)$ for $i = 0, 1, 2, \ldots$,) then either $\{z_i\}$ is finite and its last element is desirable, or else it is infinite and every accumulation point of $\{z_i\}$ is desirable.

Proof: The case of $\{z_i\}$ finite is obvious. Hence, suppose that $\{z_i\}$ is infinite and that $z_i \to z^*$ for $i \in K \subset \{0, 1, 2, \ldots\}$, with z^* not desirable. Then there exist $\varepsilon^* > 0$ and $\delta^* > 0$ and a $k \in K$ such that for all $i \geq k$, $i \in K$

5. $$\|z_i - z^*\| \leq \varepsilon^*$$

and

6. $$c(z_{i+1}) - c(z_i) \geq \delta^* \ .$$

Thus, for any two consecutive points z_i, z_{i+j}, i, $(i+j) \in K$, $i \geq k$, of the subsequence we must have

$$7. \qquad c(z_{i+j}) - c(z_i) = (c(z_{i+j}) - c(z_{i+j-1})) + (c(z_{i+j-1}) - c(z_{i+j-2}) \ldots + (c(z_{i+1}) - c(z_i)) > \delta^*$$

Now, for $i \in K$, the monotonically increasing sequence $c(z_i)$, $i = 0, 1, 2, \ldots$, must converge either because $c(\cdot)$ is continuous at z^* or because $c(z)$ is bounded from above on T. But this is contradicted by (7) and hence we are done.

In the preceeding development we had assumed that the relation between successive points, z_i, z_{i+1}, $i = 0, 1, 2, \ldots$, constructed in the search for a desirable point in T, was functional, i.e., that $z_{i+1} = a(z_i)$. In practice, however, it is usually impossible to compute $a(z_i)$ with arbitrary accuracy in finite time, and one therefore accepts a point z_{i+1} lying in an approximation set, $A_\delta(z_i)$, to $a(z_i)$. The parameter $\delta (\delta \geq 0)$ is used to express the precision with which z_{i+1} approximates $a(z_i)$; it depends on z_i and is usually driven to zero as z_i converges to a desirable point. Under these circumstances we have the following convergence condition.

8. Theorem: Suppose we are given a search function a: $T \to T$ and a stop rule c: $T \to R^1$ which satisfy the conditions of Theorem (1) and, in addition, suppose that $c(\cdot)$ is continuous.

For any $z \in T$, let

$$9. \qquad A_\varepsilon(z) = \{z' \in T \mid \|z' - a(z)\| \leq \varepsilon\}$$

Consider the following algorithm:

for every $i \in K$, $i \geq k'$ and every $y \in A_{\varepsilon',[z_i]}$

13. $$c(y) - c(z_i) \geq \varepsilon'$$

and therefore for $i \in K$, $i \geq k'$

14. $$c(z_{i+1}) - c(z_i) \geq \varepsilon'$$

Consequently, if i, i+j are two consecutive indices in K, with $i \geq k'$,

15. $$c(z_{i+j}) - c(z_i) = [c(z_{i+j}) - c(z_{i+j-1})] + \ldots + [c(z_{i+1}) - c(z_i)] > \varepsilon'$$

and hence, $c(z_i)$, $i \in K$ cannot converge to $c(z^*)$, which contradicts the continuity of $c(\cdot)$. Therefore z^* must be a desirable point and the theorem is proved.

The approximations to a(z) defined by (9) are not the only ones which can be used in a convergent algorithm for computing desirable points in T. For example, we can use approximations of the form $A(z) = \{z' \in T \mid c(z') - c(z) \geq \beta(c(a(z)) - c(z)\}$, where $\beta > 0$ is fixed. The following theorem applies to this case as well as to a number of other schemes which we shall later encounter.

16. Theorem: Let $A(\cdot)$ be a map from T into the set of all subsets of T, and let $c: T \to \mathbb{R}^1$ be a stopping rule which is either continuous at all nondesirable points in T or else c(z) is bounded from above for $z \in T$. Suppose that $z \in T$ is desirable if and only if

17. $$c(z') - c(z) \leq 0 \quad \text{for at least one} \quad z' \in A(z) \quad ,$$

and that for every nondesirable $z \in T$ there exist an $\varepsilon(z) > 0$ and a $\delta(z) > 0$ such that

18. $$c(z'') - c(z') \geq \delta(z)$$

for all $z' \in T$ such that $\|z' - z\| \leq \varepsilon(z)$ and for all $z'' \in A(z')$.

If the sequence $z_i \in T$, $i = 0, 1, 2, \ldots$, is constructed to satisfy

19.
$$z_{i+1} \in A(z_i) \qquad i = 0, 1, 2, \ldots$$
$$c(z_{i+1}) > c(z_i) \qquad i = 0, 1, 2, \ldots$$

and construction stops only if for some integer j, $c(z_{j+1}) \leq c(z_j)$, then, either $\{z_i\}$ is finite and its last element is desirable or else it is infinite and every accumulation point of $\{z_i\}$ is desirable.

The proof of this theorem follows the same steps as the proof of Theorem (1) and hence is omitted.

20. Remark: The reader must be careful not to read more into the statements of the convergence theorems presented in this section than they say. Note that these Theorems state only that if a convergent subsequence exists, then its limit point will be desirable. To ensure that convergent subsequences will exist, one must make some additional assumptions. For example, one may assume that T is compact, or else that for every real α the set $\{z \in T \mid c(z) \leq c(\alpha_0)\}$ is compact, either being sufficient to ensure the existence of convergent subsequences.

To conclude, we shall state a theorem which combines and generalizes the results contained in Theorems (8) and (16). The theorem given below will be particularly useful in proving the convergence of algorithms which use finite difference approximations to derivatives.

21. Theorem: Let $A(\cdot,\cdot)$ be a map from $\mathbb{R}^+ \times T$ into the set of all subsets of T. Let $c: T \to \mathbb{R}$ be a (stopping rule) function which is either continuous at all nondesirable $z \in T$ or else $c(z)$ is bounded from above on T.

Suppose that (i) $z \in T$ is desirable if and only if

22.
$$c(z') - c(z) \geq 0 \qquad \text{for all } z' \in A(z, \cdot) \subset A(\cdot, z)$$

(ii) for every non-desirable $z \in T$ there exist an $\varepsilon(z) > 0$, a $\delta(z) > 0$ and a

$\gamma(z) > 0$ such that

23. $$c(z'') - c(z') \geq \delta(z)$$

for all $z' \in T$, $\|z' - z\| \leq \varepsilon(z)$, and all $z'' \in A(\gamma, z')$, $0 \leq \gamma \leq \gamma(z)$. Now consider the following algorithm:

24. Algorithm: Suppose that an $\varepsilon_0 > 0$ and a $z_0 \in T$ are given.

Step 0: Set $z = z_0$

Step 1: Set $\varepsilon = \varepsilon_0$

Step 2: Compute a $y \in A(\varepsilon, z)$

Step 3: If $c(y) - c(z) \geq \varepsilon$ set $z = y$ and go to Step 1

If $c(y) - c(z) < \varepsilon$ set $\varepsilon = \varepsilon/2$ and go to Step 2.

Let $z_1, z_2, \ldots$ be the successive values assigned to z in Step 3. Then either the sequence $\{z_i\}$ is finite and its last element is desirable, or else it is infinite and every accumulation point of $\{z_i\}$ is desirable.

The proof of this theorem is a minor modification of the proof of Theorem (8) and is therefore omitted. The preceeding three convergence theorems are easily seen to be convenient special cases of Theorem (21), which correspond to $A(\cdot,\cdot)$ being of a special form, as in (8), to $A(\cdot,\cdot)$ being independent of ε as in (16), and to $A(\cdot,\cdot)$ being independent of ε and $A(\varepsilon,z)$ consisting of a unique point, as in (1). We now turn our attention to the object of our primary interest: optimization algorithms.

4. A Few Useful Properties of Continuous Functions

Throughout this text we shall make repeated use of a few properties of continuous functions. We shall summarize these properties in this section for further reference.

1. Proposition: Suppose that $f^i(\cdot)$ is a continuous function from $\mathbb{R}^n$ into $\mathbb{R}^1$ and S is a compact subset of $\mathbb{R}^n$. Then for each $z^* \in \mathbb{R}^n$, and for each $\delta > 0$ there exists an $\epsilon^* > 0$ and a $\lambda_m > 0$ such that for all $z \in B(z^*,\epsilon^*) \triangleq \{z \mid \|z-z^*\| \leq \epsilon^*\}$ and for all $h \in S$,

2. $$|f^i(z + \lambda h) - f^i(z)| \leq \delta \quad \text{for all} \quad \lambda \in [0,\lambda_m].$$

Proof: Let $\epsilon > 0$ be arbitrary, but finite and let $z^* \in \mathbb{R}^n$. Then $f^i(\cdot)$ is uniformly continuous on the compact ball $B(z^*, \epsilon) \triangleq \{z \mid \|z-z^*\| \leq \epsilon\}$ and hence there exists an $\epsilon' > 0$ such that

3. $$|f^i(z) - f^i(z')| \leq \delta$$

for all $z, z' \in B(z^*, \epsilon)$ satisfying $\|z-z'\| \leq \epsilon'$. Let $\epsilon^* = \min \{\epsilon', \epsilon/2\}$, and let $M = \max \{\|h\| \mid h \in S\}$. If we set $\lambda_m = \epsilon^*/M$, then for all $h \in S$ and for all $z \in B(z^*, \epsilon^*)$,

4. $$\|\lambda h\| \leq \epsilon^* \quad \text{for all} \quad \lambda \in [0,\lambda_m]$$

5. $$(z + \lambda h) \in B(z, \epsilon^*) \subset B(z^*, \epsilon) \quad \text{for all} \quad \lambda \in [0,\lambda_m]$$

and therefore, because of (3), for all $z \in B(z^*, \epsilon^*)$ and for all $h \in S$,

6. $$|f^i(z + \lambda h) - f^i(z)| \leq \delta \quad \text{for all } \lambda \in [0,\lambda_m] .$$

7. Proposition: Suppose that $f^i(\cdot)$ is a continuously differentiable function from $\mathbb{R}^n$ into $\mathbb{R}^1$ and S is a compact subset of $\mathbb{R}^n$. Then for each $z^* \in \mathbb{R}^n$ and for each $\delta > 0$ there exists an $\epsilon^* > 0$ and a $\lambda_m > 0$ such that for all

$z \in B(z^*, \epsilon^*) \triangleq \{z \mid \|z-z^*\| \leq \epsilon^*\}$ and for all $h \in S$

8. $$|\langle \nabla f^i(z + \zeta h), h\rangle - \langle \nabla f^i(z), h\rangle| \leq \delta \quad \text{for all } \zeta \in [0, \lambda_m] .$$

Proof: Let $\epsilon > 0$ be arbitrary, but finite and let $z^* \in \mathbb{R}^n$. Then the function $\langle \nabla f^i(\cdot), \cdot\rangle$, from $\mathbb{R}^n \times \mathbb{R}^n$ into $\mathbb{R}^1$, is uniformly continuous on $B(z^*, \epsilon) \times S$ and hence there exists an $\epsilon' > 0$ such that

9. $$|\langle \nabla f^i(z'), h'\rangle - \langle \nabla f^i(z), h\rangle| \leq \delta$$

for all (z,h), $(z',h') \in B(z^*, \epsilon) \times S$ satisfying $\|z'-z\| \leq \epsilon'$, $\|h'-h\| \leq \epsilon'$. Now, let $\epsilon^* = \{\epsilon', \epsilon/2\}$ and let $M = \max\{\|h\| \mid h \in S\}$. If we set $\lambda_m = \epsilon^*/M$, then for all $h \in S$ and for all $z \in B(z^*, \epsilon^*)$,

10. $$\|\zeta h\| \leq \epsilon^* \quad \text{for all } \zeta \in [0, \lambda_m]$$

11. $$(z + \zeta h) \in B(z, \epsilon^*) \subset B(z^*, \epsilon) \quad \text{for all } \zeta \in [0, \lambda_m]$$

and hence because of (9), for all $z \in B(z^*, \epsilon^*)$ and for all $h \in S$,

12. $$|\langle \nabla f^i(z + \zeta h), h\rangle - \langle \nabla f^i(z), h\rangle| \leq \delta \quad \text{for all } \zeta \in [0, \lambda_m] .$$

13. Proposition: Suppose that for $i = 0, 1, \ldots, m$ the functions $f^i: \mathbb{R}^n \to \mathbb{R}^1$ are continuous. Then the function $M: \mathbb{R}^n \to \mathbb{R}^1$ defined by

14. $$M(z) = \max\{f^i(z) \mid i \in \{0, 1, \ldots, m\}\}$$

is also continuous.

Proof: Let $z^* \in \mathbb{R}^n$ and $\delta > 0$ be arbitrary, then there exists an $\epsilon > 0$ (possibly depending on z^*) such that for all $z \in \overset{o}{B}(z^*, \epsilon) \triangleq \{z \mid \|z-z^*\| < \epsilon\}$

15. $$f^i(z^*) - \delta < f^i(z) < f^i(z^*) + \delta, \quad i = 0, 1, \ldots, m.$$

Hence,

16a. $$f^i(z) < M(z^*) + \delta, \quad i = 0, 1, \ldots, m,$$

and

16b. $$f^i(z^*) - \delta < M(z), \quad i = 0, 1, \ldots, m.$$

Taking the maximum over i in (16a,b), we get

17. $$M(z^*) - \delta < M(z) < M(z^*) + \delta$$

and hence $M(\cdot)$ is continuous.

18. Proposition: Let $\psi(\cdot,\cdot)$ be a continuous function from $\mathbb{R}^n \times \mathbb{R}^{n\dagger}$ into $\mathbb{R}^1$ and let S be a compact subset of $\mathbb{R}^n$. Then the function $m: \mathbb{R}^n \to \mathbb{R}^1$, defined by

19. $$m(z) = \min \{\psi(z,h) \,|\, h \in S\},$$

is also continuous.

Proof: Let $z^* \in \mathbb{R}^n$ be arbitrary. We shall show that $m(\cdot)$ is continuous at z^*. Let $\epsilon > 0$ be arbitrary, and let $B(z^*, \epsilon) = \{z \,|\, \|z-z^*\| \leq \epsilon\}$. Then $\psi(\cdot,\cdot)$ is uniformly continuous on $B(z^*, \epsilon) \times S$, and, given any $\delta > 0$, there exists an $\epsilon' > 0$ such that

20. $$|\psi(z,h) - \psi(z',h')| < \delta$$

for all (z,h), $(z',h') \in B(z^*,\epsilon) \times S$ satisfying $\|z-z'\| < \epsilon'$, $\|h-h'\| < \epsilon'$. Let $\epsilon^* = \min \{\epsilon', \epsilon\}$, then

21. $$\psi(z,h) - \delta < \psi(z^*,h) < \psi(z,h) + \delta$$

for all $h \in S$ and all $z \in \overset{\circ}{B}(z^*, \epsilon^*)$, the interior of $B(z^*,\epsilon^*)$. From (21) we now get, by minimizing the appropriate terms over $h \in S$,

22a. $$m(z) - \delta < \psi(z^*,h) \quad \text{for all } h \in S$$

22b. $$m(z^*) < \psi(z,h) + \delta \quad \text{for all } h \in S .$$

---

†This proposition is obviously valid for ψ: $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, continuous, $S \subset \mathbb{R}^{n_2}$ compact, and $m(x) = \min \{\psi(x,y) \,|\, y \in S\}$. We shall only need the form (18).

Again minimizing over $h \in S$ we now obtain from (22),

23. $$|m(z) - m(z^*)| < \delta$$

which shows that $m(\cdot)$ is continuous.

II. UNCONSTRAINED MINIMIZATION AND BOUNDARY VALUE PROBLEMS

1. Minimization: General Theory

We shall now consider the problem:

1. $$\min \{f^0(z) \mid z \in \mathbb{R}^n\} ,$$

where f^0: $\mathbb{R}^n \to \mathbb{R}^1$ is a continuously differentiable function with the property that the set

2. $$\{z \mid f^0(z) \leq \alpha\}$$

is bounded for every $\alpha \in \mathbb{R}^1$.†

A large number of algorithms for solving (1) fall into the following category. For every $z \in \mathbb{R}^n$, let $D(z)$ be an $n \times n$, positive definite (> 0) matrix whose elements are continuous functions of z. Let

3. $$h(z) \triangleq -D(z)\nabla f^0(z)$$

and let

4. $$a(z) \triangleq z + \mu(z)\, h(z) ,$$

where $\mu(z) \geq 0$ is the smallest positive scalar such that

5. $$f^0(z + \mu(z)\, h(z)) \leq f^0(z + \mu h(z)), \quad \mu \geq 0 .$$ ††

Note that for every $z \in \mathbb{R}^n$ such that $\nabla f^0(z) = 0$, $a(z) = z$. Hence, the search function $a(\cdot)$ defined in (4) and (5) can only serve for finding stationary points which are not local maxima. When the function $f^0(\cdot)$ is convex these stationary

---

† This property ensures that every sequence $\{x_i\}$ in $\mathbb{R}^n$, satisfying $f^0(x_i) < f^0(x_0)$, is compact and hence has accumulation points.

†† Algorithms of this type are variations of the method of steepest descent probably first used by Cauchy [C2] and have been in use for a very long time. For an alternative discussion see Topkis and Veinott [T1] or Zangwill [Z1].

points will be the actual minima of $f^0(\cdot)$.

6. Theorem: Let $z_0 \in \mathbb{R}^n$ be arbitrary. Suppose that the sequence z_i, $i = 0, 1, 2, \ldots$, was constructed according to

7. $$z_{i+1} = a(z_i), \qquad i = 0, 1, 2, \ldots,$$

8. $$f^0(z_{i+1}) < f^0(z_i), \quad i = 0, 1, 2, \ldots,$$

where $a(\cdot)$ is defined by (4) and (5), and construction stops only if $f^0(z_{i+1}) \geq f^0(z_i)$. Then either the sequence $\{z_i\}$ is finite, terminating at z_k with $\nabla f^0(z_k) = 0$, or else it is infinite and every accumulation point z^* of $\{z_i\}$ satisfies $\nabla f^0(z^*) = 0$.

Proof: We shall show that the assumptions of Theorem (I.3.1) are satisfied for $c \triangleq -f^0$, and $z \in \mathbb{R}^n$ defined to be desirable if and only if $\nabla f^0(z) = 0$. Obviously, $a(z) = z$ for every $z \in \mathbb{R}^n$ such that $\nabla f^0(z) = 0$. Now, if $z_k \in \mathbb{R}^n$ is such that $\nabla f^0(z_k) \neq 0$, then $\langle \nabla f^0(z_k), D(z_k) \nabla f^0(z_k) \rangle = \delta_k > 0$, and since $\nabla f^0(\cdot)$ is continuous, there exists an $\epsilon_k > 0$ such that for all $0 \leq \mu \leq \epsilon_k$,

9. $$\begin{aligned} f^0(z_k + \mu h(z_k)) &= f^0(z_k) + \mu \langle \nabla f^0(z_k + \xi h(z_k)), h(z_k) \rangle \\ &\leq f^0(z_k) - \mu\delta_k/2 , \end{aligned}$$

where $\xi \in [0, \mu]$. Consequently, $\mu(z_k) \geq \epsilon_k$ and $f^0(a(z_k)) < f^0(z_k)$, i.e., the computation stops if and only if $\nabla f^0(z_k) = 0$, so that condition (i) of Theorem (I.3.1) is satisfied.

Now, let $z^* \in \mathbb{R}^n$ be such that $\nabla f^0(z^*) \neq 0$. Then $\mu(z^*) > 0$ and we define $\theta: \mathbb{R}^n \to \mathbb{R}^1$ by

10. $$\theta(z) = f^0(z) - f^0(z + \mu(z^*)\, h(z)) .$$

By inspection, $\theta(\cdot)$ is a continuous function and

11. $$\theta(z^*) = f^0(z^*) - f^0(z^* + \mu(z^*)\, h(z^*)) = \theta^* > 0 .$$

Since $\theta(\cdot)$ is continuous, there exists an $\epsilon^* \geq 0$ such that

12. $$|\theta(z) - \theta(z^*)| \leq \theta^*/2 ,$$

i.e., such that

13. $$\theta(z) \geq \theta^*/2$$

for all $z \in \{z \mid \|z - z^*\| \leq \epsilon^*\}$.

But

14. $$f^0(z) - f^0(z + \mu(z)\, h(z)) \geq \theta(z)$$

and hence, setting $\delta(z^*) = \theta^*/2$, $\epsilon(z^*) = \epsilon^*$, we find that assumption (iii) of Theorem (I.3.1) is satisfied and we are done. (Assumption (ii) is satisfied since $f^0(\cdot)$ is continuous.)

The search function $a(\cdot)$ defined by (4) and (5) cannot be calculated in a finite number of operations because of the one-dimensional minimization indicated in (5). We now give a modification of the algorithm (7), (8) which sheds light on the extent to which the minimization in (5) can be relaxed without affecting the convergence properties of the resulting algorithm.

15. Corollary: Suppose that we construct a sequence $\{z_i\}$ in $\mathbb{R}^n$ such that

16. $$z_{i+1} = z_i + \lambda_i h(z_i), \quad \lambda_i \geq 0, \quad i = 0, 1, 2, \ldots,$$

and, for a fixed $\lambda^* > 0$,

17. $$f^0(z_i) - f^0(z_{i+1}) \geq \lambda^*(f^0(z_i) - f^0(a(z_i))) ,$$

where $h(\cdot)$ and $a(\cdot)$ are defined as in (3), and (4) and (5), respectively. Then either $\{z_i\}$ is finite and its last element, z_k, satisfies $\nabla f^0(z_k) = 0$, or $\{z_i\}$ is infinite and every accumulation point z^* of $\{z_i\}$ satisfies $\nabla f^0(z^*) = 0$.

We omit a proof of this corollary since it follows directly from Theorem (I.3.16). The computational importance of this corollary lies in the fact that if $\lambda^* > 0$ is chosen to be extremely small, then, in all probability, just about any coarse minimization of $f^0(z_i + \lambda h(z_i))$, $\lambda \geq 0$, will result in a λ_i satisfying (17) and hence in a convergent scheme. (However, the rate of convergence may be affected adversely.)

We now give an algorithm for solving (1) which does not require us to perform a one-dimensional minimization and yet can be proved to converge in the sense of Theorem (I.3.16).

18. Theorem: Suppose that we are given an $\alpha \in (0, \frac{1}{2})$ and that the sequence $\{z_i\}$ in $\mathbb{R}^n$ is constructed to satisfy, for $i = 0, 1, 2, \ldots$,

19. $$z_{i+1} = z_i + \lambda_i h(z_i) \quad ,$$

20. $$\lambda_i(1-\alpha) \langle \nabla f^0(z_i), h(z_i) \rangle \leq f^0(z_i + \lambda_i h(z_i)) - f^0(z_i)$$
$$\leq \lambda_i \alpha \langle \nabla f^0(z_i), h(z_i) \rangle \dagger$$

and

21. $$f^0(z_{i+1}) < f^0(z_i) \quad ,$$

(see Fig. 1), where $h(\cdot)$ is defined as in (3). Then, either $\{z_i\}$ is finite and its last element, z_k, satisfies $\nabla f^0(z_k) = 0$, or $\{z_i\}$ is infinite and every accumulation point z^* of $\{z_i\}$ satisfies $\nabla f^0(z^*) = 0$.

Proof: Note that if $\nabla f^0(z_i) = 0$, then $z_{i+1} = z_i$ and that if $\nabla f^0(z_i) \neq 0$, then $z_{i+1} \neq z_i$ and $f^0(z_{i+1}) < f^0(z_i)$, by simple geometric reasoning. Thus, if the construction of the sequence stops at z_k, (because $f^0(z_{k+1}) = f^0(z_k)$), then $\nabla f^0(z_k) = 0$. For the case where $\{z_i\}$ is infinite, we shall show that the assumptions of Theorem (I.3.16) are satisfied for $c \triangleq -f^0$ and, for $A(\cdot)$ defined by

---

†A step-size rule of this type was probably first used by Goldstein and Price [G2].

22. $$A(z) \triangleq \{z' = z + \lambda h(z) \mid \lambda \geq 0 \quad \text{and}$$

$$\lambda(1-\alpha)\ \langle \nabla f^0(z),\ h(z)\rangle \leq f^0(z') - f^0(z) \leq \lambda\ \alpha\ \langle \nabla f^0(z),\ h(z)\rangle\ .$$

For every $\lambda \in \mathbb{R}^1$ and every $z \in \mathbb{R}^n$, let $\psi(\lambda,z)$ be defined by

23. $$\psi(\lambda;z) = \frac{1}{\lambda}\ (f^0(z + \lambda h(z)) - f^0(z)) - (1-\alpha)\ \langle \nabla f^0(z),\ h(z)\rangle\}\ .$$

Let $\rho(z) > 0$ be the smallest positive root of the equation $\psi(\lambda;z) = 0$. Then the procedure defined by (19), (20) and (21) will always set $\lambda_i \geq \rho(z_i)$. Now note that for every $z \in \mathbb{R}^n$ such that $\nabla f^0(z) \neq 0$, $\psi(\lambda;z) < 0$ for all $\lambda \in [0,\ \rho(z)]$; that $\psi(0,z) = +\ \alpha\ \langle \nabla f^0(z),\ h(z)\rangle$; and that $\psi(\cdot;\cdot)$ is jointly continuous in λ and z.

Let $z_i \in \mathbb{R}^n$ be such that $\nabla f^0(z_i) \neq 0$ and let

24. $$\beta_i = \max\ \{\psi(\lambda;z_i) \mid \lambda \in [0,\ \tfrac{1}{2}\ \rho(z_i)]\}\ .$$

Then $\beta_i < 0$, and there exists an $\epsilon' > 0$ such that for all $z \in \{z \mid \|z - z_i\| \leq \epsilon'\}$ and $\lambda \in [0,\ \frac{1}{2}\ \rho(z_i)]$,

25. $$\psi(\lambda;z) \leq \beta_i/2\ .$$

Since $\langle \nabla f^0(\cdot),\ h(\cdot)\rangle$ is continuous, there exists an $\epsilon'' > 0$ such that

26. $$\langle \nabla f^0(z),\ h(z)\rangle \leq \frac{1}{2}\ \langle \nabla f^0(z_i),\ h(z_i)\rangle \triangleq \gamma_i/2 < 0\ ,$$

for all $z \in \{z \mid \|z - z_i\| \leq \epsilon''\}$. Let $\epsilon = \min\ \{\epsilon',\ \epsilon''\}$, then for all $z \in \{z \mid \|z - z_i\| \leq \epsilon\}$, the algorithm chooses a $\lambda_z \geq \rho(z_i)/2$ such that

27. $$f^0(z + \lambda_z\ h(z)) - f^0(z) \leq \alpha\ \lambda_z\ \langle \nabla f^0(z),\ h(z)\rangle \leq \alpha\ \frac{\rho(z_i)}{2}\ \langle \nabla f^0(z),\ h(z)\rangle$$

$$\leq \alpha\ \frac{\rho(z_i)}{2}\ \frac{\gamma_i}{2} < 0\ .$$

Setting $\epsilon(z_i) = \epsilon$ and $\delta(z_i) = -\alpha \dfrac{\rho(z_i)\,\gamma_i}{4}$, we find that the assumptions of Theorem (I.3.16) are satisfied.

28. Remark: The following is a simple method for finding a $\lambda_i > 0$ which satisfies (20) for a given z_i.† Select a suitable step length, $\rho > 0$. Let

29. $$\underline{\theta}(\lambda;z) = f^0(z + \lambda h(z)) - f^0(z) - \lambda\,\alpha\,\langle \nabla f^0(z),\, h(z)\rangle \quad .$$

For $j = 0, 1, 2, \ldots$, compute $\underline{\theta}(j\rho;\, z_i)$, stopping the calculation for the first $j > 0$ such that $\underline{\theta}(j\rho;\, z_i) \geq 0$. If $\lambda_i = j\rho$ satisfies (20) we are done. If not there is a λ_i in $[(j-1)\rho,\, j\rho]$ which satisfies (20) and it can be found by consecutive halvings of subintervals of $[(j-1)\rho,\, j\rho]$ in an obvious manner. A faster method would be to divide subintervals of $[(j-1)\rho,\, j\rho]$ into thirds.

30. Remark: The proof of Theorem (18) indicates that a sequence $\{z_i\}$ in $\mathbb{R}^n$, constructed to satisfy

31. $$z_{i+1} = z_i + \lambda_i\, h(z_i), \quad i = 0, 1, 2, \ldots$$

with λ_i computed as shown below, has the same convergence properties as the sequence $\{z_i\}$ constructed in (18).

32. Algorithm for λ_i: Choose a step size $\rho > 0$ and an $\alpha \in (0, 1)$. Let

33. $$\bar{\theta}(\lambda;z) = f^0(z + \lambda h(z)) - f^0(z) - \lambda(1-\alpha)\,\langle \nabla f^0(z),\, h(z)\rangle \quad ,$$

where $h(z)$ is as defined in (3).

Step 1: Set $\lambda = \rho$.

Step 2: Compute $\bar{\theta}(\lambda;z_i)$.

Step 3: If $\bar{\theta}(\lambda;z_i) \leq 0$, set $\lambda_i = \lambda$.

If $\bar{\theta}(\lambda;z_i) > 0$, set $\lambda = \dfrac{\lambda}{2}$ and go to Step 2.*

---

†It is an obvious adaptation of the regula falsi method for finding a root.

*This procedure was suggested to the author by M. G. Meyer.

Note that (31) and (32) now define an (ordinary) function a: $\mathbb{R}^n \to \mathbb{R}^n$.

To conclude this section, we indicate a somewhat more general class of algorithms which is used less frequently.

Let $h(\lambda;z)$ be a continuous function from $\mathbb{R}^1 \times \mathbb{R}^n$ into $\mathbb{R}^n$ such that for every $z \in \mathbb{R}^n$ satisfying $\nabla f^0(z) \neq 0$,

34. $$\langle \nabla f^0(z), h(0;z) \rangle < 0 ,$$

then the search function a: $\mathbb{R}^n \to \mathbb{R}^n$ defined by

35. $$a(z) = z + h(\lambda(z);z)$$

36. $$f^0(z + h(\lambda(z);z)) = \min f^0(z + h(\lambda;z))$$

can be used to compute points $z^* \in \mathbb{R}^n$, satisfying $\nabla f^0(z^*) = 0$, as cluster points of a sequence $\{z_i\}$ constructed to satisfy

37. $$z_{i+1} = a(z_i)$$

38. $$f^0(z_{i+1}) < f^0(z_i) .$$

The proof of this should be obvious by now, as well as the various ways in which the minimization in (36) can be relaxed.

2. Boundary Value Problems: General Theory

We now address ourselves to the problem of finding a vector $z \in \mathbb{R}^n$ such that $g(z) = 0$, where g: $\mathbb{R}^n \to \mathbb{R}^m$ is a continuously differentiable function. Obviously we can convert it to the form

1. $$\min \{f^0(z) \triangleq \|g(z)\|^2 \mid z \in \mathbb{R}^n\}$$

and apply to it any one of the general algorithms described in the previous subsection. For example, let a(z) be defined by (1.4) and (1.5) for some continuous

$n \times n$ matrix $D(z) > 0$, i.e.,

$$2. \qquad a(z) = z - \mu(z)\, D(z) \left(\frac{\partial g(z)}{\partial z}\right)^T g(z) \triangleq z + \mu(z)\, h(z)$$

with $\mu(z) > 0$ being the smallest positive real such that

$$3. \qquad \|g(a(z))\|^2 \leq \|g(z + \mu h(z))\|^2 \quad , \text{ for all } \mu \geq 0$$

As we have already pointed out, this search function can be used only to find points z such that $h(z) = -D(z)(\partial g(z)/\partial z)^T g(z) = 0$. Since $D(z) > 0$ is non-singular, $h(z) = 0$ if and only if $(\partial g(z)/\partial z)^T g(z) = 0$. Now suppose that $\partial g(z)/\partial z$ has maximum rank for all $z \in \mathbb{R}^n$ such that $g(z) \neq 0$. Then the algorithm, $z_{i+1} = a(z_i)$, $f^0(z_{i+1}) < f^0(z_i)$, $i = 0, 1, 2, \ldots$, will indeed compute sequences $\{z_i\}$ in $\mathbb{R}^n$ whose accumulation points z^* satisfy $g(z^*) = 0$. This conclusion also remains valid for the other algorithms discussed in the preceeding section. Note that the assumption that $(\partial g(z)/\partial z)$ has maximum rank at all $z \in \mathbb{R}^n$ for which $g(z) \neq 0$, produces the same effect for the problem (1) as did the assumption of convexity of $f^0(\cdot)$ for the problem (1.1), i.e., the algorithms presented will compute points z such that $g(z^*) = 0$ (or such that $f^0(z^*) \leq f^0(z)$ for all $z \in \mathbb{R}^n$.

Now let us consider the following problem,

$$4. \qquad \min \{f^0(z) \mid r(z) = 0\}$$

where f^0: $\mathbb{R}^n \to \mathbb{R}^1$ and r: $\mathbb{R}^n \to \mathbb{R}^\ell$ are continuously differentiable functions. For this case, Theorem (I.2.1) indicates that if $\hat{z} \in R^n$ is optimal, then there exist multipliers $\psi^0, \psi^1, \ldots, \psi^\ell$, not all zero, such that

$$5. \qquad \psi^0 \nabla f^0(\hat{z}) + \sum_{i=1}^{\ell} \psi^i \nabla r^i(\hat{z}) = 0 \quad .$$

(Note that the condition $\mu^0 \triangleq \psi^0 \leq 0$ in (I.2.1) looses its significance when there are no inequality constraints and is therefore omitted). We can use algorithms in the class described previously to find points $z \in \mathbb{R}^n$ satisfying (5), for some $\psi^0, \psi^1, \ldots, \psi^\ell$ not all zero, as follows.

There are two possibilities. Either we must set $\psi^0 = 0$ in (5) or not. If we must set $\psi^0 = 0$, then, setting $\psi = (\psi^1, \psi^2, \ldots, \psi^\ell)$, $\zeta = (z, \psi)$ and defining $g: \mathbb{R}^n \times \mathbb{R}^\ell \to \mathbb{R}^n \times \mathbb{R}^\ell$ by

6. $$g(\zeta) = \begin{pmatrix} \sum_{i=1}^{\ell} \psi^i \nabla r^i(z) \\ r(z) \end{pmatrix},$$

we can try to compute points ζ satisfying $g(\zeta) = 0$ as previously explained.

If $\psi^0 \neq 0$, then we may set $\psi^0 = -1$ and define

7. $$g(\zeta) = \begin{pmatrix} \sum_{i=1}^{\ell} \psi^i \nabla r^i(z) - \nabla f^0(z) \\ r(z) \end{pmatrix}.$$

Suppose that for all z such that $r(z) = 0$, the vectors $\nabla r^i(z)$, $i = 1, 2, \ldots, \ell$, are linearly independent. Then $g(\zeta) = 0$, with g defined by (6), cannot have a solution. However, $g(\zeta) = 0$, with g defined by (7), may have a solution. If one does not know whether the vectors $\nabla r^i(z)$ are linearly independent or not at all points z satisfying $r(z) = 0$, one may try to solve $g(\zeta) = 0$ with $g(\cdot)$ defined by (7) and switch to the formula given by (6) if $\|g(\zeta)\|^2$ does not converge to zero but to a positive value.

8. Remark: Suppose that for all $z \in \mathbb{R}^n$ such that $r(z) = 0$, the vectors $\nabla r^i(z)$, $i = 1, 2, \ldots, \ell$, are linearly dependent. Then for any z satisfying $r(z) = 0$, we

can satisfy (5) with $\psi^0 = 0$, i.e., in this case condition (5) is trivial and, in all likelihood our computations will only produce a non-optimal point z satisfying $r(z) = 0$, which could have been obtained more easily by solving $r(z) = 0$ directly.

9. Remark: Consider now the problem

10. $$\min \{f^0(z) \mid r^i(z) = 0,\ i = 1, 2, \ldots, \ell\ ,$$
$$f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}\ ,\quad z \in \mathbb{R}^n,$$

first defined in (I.1.11), and suppose that all the functions in (10) are continuously differentiable. This problem can be converted to the form (4) as follows. For $i = 1, 2, \ldots, m$, let

11. $$r^{\ell+i}(z) \triangleq (\max \{0, f^i(z)\})^2\ .$$

Then $r^{\ell+i}(z) = 0$ if and only if $f^i(z) \leq 0$, $i = 1, 2, \ldots, m$, and therefore (10) is equivalent to

12. $$\min \{f^0(z) \mid r^i(z) = 0,\ i = 1, 2, \ldots, \ell+m\}$$

However, for any z such that $r^i(z) = 0$, $i = 1, 2, \ldots, \ell+m$, and which is in the interior of the set $\{z \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$, we find that $\nabla r^i(z) = 0$, for $i = \ell+1, \ell+2, \ldots, \ell+m$. Thus, for all such z, the condition (5) is trivially satisfied, and hence the form (12) is an unproductive transcription of (10) as far as obtaining solutions to (10) is concerned.

3. Quasi-Newton Methods

We shall now consider a few specific algorithms which are presented in the class in the preceeding two sections. These methods are primarily characterized by the particular choice of the matrix D(z) in (1.3). The step length ($\mu(z)$ or λ_i) to be used can be determined according to any one of the rules previously indicated: (1.5), (1.20), or (1.32).

i.e., $\min\{f^0(z) \mid z \in \mathbb{R}^n\}$

1. Steepest Descent Method: For the problem (1.1), this method sets $D(z) = I$, the identity matrix, so that

$$h(z) = -\nabla f^0(z) \quad .$$

This algorithm is credited to Cauchy [C2], and more recently to Curry [C5].

2. Modified Newton-Raphson Method: For the problem (1.1), this method sets $D(z) = (\partial^2 f^0(z)/\partial z^2)^{-1}$ (assuming that this inverse matrix exists and that it is continuous and positive definite), so that,

3. $$h(z) = -\left(\frac{\partial^2 f^0(z)}{\partial z^2}\right)^{-1} \nabla f^0(z) \quad .$$

Consequently this method can only be used to solve the problem (1.1) when $f^0(\cdot)$ is strictly convex. This algorithm was also described by McCormack and Zangwill [M1], and again by Zangwill in [Z1].†

To solve problems of the form $g(z) = 0$, or $\nabla f^0(z) = 0$, assuming that $(\partial g(z)/\partial z)^{-1}$, or $(\partial^2 f^0(z)/\partial z^2)^{-1}$, exists, the modified Newton-Raphson method† sets

4. $$D(z) = \left(\frac{\partial g(z)}{\partial z}\right)^{-1} \left(\frac{\partial g(z)}{\partial z}\right)^{T-1} ,$$

or

5. $$D(z) = \left(\frac{\partial^2 f^0(z)}{\partial z^2}\right)^{-1} \left(\frac{\partial^2 f^0(z)^T}{\partial z^2}\right)^{T-1} ,$$

respectively. Thus defined, $D(z)$ is positive definite and, for the problem $g(z) = 0$,

6. $$h(z) = -D(z)\left(\frac{\partial}{\partial z} \frac{1}{2}\|g(z)\|^2\right)^T$$

$$= -D(z)\left(\frac{\partial g(z)}{\partial z}\right)^T g(z)$$

$$= -\left(\frac{\partial g(z)}{\partial z}\right)^{-1} g(z) ,$$

†For the best classical treatment of the Newton-Raphson method, see Kantorovich [K2].

while for the problem $\min \{f^0(z) \mid z \in \mathbb{R}^n\}$,

7. $$h(z) = -D(z) \left(\frac{\partial}{\partial z} \frac{1}{2} \|\nabla f^0(z)\|^2 \right)^T$$
$$= -D(z) \left(\frac{\partial^2 f^0(z)}{\partial z^2} \right)^T \nabla f^0(z)$$
$$= - \left(\frac{\partial^2 f^0(z)}{\partial z^2} \right)^{-1} \nabla f^0(z) .$$

In this case the step length is again set as in (5), (20) or (32), but with $\frac{1}{2}\|g(z)\|$ or $\frac{1}{2}\|\nabla f^0(z)\|$ taking the place of $f^0(z)$.

8. Remark: In Section II.2 we pointed out that the assumption that $f^0(\cdot)$ is convex ensured that the algorithms would compute points z^* such that $f^0(z^*) \leq f^0(z)$ for all $z \in \mathbb{R}^n$. Note that for the modifications of the Newton-Raphson method, the assumption that $f^0(\cdot)$ is strictly convex, must be made to ensure the same results.

The computation of $(\partial g(z)/\partial z)^{-1}$ or $(\partial^2 f^0(z)/\partial z^2)^{-1}$ can be quite expensive. Hence one may wish to use the matrix $(\partial g(z_j)/\partial z)^{-1}$ (or $(\partial^2 f^0(z_j)/\partial z^2)^{-1}$) for $j = i, i+1, \ldots, i+k$ steps, say, and then recompute it again. It is easy to show that this will also result in a convergent algorithm for solving $\min \{\frac{1}{2}\|g(z)\|^2 \mid z \in \mathbb{R}^n\}$, or $(\min \{\frac{1}{2}\|\nabla f^0(z)\|^2 \mid z \in \mathbb{R}^n\}$, provided there exists an $\alpha > 0$ such that

9. $$\left\| \left(\frac{\partial g(z)}{\partial z} \right)^{-1} \left(\frac{\partial g(z)}{\partial z}^T \right)^{-1} \right\| \geq \alpha$$

for all

10. $$z \in \{z \mid \|g(z)\|^2 \leq \|g(z_0)\|^2\};$$

or

11. $$\left\| \left(\frac{\partial^2 f^0(z)}{\partial z^2} \right)^{-1} \left(\frac{\partial^2 f^0(z)}{\partial z^2}^T \right)^{-1} \right\| \geq \alpha$$

for all

12. $$z \in \{z \mid \|\nabla f^0(z)\|^2 \leq \|\nabla f^0(z_0)\|^2\} \quad .$$

In fact, one can generalize the above observation even further, as follows.

13. Theorem: Consider the problem

14. $$\min \{f^0(z) \mid z \in \mathbb{R}^n\}$$

where f^0: $\mathbb{R}^n \to \mathbb{R}^1$ is continuously differentiable.

Suppose that the sequence $z_i \in \mathbb{R}^n$, $i = 0, 1, 2, \ldots$, is constructed to satisfy

15. $$z_{i+1} = z_i + \lambda_i h_i$$

16. $$f^0(z_{i+1}) < f^0(z_i) \quad ,$$

with the $h_i \neq 0$ chosen so as to satisfy

17. $$-\langle \nabla f^0(z_i), h_i \rangle \geq \rho \|\nabla f^0(z_i)\| \, \|h_i\| \quad ,$$

for some fixed $\rho > 0$, and with the $\lambda_i \geq 0$ chosen either to be the smallest real satisfying

18. $$f^0(z_i + \lambda_i h_i) = \min f^0(z_i + \lambda h_i), \quad \lambda \geq 0 \quad ,$$

or else to satisfy

19. $$\lambda_i(1-\alpha) \langle \nabla f^0(z_i), h_i \rangle \leq f^0(z_i + \lambda_i h_i) - f^0(z_i) \leq \lambda_i \alpha \langle \nabla f^0(z_i), h_i \rangle \quad ,$$

for some $\alpha \in (0, \frac{1}{2})$. Then either the sequence $\{z_i\}$ is finite and its last element z_k satisfies $\nabla f^0(z_k) = 0$ or else it is infinite and every accumulation

point z^* of $\{z_i\}$ satisfies $\nabla f^0(z^*) = 0$.†

This theorem can be proved by making use of Theorem (I.3.16) and of the continuity properties of $f^0(\cdot)$. Theorem (13) bears not only on some of the modified Newton-Raphson methods sketched out above, but also on conjugate gradient methods which we shall discuss in the next section.

20. Remark: The modified Newton-Raphson methods described in this section differ from the classical form in one important respect. They require one additional operation -- the computation of $\mu(z_i)$ or λ_i -- to obtain z_{i+1} from z_i. Although this means that these methods involve more work, their convergence is not limited to a small region about the desirable points, which is the case with the classical form of the Newton-Raphson method. As far as rate of convergence is concerned, it is not difficult to see that with the modified Newton-Raphson methods $f^0(z_i) \to f^0(\hat{z})$, or $\|g(z_i)\| \to 0$, as $i \to \infty$, at a rate at least as fast as when the classical form is used.

---

†The convergence properties of the sequence $\{z_i\}$ are unaffected by whether one always uses (19), or always (20), or whether one alternates between them in any manner, to compute λ_i.

4. Conjugate Gradient Methods

Although the rate of convergence of the quasi-Newton methods is quite satisfactory, they require the calculation of second partial derivatives as well as the inversion of possibly large matrices $\partial^2 f^0(z)/\partial z^2$ in the minimization of $f^0(z)$, $z \in \mathbb{R}^n$. We shall now describe a class of methods which require the computation of only the gradient of $f^0(z)$ and whose convergence rates are nevertheless considerably superior to that of the method of steepest descent, though not quite as good as those of some of the quasi-Newton methods.

These methods are called conjugate gradient methods and they all have in common the feature that they require at most n steps to solve the problem

$$1. \qquad \min \{\langle z, Hz\rangle \mid z \in \mathbb{R}^n\} ,$$

where H is an n x n positive definite symmetric matrix. Consider for the moment the problem $\min \{f^0(z) \mid z \in \mathbb{R}^n\}$ and suppose that $f^0(\cdot)$ is twice continuously differentiable. Let $\hat{z}$ be an optimal point for this problem, then $\nabla f^0(\hat{z}) = 0$ and for δz small, i.e., for $z = \hat{z} + \delta z$ in a small ball about $\hat{z}$, we have

$$2. \qquad f^0(z) - f^0(\hat{z}) \doteq \langle \delta z, H(\hat{z})\, \delta z\rangle ,$$

where $H(\hat{z}) = \partial^2 f^0(z)/\partial z^2$. Thus, it is heuristically clear that any method which is efficient in solving the problem (1), very likely, will also be efficient in solving the problem $\min \{f^0(z) \mid z \in \mathbb{R}^n\}$, provided it is convergent for this problem, and provided, of course, that $\partial^2 f^0(z)/\partial z^2$ is positive definite in a neighborhood of the optimal points $\hat{z}$.

We begin the description of these methods by a discussion of a biorthogonalization process.† Suppose that we wish to construct two sequences in $\mathbb{R}^n$, $g_0, g_1, \ldots, g_{n-1}$ and $h_0, h_1, \ldots, h_{n-1}$, such that

---

†We follow here the presentation of Hestenes [H2], see also [H1].

3. $\langle g_i, g_j \rangle = 0$ for all $i \neq j$

and

4. $\langle h_i, Hh_j \rangle = 0$ for all $i \neq j$

where H is an n x n positive definite, symmetric matrix. We can do this by means of the Gramm-Schmidt orthogonalization method as follows.

Let g_0 be arbitrary. Set $h_0 = g_0$. Now, let

4a. $$g_1 = g_0 + \lambda_o Hh_0, \quad \text{with} \quad \lambda_0 = - \frac{\langle g_0, g_0 \rangle}{\langle g_0, Hh_0 \rangle} ,$$

which ensures that $\langle g_0, g_1 \rangle = 0$. Next, set

4b. $$h_1 = g_1 + \gamma_0 h_0, \text{ with } \gamma_0 = - \frac{\langle Hh_0, g_1 \rangle}{\langle Hh_0, h_0 \rangle}$$

This process can obviously be continued as follows:

4c. $$g_2 = \beta_{20} g_0 + g_1 + \lambda_1 Hh_1; \quad h_2 = g_2 + \gamma_1 h_1 + \alpha_{20} h_0 ,$$

where λ_1, β_{20} are chosen to make $\langle g_0, g_2 \rangle = \langle g_1, g_2 \rangle = 0$, and γ_1, α_{20} are chosen to make $\langle h_0, Hh_2 \rangle = \langle h_1, Hh_2 \rangle = 0$, and so forth, until for some $m \leq n$, $g_m = h_m = 0$. The interesting thing about this construction is that the coefficients β_{20} and α_{20}, etc., are zero, as we now show.

5. Theorem: Suppose that for i = 0, 1, 2,...,

6. $$g_{i+1} = g_i + \lambda_i Hh_i, \quad h_{i+1} = g_{i+1} + \gamma_i h_i, \quad g_0 = h_0$$

with

7. $$\lambda_i = - \frac{\langle g_i, g_i \rangle}{\langle g_i, Hh_i \rangle} \qquad \gamma_i = - \frac{\langle Hh_i, g_{i+1} \rangle}{\langle Hh_i, h_i \rangle} .$$

where H is an n × n symmetric positive definite matrix. Then, for i, j = 0, 1, 2,...,

8. $$\langle g_i, g_j \rangle = \delta_{ij} \|g_i\|^2 \quad \text{and} \quad \langle h_i, Hh_j \rangle = \delta_{ij} \|h_i\|_H^2$$

where δ_{ij} is the Kronecker delta.

Proof: We give a proof by induction. By construction, $\langle g_0, g_1 \rangle = \langle g_0, Hh_1 \rangle = 0$. Now, suppose that for some integer k, $0 < k \leq n-1$

8a. $$\langle h_i, Hh_j \rangle = \langle g_i, g_j \rangle = 0 \quad \text{for all} \quad i \neq j,\ i, j \leq k$$

and let $i \in \{1, 2, \ldots, k-1\}$, then

8b. $$\begin{aligned} \langle g_{k+1}, g_i \rangle &= \langle g_k + \lambda_k Hh_k, g_i \rangle \\ &= \lambda_k \langle Hh_k, g_i \rangle \\ &= \lambda_k \langle Hh_k, (h_i - \gamma_{i-1} h_{i-1}) \rangle = 0 . \end{aligned}$$

Also, $\langle g_k, g_{k+1} \rangle = 0$ by the choice of λ_k and

8c. $$\begin{aligned} \langle g_{k+1}, g_0 \rangle &= \langle g_k + \lambda_k Hh_k, g_0 \rangle \\ &= \lambda_k \langle Hh_k, g_0 \rangle = 0 \quad (g_0 = h_0). \end{aligned}$$

Similarly, $\langle h_{k+1}, Hh_k \rangle = 0$ by construction of γ_k. Hence, let $i \in \{0, 1, 2, \ldots, k-1\}$, then

8d. $$\begin{aligned} \langle h_{k+1}, Hh_i \rangle &= \langle g_{k+1} + \gamma_k h_k, Hh_i \rangle \\ &= \langle g_{k+1}, Hh_i \rangle \\ &= \langle g_{k+1}, \frac{1}{\lambda_i} (g_{i+1} - g_i) \rangle = 0 , \end{aligned}$$

i.e., if (8a) is true for k it is also true for k+1. But it is true for k = 0, hence it is true for k = 0, 1, 2,... which completes our proof.

For the vectors constructed above, it is easy to show that the following hold:

9. $$\langle h_i, g_k \rangle = 0 \qquad \text{for all} \qquad i \in \{0, 1, 2, \ldots, k-1\}$$

10. $$\gamma_i = -\frac{\langle Hh_i, g_{i+1} \rangle}{\langle Hh_i, h_i \rangle} = \frac{\langle g_{i+1}, g_{i+1} \rangle}{\langle g_i, g_i \rangle}$$

$$= \frac{\langle g_{i+1}, g_{i+1} \rangle \mp \langle g_{i+1}, g_i \rangle}{\langle g_i, g_i \rangle}$$

$i = 0, 1, \ldots, n-1.$

11. $$\lambda_i = -\frac{\langle g_i, g_i \rangle}{\langle g_i, Hh_i \rangle} = \frac{\langle h_i, g_i \rangle}{\langle h_i, Hh_i \rangle}$$

To establish (9), we note that for $0 \leq i < k$,

11a. $$\langle h_i, g_k \rangle = \langle h_i, g_{i+1} \rangle$$

Now, $\langle h_0, g_1 \rangle = 0$ since $h_0 = g_0$. Suppose therefore that $\langle h_i, g_{i+1} \rangle = 0$ for all $i \in \{0, 1, \ldots, k-1\}$, with $k \leq n-1$. We shall show that it must then also be true for $i = k$.

11b. $$\begin{aligned} \langle h_k, g_{k+1} \rangle &= \langle h_k, g_k + \lambda_k Hh_k \rangle \\ &= \langle h_k, g_k \rangle + \lambda_k \langle h_k, Hh_k \rangle \\ &= \langle g_k, g_k + \gamma_{k-1} h_{k-1} \rangle + \lambda_k \langle Hh_k, h_k \rangle \\ &= \langle g_k, g_k \rangle + \lambda_k \langle Hh_k, h_k \rangle = 0 \end{aligned}$$

To establish 10, we proceed as follows.

11c. $$\begin{aligned} -\frac{\langle Hh_i, g_{i+1} \rangle}{\langle Hh_i, h_i \rangle} &= -\frac{\frac{1}{\lambda_i} \langle g_{i+1} - g_i, g_{i+1} \rangle}{\frac{1}{\lambda_i} \langle g_{i+1} - g_i, h_i \rangle} \\ &= \frac{\langle g_{i+1}, g_{i+1} \rangle}{\langle g_i, h_i \rangle} \\ &= \frac{\langle g_{i+1}, g_{i+1} \rangle}{\langle g_i, g_i + \gamma_{i-1} h_{i-1} \rangle} \\ &= \frac{\langle g_{i+1}, g_{i+1} \rangle}{\langle g_i, g_i \rangle} \\ &= \frac{\langle g_{i+1}, g_{i+1} \rangle \pm \langle g_{i+1}, g_i \rangle}{\langle g_i, g_i \rangle} \end{aligned}$$

To establish (11), we observe that

11d. $$\frac{\langle g_i, g_i \rangle}{\langle g_i, Hh_i \rangle} = \frac{\langle h_i - \gamma_{i-1} h_{i-1}, g_i \rangle}{\langle h_i - \gamma_{i-1} h_{i-1}, Hh_i \rangle} = \frac{\langle h_i, g_i \rangle}{\langle h_i, Hh_i \rangle}$$

Now, suppose that we wish to minimize $f^0(z) \equiv \langle z, Hz\rangle$, where H is an $n \times n$ symmetric positive definite matrix, starting with a given z_0. Suppose that we construct $z_1, z_2, \ldots$ according to

12.
$$\begin{cases} z_{i+1} = z_i + \lambda_i h_i, & i = 0, 1, 2, \ldots \\ h_{i+1} = -Hz_{i+1} + \gamma_i h_i, \; h_0 = -Hz_0, & i = 0, 1, 2, \ldots \end{cases}$$

with γ_i chosen so that $\langle h_i, Hh_{i+1}\rangle = 0$ and λ_i chosen so as to minimize $f^0(z_i + \lambda h_i)$ for $\lambda \geq 0$, i.e., since $\nabla f^0(z) = Hz$, λ_i must satisfy $\langle h_i, Hz_{i+1}\rangle = 0$.

Hence,

13.
$$\lambda_i = -\frac{\langle h_i, Hz_i\rangle}{\langle h_i, Hh_i\rangle}, \qquad \gamma_i = +\frac{\langle Hh_i, Hz_{i+1}\rangle}{\langle Hh_i, h_i\rangle}$$

Now, let $g_i = -\nabla f^0(z_i) = -Hz_i$, then, for $i = 0, 1, 2, \ldots, n-1$, the vectors g_i, h_i satisfy (6) and (7) (see (10), (11)). Also, since (by (9)) $\langle h_j, g_i\rangle = 0$ for all $0 \leq j < i$, and since $i = 0, 1, 2, \ldots$, $g_m = Hz_m = 0$ for some $m \leq n$. Consequently $z_m = 0$, which is the desired solution.

There are two convergent adaptations of the above procedure to problems of the form $\min \{f^0(z) \mid z \in \mathbb{R}^n\}$.

14. Assumption: For the purpose of the conjugate gradient algorithms we are about to describe, we shall assume (in addition to requirement stated in Section II.1) that the set $\{z \mid f^0(z) \leq f^0(z_0)\}$ is bounded, that $f^0(\cdot)$ is twice continuously differentiable, and that there exist $\delta_n > \delta_1 > 0$ such that for all x, $z \in \{z \mid f^0(z) \leq f^0(z_0)\}$,

$$\delta_1 \|x\|^2 \leq \left\langle x, \left(\frac{\partial^2 f^0(z)}{\partial z^2}\right) x\right\rangle \leq \delta_n \|x\|^2 .$$

15. Fletcher-Reeves Algorithm [F4]:

(i) Take $z_0 \in \mathbb{R}^n$ to be a good guess at the minimum of $f^0(z)$.

(ii) For $i = 1, 2, 3, \ldots$, compute z_i, h_i, g_i, according to the rule

16.
$$\begin{cases} z_{i+1} = z_i + \lambda_i h_i \\ g_i = -\nabla f^0(z_i) \\ h_{i+1} = g_{i+1} + \gamma_i h_i, \; h_0 = g_0 , \end{cases}$$

where λ_i is the smallest positive λ such that

17. $$f^0(z_i + \lambda_i h_i) \leq f^0(z_i + \lambda h_i) \qquad \lambda \geq 0, \; i = 0, 1, 2, \ldots .$$

(which implies that $\langle \nabla f^0(z_i + \lambda_i h_i), h_i \rangle = \langle g_{i+1}, h_i \rangle = 0$), and

18. $$\gamma_i = \frac{\langle g_{i+1}, g_{i+1} \rangle}{\langle g_i, g_i \rangle} , \qquad i = 0, 1, 2, \ldots$$

Sequences $\{z_i\}$ constructed by this algorithm can be proved to converge to points z^* such that $\nabla f^0(z^*) = 0$. However, to prove convergence requires a more complicated convergence theorem than the ones described in Section I.3. On the other hand, the following simple modification of (15) can be readily treated by means of the convergence results already established.

19. Polak-Ribière Algorithm [P4]:

(i) Take $z_0 \in \mathbb{R}^n$ to be a good guess at the minimum of $f^0(z)$.

(ii) For $i = 1, 2, 3, \ldots$, compute z_i according to (16) and (17), but with

20. $$\gamma_i = \frac{\langle g_{i+1} - g_i, g_{i+1} \rangle}{\langle g_i, g_i \rangle} , \qquad i = 0, 1, 2, \ldots$$

We shall now prove the convergence of (19).

21. Theorem: If $z_0, z_1, z_2, \ldots$, is a sequence in $\mathbb{R}^n$ constructed according to (19), then there exists an $\alpha > 0$ such that

22. $$-\langle g_i, h_i \rangle \geq \alpha \|g_i\| \, \|h_i\| ,$$

and either $\{z_i\}$ is a finite sequence, terminating at z_k, with $\nabla f^0(z_k) = g_k = 0$, or else every accumulation point z^* of $\{z_i\}$ satisfies $\nabla f^0(z^*) = 0$.

Proof: We only need to prove (22) since the second assertion then follows from Theorem (3.13). Let $H(z) \equiv \partial^2 f^0(z)/\partial z^2$ and let $g(z) = -\nabla f^0(z)$. Expanding $-g(z)$ about any point z_i, $i = 0, 1, 2, \ldots$, in the sequence generated by the algorithm, we find that

23.
$$- g_{i+1} = - g(z_{i+1}) = - g(z_i + \lambda_i h_i)$$
$$= - g_i + \lambda_i \left(\int_0^1 H(z_i + \zeta \lambda_i h_i) d\zeta \right) h_i$$

Since $\langle h_i, g_{i+1} \rangle = 0$, we get

24.
$$\lambda_i = \frac{\langle h_i, g_i \rangle}{\langle h_i, \tilde{H}_i h_i \rangle} = \frac{\langle g_i, g_i \rangle}{\langle h_i, \tilde{H}_i h_i \rangle}$$

where

$$\tilde{H}_i = \left(\int_0^1 H(z_i + \zeta \lambda_i h_i) d\zeta \right) .$$

Note that for all $z \in \mathbb{R}^n$, $\delta_1 \|z\|^2 \leq \langle z, \tilde{H}_i z \rangle \leq \delta_n \|z\|^2$. Now, from (20) and (23) and (24), we get that

25.
$$\gamma_i = - \frac{\langle g_{i+1}, \tilde{H}_i h_i \rangle}{\langle g_i, g_i \rangle} \cdot \frac{\langle g_i, g_i \rangle}{\langle h_i, \tilde{H}_i h_i \rangle}$$
$$= - \frac{\langle g_{i+1}, \tilde{H}_i h_i \rangle}{\langle h_i, \tilde{H}_i, h_i \rangle}$$

so that

26.
$$|\gamma_i| \leq \frac{\|g_{i+1}\| \cdot \|\tilde{H}_i\| \cdot \|h_i\|}{\delta_1 \|h_i\|^2} \leq \frac{\|g_{i+1}\|}{\|h_i\|} \quad \frac{\delta_n}{\delta_1}$$

Now,

27. $$\|h_{i+1}\| \leq \|g_{i+1}\| + |\gamma_i| \, \|h_i\|$$

which, because of (26), becomes

28. $$\|h_{i+1}\| \leq \|g_{i+1}\| \left(1 + \frac{\delta_n}{\delta_1}\right)$$

But

29. $$\langle h_{i+1}, g_{i+1}\rangle = \langle g_{i+1} + \gamma_i h_i, g_{i+1}\rangle = \langle g_{i+1}, g_{i+1}\rangle$$

Hence,

30. $$\frac{\langle h_{i+1}, g_{i+1}\rangle}{\|h_{i+1}\| \, \|g_{i+1}\|} = \frac{\|g_{i+1}\|^2}{\|h_{i+1}\| \cdot \|g_{i+1}\|} = \frac{\|g_{i+1}\|}{\|h_{i+1}\|}$$

$$\geq \frac{\|g_{i+1}\|}{\|g_{i+1}\|(1 + \delta_n/\delta_1)}$$

$$= \frac{1}{\left(1 + \frac{\delta_n}{\delta_1}\right)}$$

which is the desired result.

We conclude this section with the Fletcher-Powell [F3] version of the Davidon [D2] algorithm. This is a very efficient conjugate directions algorithm which approximates the Newton-Raphson method but whose application to control problems is limited by the fact that the dimension of control problems produces unreasonably large core storage requirements.

31. Davidon-Fletcher-Powell Algorithm:

(i) Let $H_0 = I$, the n x n identity matrix. Take z_0 to be a good guess at the

minimum of $f^0(z)$.

(ii) For $i = 1, 2, \ldots$ compute z_i, H_i according to the rule

$$z_{i+1} = z_i + \lambda_i H_i g_i$$

32.

$$H_{i+1} = H_i - \frac{\Delta z_i \rangle\langle \Delta z_i}{\langle \Delta z_i, \Delta g_i \rangle} - \frac{H_i \Delta g_i \rangle\langle H_i \Delta g_i}{\langle \Delta g_i, H_i \Delta g_i \rangle}$$

where $\lambda_i \geq 0$ is the smallest positive real such that

33. $$f^0(z_{i+1}) \leq f^0(z_i + \lambda H_i g_i) \qquad \text{for all } \lambda \geq 0 \ .$$

In (32), $\Delta z_i = z_{i+1} - z_i$, $g_i = -\nabla f^0(z_i)$, $\Delta g_i = g_{i+1} - g_i$, and, for $y \in \mathbb{R}^n$, $x \rangle\langle y$ is a dyad, i.e., an $n \times n$ matrix whose ij^{th} element is $x^i y^j$.

34. Lemma: For $i = 0, 1, 2, \ldots$, the matrices H_i are symmetric and positive definite.

Proof: For $i = 0$, $H_i = I$, a symmetric positive definite matrix. By (32), H_{i+1} is symmetric if H_i is symmetric, hence we only need to prove that the H_i are positive definite. We give a proof by induction. Suppose $H_i > 0$. Then, for any nonzero vector $z \in \mathbb{R}^n$,

35. $$\langle z, H_{i+1} z \rangle = \langle z, H_i z \rangle - \frac{\langle z, \Delta z_i \rangle^2}{\langle \Delta z_i, \Delta g_i \rangle} - \frac{\langle z, H_i \Delta g_i \rangle^2}{\langle \Delta g_i, H_i \Delta g_i \rangle}$$

Since $H_i > 0$, $H_i^{1/2}$ is a well defined symmetric positive definite let $p = H_i^{1/2} z$ and let $q = H_i^{1/2} \Delta g_i$. Then (35) becomes

36. $$\langle z, H_{i+1} z \rangle = \frac{\langle p,p\rangle\langle q,q\rangle - \langle p,q\rangle^2}{\langle q,q\rangle} - \frac{\langle z, \Delta z_i \rangle^2}{\langle \Delta z_i, \Delta g_i \rangle}$$

Applying Schwartz's inequality we obtain $\langle p,p\rangle\langle q,q\rangle = \|p\|^2 \|q\|^2 \geq \langle p,q\rangle^2$, and hence

37. $$\langle z, H_{i+1} z \rangle \geq - \frac{\langle z, \Delta z_i \rangle^2}{\langle \Delta z_i, \Delta q_i \rangle}$$

Now, since $\langle \Delta z_i, g_{i+1} \rangle = 0$,

38. $$- \langle \Delta z_i, \Delta g_i \rangle = - \langle \Delta z_i, g_{i+1} \rangle + \langle \Delta z_i, g_i \rangle$$

$$= + \langle \Delta z_i, g_i \rangle$$

$$= \lambda_i \langle H_i g_i, g_i \rangle > 0$$

and hence,

39. $$\langle z, H_{i+1} z \rangle \geq 0 \quad .$$

Now suppose that $z \neq 0$ but $\langle z, H_{i+1} z \rangle = 0$. Then, from (36) and (38), we must have that

(i) $\langle z, \Delta z_i \rangle = 0$ and

(ii) $\langle p,p \rangle \langle q,q \rangle = \langle p,q \rangle^2$.

But (ii) implies that $p = \alpha q$, for some α real, i.e., that $z = \alpha \Delta g_i$ and hence by (i) we should have that $\langle \Delta g_i, \Delta z_i \rangle = 0$, which contradicts (38). Therefore $H_{i+1} > 0$, which completes our proof.

40. Theorem: For $i = 0, 1, 2, \ldots$, let $\mu_i^n > 0$ be the largest eigenvalue of H_i and let $\mu_i^1 > 0$ the smallest eigenvalue of H_i. Suppose that there exist real M and m such that $\mu_i^n \leq M$ and $\mu_i^1 \geq m$ for $i = 0, 1, 2, \ldots$ Then, either the sequence $\{z_i\}$ generated by (32) and (33) is finite, terminating at z_k, with $\nabla f^0(z_k) = 0$, or else $\{z_i\}$ is infinite and every accumulation point z^* of $\{z_i\}$ satisfies $\nabla f^0(z^*) = 0$.

This theorem also follows directly from Theorem (3.13) and we therefore omit its proof. (It was also presented by Daniel [D1]).

41. Remark: At present there are no known assumptions on $f^0(\cdot)$ which guarantee that the sequence of matrices H_i generated by (32) and (33) will have eigenvalues bounded both from below and from above. In current practice, the sequence H_i is restarted periodically from $H_0 = I$, to prevent the eigenvalues of H_i from

becoming excessively spread out. When the sequence H_i is constructed using (32) and (33), with the additional feature that H_i is set equal to H_0 for $i = 0, k, 2k, \ldots$, it follows directly from Theorem (3.13) that every accumulation point z^* of $\{z_i\}$ will satisfy $\nabla f^0(z^*) = 0$.

In conclusion, we wish to point out that there exist a large number of variants of the methods presented in this section, all of which are heuristically derived, and whose computational merits are briefly discussed in a survey paper by M.J.D.Powell [P6], to which the interested reader is referred.

5. Applications to Optimal Control

As we have already seen in Chapter I, discrete optimal control problems are mathematical programming problems -- usually of large dimension. Therefore all the algorithms discussed so far are, at least in principle, applicable to these problems. In this section we shall discuss how the immense structure of the optimal control problem can be utilized to produce simplifications in the calculation of the vector $h(z)$, without which our task may well be hopeless. In general, these simplifications will consist of substituting a sequence of "low dimensional" operations for a single "high dimensional" operation.

We begin by considering the free end optimal control problem:

$$1. \qquad \min \sum_{i=0}^{k-1} f_i^0(x_i,u_i) + \varphi(x_k),\ x_i \in \mathbb{R}^\nu,\ u_i \in \mathbb{R}^\mu ,$$

subject to

$$2. \qquad x_{i+1} - x_i = f_i(x_i,u_i) \qquad i = 0, 1, \ldots, k-1 \text{ with } x_0 = \hat{x}_0$$

This form of problem usually arises when penalty functions, to be discussed later, are used to cope with inequality constraints on the states x_i and on the controls u_i.

Since $x_0 = \hat{x}_0$ is given, the x_i, $i = 1, 2, \ldots, k$, are uniquely determined by the sequence $z = (u_0, u_1, \ldots, u_{k-1})$ and (2), i.e., $x_i = x_i(z)$. Therefore (1) is a problem of the form

3. $$\min f^0(z),\ z \in \mathbb{R}^n \qquad (n = k\mu)$$

where

4. $$f^0(z) = \sum_{i=0}^{k-1} f_i^0(x_i(z), u_i) + \varphi(x_k(z))$$

In what follows, when we write x_i, we mean $x_i(z)$.

We now show a method for computing $\nabla f^0(z)$. Note that

5. $$\nabla f^0(z)^T \triangleq \frac{\partial f^0(z)}{\partial z} = \left(\frac{\partial f^0(z)}{\partial u_0} \mid \frac{\partial f^0(z)}{\partial u_1} \mid \cdots \mid \frac{\partial f^0(z)}{\partial u_{k-1}} \right)^\dagger$$

and that for $i = 0, 1, 2, \ldots, k-1$,

6. $$\frac{\partial f^0(z)}{\partial u_i} = \frac{\partial f_i^0(x_i,u_i)}{\partial u_i} + \sum_{j=i+1}^{k-1} \frac{\partial f_j^0(x_j,u_j)}{\partial x_j} \frac{\partial x_j(z)}{\partial u_i} + \frac{\partial \varphi(x_k)}{\partial x_k} \frac{\partial x_k(z)}{\partial u_i} ,$$

Now, for $k \geq j \geq i$, and $i = 0, 1, 2, \ldots, k$, let $\Phi_{j,i}$ be an $n \times n$ matrix satisfying $\Phi_{i,i} = I$, the identity matrix, and

7. $$\Phi_{j+1,i} - \Phi_{j,i} = \frac{\partial f_j(x_j,u_j)}{\partial x_j} \Phi_{j,i}, \qquad j = i, i+1, \ldots, k-1 .$$

Then

8. $$\frac{\partial x_j(z)}{\partial u_i} = \Phi_{j,i+1} \frac{\partial f_i(x_i,u_i)}{\partial u_i} , \quad \text{for} \quad j = i+1, i+2, \ldots, k$$

$$= 0 \quad \text{for } j = 1, 2, \ldots, i.$$

---

†Note that $\nabla f^0(z)$ is a column vector, but $\frac{\partial f^0(z)}{\partial z}$ is a $1 \times n$ Jacobian matrix, i.e., a row vector.

Hence, for i = 0, 1, 2,..., k-1,

$$9. \qquad \frac{\partial f^0(z)}{\partial u_i} = \frac{\partial f_i^0(x_i,u_i)}{\partial u_i} + \left(\sum_{j=i+1}^{k-1} \frac{\partial f_j^0(x_j,u_j)}{\partial x_j} \Phi_{j,i+1} \right) \frac{\partial f_i(x_i,u_i)}{\partial u_i}$$

$$+ \frac{\partial \varphi(x_k)}{\partial x_k} \Phi_{k,i+1} \frac{\partial f_i(x_i,u_i)}{\partial u_i} \quad .$$

Since we are in the habit of working only with column vectors, we transpose (9) to get

$$10. \qquad \left(\frac{\partial f^0(z)}{\partial u_i}\right)^T = \left(\frac{\partial f_i^0(x_i,u_i)}{\partial u_i}\right)^T + \left(\frac{\partial f_i(x_i,u_i)}{\partial u_i}\right)^T \left(\sum_{j=i+1}^{k-1} \Phi_{j,i+1}^T \left(\frac{\partial f_j^0(x_j,u_j)}{\partial x_j}\right)^T \right)$$

$$+ \left(\frac{\partial f_i(x_i,u_i)}{\partial u_i}\right)^T \Phi_{k,i+1}^T \left(\frac{\partial \varphi(x_k)}{\partial x_k}\right)^T \quad .$$

Now, for i = 1, 2,..., k, let p_i be the solution of

$$11. \qquad p_i - p_{i+1} = \left(\frac{\partial f_i(x_i,u_i)}{\partial x_i}\right)^T p_{i+1} + \left(\frac{\partial f_i^0(x_i,u_i)}{\partial x_i}\right)^T ,$$

$$\text{with} \qquad p_k = \left(\frac{\partial \varphi(x_k)}{\partial x_k}\right)^T \quad .$$

Then

$$12. \qquad p_i = \Phi_{k,i}^T \left(\frac{\partial \varphi(x_k)}{\partial x_k}\right)^T + \sum_{j=i}^{k-1} \Phi_{j,i}^T \left(\frac{\partial f_j^0(x_j,u_j)}{\partial x_j}\right)^T$$

and

$$13. \qquad \left(\frac{\partial f^0(z)}{\partial u_i}\right)^T = \left(\frac{\partial f_i^0(x_i,u_i)}{\partial u_i}\right)^T + \left(\frac{\partial f_i(x_i,u_i)}{\partial u_i}\right)^T p_{i+1}$$

Thus, to compute $\nabla f^0(z)$, we first solve (11) to obtain the vectors $p_1, p_2, \ldots, p_k$ and then use (13) to complete the evaluation.

With the procedure for computing $\nabla f^0(z)$ developed, it should now be obvious how one can use the method of steepest descent (3.1), or any one of the conjugate gradient methods (4.15), (4.19), (4.31), for solving (1).

We shall now show how to apply the modified Newton-Raphson method (3.2) to an optimal control boundary value problem. Note that the procedure we are about to develop also applies to the solution of problems of the form (1). Consider the problem

$$\text{14.} \qquad \min \sum_{i=0}^{k-1} f_i^0(x_i,u_i) + \varphi(x_k),\ x_i \in \mathbb{R}^\nu,\ u_i \in \mathbb{R}^\mu ,$$

subject to

$$\text{15.} \qquad x_{i+1} - x_i = f_i(x_i,u_i), \qquad i = 0, 1, \ldots, k-1$$

where x_0 and $g_k(x_k) = 0$ ($g_k\colon \mathbb{R}^\nu \to \mathbb{R}^\ell$) are given. We assume that all the derivatives we shall use are continuous in all their arguments.

As we recall, the modified Newton-Raphson Method can be used to obtain a control sequence $\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{k-1}$, and a corresponding trajectory $\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_k$ satisfying the necessary conditions of optimality for (14), (15), developed in Section I.2, i.e., it can be used to find vectors $\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_k$, $\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{k-1}$, and multiplier vectors $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k$, π such that

$$\text{16.} \qquad \hat{x}_{i+1} - \hat{x}_i - f_i(\hat{x}_i,\hat{u}_i) = 0 \qquad \text{for } i = 0, 1, \ldots, k-1 ;$$

$$\text{17.} \qquad g_k(\hat{x}_k) = 0$$

$$\text{18.} \qquad \hat{p}_i - \hat{p}_{i+1} - \left(\frac{\partial f_i(\hat{x}_i,\hat{u}_i)}{\partial x_i}\right)^T \hat{p}_{i+1} + \left(\frac{\partial f_i^0(\hat{x}_i,\hat{u}_i)}{\partial x_i}\right)^T = 0$$

$$\text{for } i = 1, 2, \ldots, k-1 ;$$

19. $$p_k - \left(\frac{\partial g_k(\hat{x}_k)}{\partial x_k}\right)^T \pi + \left(\frac{\partial \varphi(\hat{x}_k)}{\partial x_k}\right)^T = 0$$

20. $$-\left(\frac{\partial f_i^0(\hat{x}_i,\hat{u}_i)}{\partial u_i}\right)^T + \left(\frac{\partial f_i(\hat{x}_i,\hat{u}_i)}{\partial u_i}\right)^T p_{i+1} = 0$$

for $i = 0, 1, \ldots, k-1$.

Note: We have assumed that the problem (14), (15) is not "degenerate", i.e., that the multiplier of $(\partial f_i^0(\hat{x}_i,\hat{u}_i)/\partial x_i)^T$ in (18) is -1 and not zero.

Let $z = (x_1, x_2, \ldots, x_k, p_1, p_2, \ldots, p_k, \pi, u_0, u_1, \ldots, u_{k-1})$, then (16), (17), (18), (19) and (20) define a map $r: \mathbb{R}^{k\nu} \times \mathbb{R}^{k\nu} \times \mathbb{R}^{\ell} \times \mathbb{R}^{k\mu} \to \mathbb{R}^{k\nu} \times \mathbb{R}^{\ell} \times \mathbb{R}^{(k-1)\nu} \times \mathbb{R}^{\nu} \times \mathbb{R}^{(k)\mu}$, i.e., if we set $n = 2k\nu + \ell + k\mu$, then

$r\colon\ \mathbb{R}^n \to \mathbb{R}^n$. We wish to find a $\hat{z} \in \mathbb{R}^n$ such that $r(\hat{z}) = 0$.

As we recall from (3.7), that $h(z)$, the "direction" of motion for the modified Newton-Raphson method, is given by

21. $$h(z) = -\left(\frac{\partial r(z)}{\partial z}\right)^{-1} r(z) \quad .$$

(We obviously assume that $(\partial r(z)/\partial z)^{-1}$ exists.) Hence $h(z)$ is the solution to the equation

22. $$\frac{\partial r(z)}{\partial z}\ h(z) = r(z) \quad .$$

With $r(z)$ defined as above, and $h(z) \triangleq (\delta x_1, \delta x_2, \ldots, \delta x_k, \delta p_1, \delta p_2, \ldots, \delta p_k, \delta\pi, \delta u_0, \delta u_1, \ldots, \delta u_{k-1})$, (22) can be expanded using (16), (17), (18) and (19). Thus, from (16) we get

23. $$\delta x_{i+1} - \delta x_i = \frac{\partial f_i(x_i,u_i)}{\partial x_i}\ \delta x_i + \frac{\partial f_i(x_i,u_i)}{\partial u_i}\ \delta u_i - v_i, \quad i = 0, 1, 2, \ldots, k-1 \ ,$$

where $\delta x_0 = 0$ and

24. $$v_i = x_{i+1} - x_i - f_i(x_i,u_i) \ ;$$

from (18),

25. $$\begin{aligned} \delta p_i - \delta p_{i+1} = {} & \left(\frac{\partial f_i(x_i,u_i)}{\partial x_i}\right)^T \delta p_{i+1} \\ & + \frac{\partial}{\partial x_i}\left[\left(\frac{\partial f_i(x_i,u_i)}{\partial x_i}\right)^T p_{i+1}\right] \delta x_i \\ & + \frac{\partial}{\partial u_i}\left[\left(\frac{\partial f_i(x_i,u_i)}{\partial x_i}\right)^T p_{i+1}\right] \delta u_i \\ & - \left(\frac{\partial^2 f_i^0(x_i,u_i)}{\partial x_i^2}\right)^T \delta x_i - \left(\frac{\partial^2 f_i^0(x_i,u_i)}{\partial u_i\, \partial x_i}\right)^T \delta u_i - w_i \ , \end{aligned}$$

for $i = 1, 2, \ldots, k-1$, where

26. $$w_i = p_i - p_{i+1} - \left(\frac{\partial f_i(x_i,u_i)}{\partial x_i}\right)^T p_{i+1} + \left(\frac{\partial f_i^0(x_i,u_i)}{\partial x_i}\right)^T$$

From (19) we get

27. $$\delta p_k - \left(\frac{\partial g_k(x_k)}{\partial x_k}\right)^T \delta\pi - \frac{\partial}{\partial x_k}\left(\left(\frac{\partial g_k(x_k)}{\partial x_k}\right)^T \pi - \left(\frac{\partial \varphi(x_k)}{\partial x_k}\right)^T\right) \delta x_k - y_k = 0$$

where

28. $$y_k = p_k - \left(\frac{\partial g_k(x_k)}{\partial x_k}\right)^T \pi + \left(\frac{\partial \varphi(x_k)}{\partial x_k}\right)^T$$

From (20) we get, for i = 0, 1, 2,..., k-1,

$$- \left(\frac{\partial^2 f_i^0(x_i,u_i)}{\partial x_i\, u_i}\right)^T \delta x_i - \left(\frac{\partial^2 f_i^0(x_i,u_i)}{\partial u_i^2}\right)^T \delta u_i$$

29. $$+ \frac{\partial}{\partial x_i}\left[\left(\frac{\partial f_i(x_i,u_i)}{\partial u_i}\right)^T p_{i+1}\right] \delta x_i + \frac{\partial}{\partial u_i}\left[\left(\frac{\partial f_i(x_i,u_i)}{\partial u_i}\right)^T p_{i+1}\right] \delta u_i$$

$$+ \left(\frac{\partial f_i(x_i,u_i)}{\partial u_i}\right)^T \delta p_{i+1} - t_i = 0 ,$$

where

30. $$t_i = \left(\frac{\partial f^0(x_i,u_i)}{\partial u_i}\right)^T - \left(\frac{\partial f_i(x_i,u_i)}{\partial u_i}\right)^T p_{i+1}$$

From (29), for $i = 0, 1, 2, \ldots, k-1$, δu_i can be solved for in terms of δx_i and δp_{i+1}. Since (29) is linear in the variational arguments (i.e., the δx_i, δp_i, and δu_i), the expressions for the δu_i will be linear in their arguments (δx_i, δp_{i+1}). Substituting for δu_i into (23) and (25), we now have to solve a boundary value problem of the form below to complete our calculations of $h(z) = \delta z$.

$$31a. \qquad \delta x_{i+1} = A_i \delta x_i + B_i \delta p_{i+1} - v_i \qquad i = 0, 1, 2, \ldots, k-1$$

$$31b. \qquad \delta p_i = C_i \delta x_i + D_i \delta p_{i+1} - w_i \qquad i = 1, 2, \ldots, k-1$$

$$31c. \qquad \delta x_0 = 0, \; G_k \delta x_k + g_k = 0, \; \delta p_k = G_k^T \delta \pi + M_k \delta x_k + y_k$$

We may now proceed in one of several ways to solve (31) for δx_i, $i = 1, 2, \ldots, k$, δp_i, $i = 0, 1, \ldots, k$, and $\delta \pi$. The most straightforward one is to simply set up (31) as a combined array and try to invert the matrix on the left, i.e.,

32.
$$\begin{pmatrix} I & 0 & 0 & \cdots & 0 & 0 & 0 & -B_0 & 0 & \cdots & 0 \\ -A_1 & I & 0 & \cdots & 0 & 0 & 0 & 0 & -B_1 & 0 & \cdots & 0 \\ 0 & -A_2 & I & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -B_2 & 0 \cdots 0 \\ \vdots & & & & & & & & & & & \\ 0 & 0 & 0 & -A_{k-1} & I & 0 & \cdots & & & 0 & -B_{k-1} & 0 \\ -C_1 & 0 & 0 & & \cdots & 0 & I & -D_1 & 0 & \cdots & 0 & 0 \\ 0 & -C_2 & 0 & \cdots & & 0 & 0 & I & -D_2 & 0 & \cdots & 0 \\ \vdots & & & & & & & & & & & \\ 0 & \cdots & 0 & -C_{k-1} & 0 & \cdots & & & 0 & I & -D_{k-1} & 0 \\ 0 & \cdots & 0 & G_k & 0 & \cdots & & & & & & 0 \\ 0 & \cdots & -M_k & \cdots & & & & & 0 & I & -G_k^T & \end{pmatrix} \begin{pmatrix} \delta x_1 \\ \delta x_2 \\ \vdots \\ \delta x_k \\ \delta p_1 \\ \vdots \\ \delta p_{k-1} \\ \delta p_k \\ \delta \pi \end{pmatrix} = \begin{pmatrix} -v_0 + A_0 \delta x_0 \\ -v_1 \\ \vdots \\ -v_{k-1} \\ -w_1 \\ \\ -w_{k-1} \\ -g_k \\ y_k \end{pmatrix}$$

(32)

Obviously, this is an enormous matrix in any practical situation, and cannot be solved (and even, possibly, stored in core) without a certain amount of cunning which takes its structure into consideration.

In what follows we shall assume that the inverse matrices used do indeed exist, and shall develop a method for solving (32) which exploits its structure to the fullest extent. When these inverses do not exist, it obviously becomes much harder to utilize the structure of (32) in its solution.

Now, for $i = k-1$, we have from (31) that

33. $$\delta p_{k-1} = C_{k-1} \delta x_{k-1} + D_{k-1} \delta p_k - w_{k-1} .$$

If we treat, for the time being, δp_k as a known constant, we see that

34. $$\delta p_{k-1} = K_{k-1} \delta x_{k-1} + \varkappa_{k-1}$$

Where $K_{k-1} = C_{k-1}$ and $\varkappa_{k-1} = D_{k-1} \delta p_k - w_{k-1}$. Hence let us suppose that for $j = i+1$,

35. $$\delta p_{i+1} = K_{i+1}\,\delta x_{i+1} + \varkappa_{i+1}$$

and find the corresponding expression for δp_i. From (31a) and (35)

36. $$\delta x_{i+1} = A_i \delta x_i + B_i K_{i+1}\,\delta x_{i+1} + B_i \varkappa_{i+1} - \nu_i$$

Solving for x_{i+1}, we get

36b. $$\delta x_{i+1} = (I - B_i K_{i+1})^{-1} [A_i \delta x_i + B_i \varkappa_{i+1} - \nu_i] .$$

Substituting from (35) for δp_{i+1} in (31b), we get,

37. $$\begin{aligned} \delta p_i &= C_i \delta x_i + D_i K_{i+1}(I - B_i K_{i+1})^{-1} [A_i \delta x_i + B_i \varkappa_{i+1} + \nu_i] + D_i \varkappa_{i+1} - w_i \\ &= [C_i + D_i K_{i+1}(I - B_i K_{i+1})^{-1} A_i]\,\delta x_i \\ &+ [D_i + D_i K_{i+1}(I - B_i K_{i+1})^{-1} B_i]\,\varkappa_{i+1} \\ &- D_i K_{i+1}(I - B_i K_{i+1})^{-1} \nu_i - w_i \end{aligned}$$

i.e., we find that

38. $$\delta p_i = K_i \delta x_i + \varkappa_i ,$$

where for $i = k-1, \ldots, 1$, K_i and $\varkappa_i$ satisfy

39. $$K_i = [C_i + D_i K_{i+1}(I - B_i K_{i+1})^{-1} A_i]$$

40. $$\begin{aligned} \varkappa_i &= D_i [I + K_{i+1}(I - B_i K_{i+1})^{-1} B_i]\,\varkappa_{i+1} \\ &- D_i K_{i+1}(I - B_i K_{i+1})^{-1} \nu_i - w_i \end{aligned}$$

Since (35) is true for $j = k-1$ (i.e., for $i = k-2$), we conclude by induction that it must also be true for $j = k-2, k-3, \ldots, 1$. Note that we still must give boundary conditions for (39) and (40). Since $K_{k-1} = C_{k-1}$, we set $K_k = 0$, hence from (35), $\varkappa_k = \delta p_k$.

To solve (31) with the boundary conditions given, we compute the K_i, $i = k-1, k-2, \ldots, 1$, from (39), with $K_k = 0$. Then, from (39), we compute $\varkappa_i$, in the form

41. $$\varkappa_i = F_i \delta p_k + \xi_i, \qquad i = k-1, k-2, \ldots, 1 ,$$

(actually, we only compute the matrices F_i and the vectors ξ_i). Substituting from (41) into (36a), and using the fact that $\delta x_0 = 0$, we obtain δx_k in the form

42. $$\delta x_k = E \delta p_k + \xi .$$

Now making use of the boundary conditions (31c) for δx_k and δp_k, we obtain,

43. $$G_k \delta x_k = G_k (I - EM_k)^{-1} (E G_k^T \delta\pi + E y_k + \xi) = g_k$$

which yields

44. $$\delta\pi = (G_k (I-EM_k)^{-1} E G_k^T)^{-1} (g_k - G_k (I-EM_k)^{-1} (E y_k + \xi))$$

With $\delta\pi$ determined we can now obtain all the required quantities.

There are a number of cases in which the amount of labor involved in solving the necessary condition equations (16) to (20) is considerably less than for the general case we have described above. We shall now consider two of these rather important special cases.

Case I

45. $$\text{minimize } \frac{1}{2} \sum_{i=0}^{k-1} \|u_i\|^2 + \frac{1}{2} \sum_{i=1}^{k-1} \|x_i - x_i^*\|^2$$

subject to

46. $$x_{i+1} = Ax_i + Bu_i, \qquad i = 0, 1, 2, \ldots, k-1 ,$$

with $x_i \in \mathbb{R}^\nu$, $u_i \in \mathbb{R}^\mu$, and $x_0 = \hat{x}_0$, $x_k = \hat{x}_k$ given.

Assuming that $p^0 = -1$ is true, the necessary conditions (16) to (20) require for this problem, that the optimal $\hat{u}_i$ and $\hat{x}_i$ satisfy

47. $\hat{x}_{i+1} = A\hat{x}_i + B\hat{u}_i, \quad i = 0, 1,\ldots, k-1, \ \hat{x}_0 = \hat{x}_0, \ \hat{x}_k = \hat{x}_k$;

48. $\hat{p}_i = A^T\hat{p}_{i+1} - (\hat{x}_i - x_i^*), \quad i = 1, 2,\ldots, k$;

49. $-\hat{u}_i + B^T\hat{p}_{i+1} = 0, \quad i = 0, 1,\ldots, k-1$.

From (49), $\hat{u}_i = B^T\hat{p}_{i+1}$, $i = 0, 1,\ldots, k-1$, and hence the optimal $\hat{u}_0, \hat{u}_1,\ldots, \hat{u}_{k-1}$ and $\hat{x}_0, \hat{x}_1,\ldots, \hat{x}_k$, and $\hat{p}_1, \hat{p}_2,\ldots, \hat{p}_k$ satisfy (49) and

50a. $\hat{x}_{i+1} = A\hat{x}_i + BB^T\hat{p}_{i+1}, \quad i = 0, 1,\ldots,k-1, \quad \hat{x}_0 = \hat{x}_0, \ \hat{x}_k = x_k$

50b. $\hat{p}_i = A^T\hat{p}_{i+1} - (\hat{x}_i - x_i^*)$

Obviously, there is no need to use the modified Newton-Raphson method to solve (50), since it is linear and of the form of equations we already know how to solve.

Case II

51. $$\text{minimize} \sum_{i=0}^{k-1} f_i^0(x_i) + \frac{1}{2} \|u_i\|^2$$

subject to

52. $x_{i+1} - x_i = f_i(x_i) + Bu_i, \quad i = 0, 1,\ldots, k-1$,

$x_i \in \mathbb{R}^\nu$, $u_i \in \mathbb{R}^\mu$, with $x_0 = \hat{x}_0$, $x_k = \hat{x}_k$, given, Again assuming that $p^0 = -1$, the necessary conditions (16) to (20) require for this problem that the optimal $\hat{u}_i$ and $\hat{x}_i$ must satisfy

53. $\hat{x}_{i+1} - \hat{x}_i = f_i(\hat{x}_i) + B\hat{u}_i, \quad i = 0, 1,\ldots, k-1, \ \hat{x}_0 = \hat{x}_0, \ \hat{x}_k = \hat{x}_k$;

54. $$\hat{p}_i - \hat{p}_{i+1} = \frac{\partial f_i(\hat{x}_i)^T}{\partial x_i}\,\hat{p}_{i+1} - \nabla f_i^0(\hat{x}_i), \qquad i = 1, 2, \ldots, k \ ;$$

55. $$-\hat{u}_i + B^T \hat{p}_{i+1} = 0, \qquad i = 0, 1, \ldots, k-1 \ .$$

Since $\hat{u}_i$ can be obtained in terms of $\hat{p}_{i+1}$ from (55), we eliminate (55) before applying the modified Newton-Raphson method to the following resulting system:

56. $$x_{i+1} - x_i = f_i(x_i) + BB^T p_{i+1}, \quad i = 0, 1, \ldots, k-1, \ x_0 = \hat{x}_0, \ x_k = \hat{x}_k$$

57. $$p_i - p_{i+1} = \left(\frac{\partial f_i(x_i)}{\partial x_i}\right)^T p_{i+1} - \nabla f_i^0(x_i), \quad i = 1, 2, \ldots, k-1.$$

In the above derivation, note the great saving in labor which results from the fact that u_i may be expressed in terms of p_{i+1}, as compared with the general case presented in the beginning of this section.

For further reading on the use of the Newton-Raphson method in control problems, the reader should consult [B4] and [L1]. Also, to see how dynamic programming can be used to develop a related algorithm, the reader should consult [M2].

6. Minimization Without Calculation of Derivatives

The calculation of the derivatives of a function may be quite costly and one may therefore wish to avoid it. There are basically two types of algorithms for minimization which avoid the calculation of derivatives. The first type approximates derivatives by finite differences, the second type is intrinsically independent of derivative calculations. We shall now give a few examples of algorithms which avoid the computation of derivatives.

Consider again the problem

1. $$\min \{f^0(z) \mid z \in \mathbb{R}^n\}$$

where f^0: $\mathbb{R}^n \to \mathbb{R}^1$ is at least once continuously differentiable, and for every α real, the set $\{z \mid f^0(z) \leq \alpha\}$ is bounded. The following three methods approximate derivatives by means of finite differences.

2. Modified Steepest Descent: Assume that a $z_0 \in \mathbb{R}^n$ and an $\epsilon_0 > 0$ are given.

Step 0: Set $z = z_0$

Step 1: Set $\epsilon = \epsilon_0$

Step 2: Compute the vector $h_\epsilon(z) \in \mathbb{R}^n$, whose i^{th} component, $h_\epsilon^i(z)$, is defined by

3. $$h_\epsilon^i(z) \triangleq -\frac{1}{\epsilon}(f^0(z + \epsilon\, \xi_i) - f^0(z)), \quad i = 1, 2, \ldots, n$$

and ξ_i is the i^{th} column of a $n \times n$ unit matrix, i.e., $\xi_1 = (1, 0, \ldots, 0)$, $\xi_2 = (0, 1, 0, \ldots, 0)$ etc.

Step 3: Compute a $\mu_\epsilon(z) \geq 0$ to be the smallest real such that

4. $$f^0(z + \mu_\epsilon(z)\, h_\epsilon(z)) \leq f^0(z + \mu h_\epsilon(z)) \quad \text{for all } \mu \geq 0$$

Step 4: If $f^0(z + \mu_\epsilon(z)\, h_\epsilon(z)) - f^0(z) \leq -\epsilon$, set $z = z + \mu_\epsilon(z)\, h_\epsilon(z)$ and go to Step 1. If $f^0(z + \mu_\epsilon(z)\, h_\epsilon(z)) - f^0(z) > -\epsilon$, set $\epsilon = \epsilon/2$ and go to Step 2.

We can use Theorem (I.3.21) to show that if $\{z_i\}$ is a sequence constructed by the algorithm (2) (i.e., $z_1, z_2, \ldots,$ are the consecutive values assigned to z in Step 4), then either $\{z_i\}$ is finite, i.e., the algorithm jams up at z_k, and $\nabla f^0(z_k) = 0$, or else it is infinite and every accumulation point z^* of $\{z_i\}$ satisfies $\nabla f^0(z^*) = 0$.

5. Remark: Instead of choosing $\mu_\varepsilon(z)$ to satisfy (4), we may take $\mu_\varepsilon(z)$ to be any value satisfying, for some $\alpha \in (0, \frac{1}{2})$,

6. $$-\mu_\varepsilon(z)(1-\alpha)\langle h_\varepsilon(z), h_\varepsilon(z)\rangle \leq f^0(z + \mu_\varepsilon(z)\, h_\varepsilon(z)) - f^0(z) \leq$$
$$\leq \mu_\varepsilon(z)\,\alpha \quad \langle h_\varepsilon(z), h_\varepsilon(z)\rangle \ .$$

For a comparison see (II.1.18).

7. Modified Newton-Raphson Method (Goldstein and Price [G2]

Suppose that $f^0(\cdot)$ is strictly convex and twice continuously differentiable; and that a $z_0 \in \mathbb{R}^n$ and an $\varepsilon_0 > 0$ are given.

Step 0: Set $z = z_0$.

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

Step 2: Compute the n x n matrix $H_\varepsilon(z)$ whose i^{th} column is $\frac{1}{\varepsilon}[\nabla f^0(z + \varepsilon\,\xi_i) - \nabla f^0(z)]$, where ξ_i is the i^{th} column of the n x n unit matrix.

Step 3: If $H_\varepsilon(z)^{-1}$ exists, and $\langle \nabla f^0(z), H_\varepsilon(z)^{-1}\nabla f^0(z)\rangle > 0$, compute $\mu_\varepsilon(z)$ according to (6) (or (4)), with $h_\varepsilon(z) \triangleq -H_\varepsilon(z)^{-1}\nabla f^0(z)$, and go to Step 4. Otherwise, set $\varepsilon = \frac{\varepsilon}{2}$ and go to Step 3.

Step 4: If $f^0(z + \mu_\varepsilon(z)\, h_\varepsilon(z)) - f^0(z) \leq -\varepsilon$, set $z = z + \mu_\varepsilon(z)\, h_\varepsilon(z)$ and go to Step 1. If $f^0(z + \mu_\varepsilon(z)\, h_\varepsilon(z)) - f^0(z) > -\varepsilon$, set $\varepsilon = \frac{\varepsilon}{2}$ and go to Step 2.

We can again show by means of Theorem (I.3.21) that either the sequence $\{z_i\}$, of consecutive values assigned to z in Step 4, is finite, terminating at z_k and

$\nabla f^0(z_k) = 0$, or else $\{z_i\}$ is infinite and all the accumulation points z^* of $\{z_i\}$ satisfy $\nabla f^0(z^*) = 0$. Since a strictly convex function has a unique minimum, $\hat{z}$, we conclude that $z_i \to \hat{z}$.

We conclude this chapter with a very simple method for unconstrained minimization which does not require any derivative evaluations. It is particularly effective when the function $f^0(\cdot)$, which one wishes to minimize on $\mathbb{R}^n$, is of the form

8. $$f^0(z) = \sum_{i=1}^{n} f_i^0(z^i) \quad .$$

For $i = 1, 2, \ldots, n$, let ξ_i be the i^{th} column of the $n \times n$ unit matrix, i.e., the ξ_i are the usual co-ordinates for $\mathbb{R}^n$. Let $\nu_1, \nu_2, \ldots, \nu_{2n}$ in $\mathbb{R}^n$, be defined by, $\nu_1 = \xi_1$, $\nu_2 = -\xi_1$, $\nu_3 = \xi_2$, $\nu_4 = -\xi_2, \ldots, \nu_{2n-1} = \xi_n$, $\nu_{2n} = -\xi_n$. We can now state the method of local variations.

9. The Method of Local Variations:† Suppose a $z_0 \in \mathbb{R}^n$, and a $\rho_1 > 0$ are given.

Step 0: Set $z = z_0$, $\rho = \rho_1$.

Step 1: Set $i = 1$.

Step 2: Compute $f^0(z + \rho\nu_i)$.

Step 3: If $f^0(z + \rho\nu_i) < f^0(z)$, go to Step 4.

If $f^0(z + \rho\nu_i) \geq f^0(z)$ and $i < 2n$, set $i = i+1$ and go to Step 2.

If $f^0(z + \rho\nu_i) \geq f^0(z)$ and $i = 2n$, set $\rho = \rho/2$ and go to Step 1.

Step 4: Set $z = z + \rho\nu_i$ and go to Step 1.

10. Remark: It is clear that this procedure can be made to be somewhat more efficient by using past information in choosing the first $i \in \{1, 2, \ldots, 2n\}$ for the cyclic scan of the values $f^0(z + \rho\nu_i)$. The particulars of such modifications are best worked out with respect to the specific class of problems in which one is interested.

---

†This method seems to have been known for quite a while. It was described in [R1] and [C3].

The Algorithm (9) generates a sequence of points $\{z_j\}$ which lie in the bounded set $\{z \mid f^0(z) \leq f^0(z_0)\}$, and hence $\{z_j\}$ contains convergent subsequences. The point z_j is the j^{th} consecutive value assigned to z in Step (4) of (9).

11. Definition: We shall say that a subsequence $\{z_k\}$ of a sequence $\{z_j\}$ generated by the Algorithm (9) is ρ-stationary if

$$f^0(z_k + \rho_k v_i) \geq f^0(z_k), \quad i = 1, 2, \ldots, 2n ,$$

where ρ_k is the largest value of ρ used concurrently with z_k in Step 3 of (9). (i.e., $z_k = z_{k-1} + \rho_k v_i$, for some $i \in \{1, 2, \ldots, 2n\}$.)

12. Theorem: Let $\{z_j\}$ be an infinite sequence of points in $\mathbb{R}^n$ generated by the Algorithm (9). Then $\{z_j\}$ contains ρ-stationary subsequences, and each limit point of a ρ-stationary subsequence is a local minimum or saddle point of $f^0(\cdot)$.

Proof: Since the set $\{z \mid f^0(z) \leq f^0(z_0)\}$ is bounded, starting with z_0, it is possible to construct only a finite number of points z_i satisfying, for a fixed $\rho > 0$, $z_{i+1} = z_i + \rho v_j$, $j \in \{1, 2, \ldots, 2n\}$, and $f^0(z_{i+1}) < f^0(z_i)$. Hence, after a finite number of steps, the Algorithm (9) will construct a ρ-stationary point z_k. Pursuing this argument, we see easily that $\{z_i\}$ must contain ρ-stationary subsequences $\{z_k\}$, $k \in K \subset \{0, 1, 2, \ldots\}$, such that the associated values of ρ_k, i.e., the ρ, converge to zero as $k \to \infty$.

Now, let $\{z_k\}$, $k \in K \subset \{0, 1, 2, \ldots\}$ be any ρ-stationary, convergent subsequence constructed by (9) and let z^* denote its limit point. Then, we have, for each $k \in K$,

13. $$f^0(z_k + \rho_k \xi_j) \geq f^0(z_k), \qquad j = 1, 2, \ldots, n$$

14. $$f^0(z_k - \rho_k \xi_j) \geq f^0(z_k), \qquad j = 1, 2, \ldots, n$$

Applying the Taylor expansion to (13) and (14), we obtain,

15. $$f^0(z_k) + \rho_k \frac{\partial f^0(z_k + \lambda^j \rho_k \xi_j)}{\partial z^j} \geq f^0(z_k), \quad j = 1, 2, \ldots, n, \quad \lambda^j \in [0,1]$$

16. $$f^0(z_k) - \rho_k \frac{\partial f^0(z_k - \mu^j \rho_k \xi_j)}{\partial z^j} \geq f^0(z_k), \quad j = 1, 2, \ldots, n, \quad \mu^j \in [0,1] \ .$$

Hence, since $z_k \to z^*$, $\rho_k \to 0$ and since $f^0(\cdot)$ is continuously differentiable, we conclude that $\partial f^0(z^*)/\partial z^j = 0$ for $j = 1, 2, \ldots, n$, i.e., that $\nabla f^0(z^*) = 0$, so that z^* is a stationary point. The fact that it must be a point of local minimum or a saddle point now follows by inspection.

III. CONSTRAINED MINIMIZATION PROBLEMS

1. Penalty Function Methods

Penalty function methods for solving problems of the form $\min \{f^0(z) \mid z \in \Omega \in \mathbb{R}^n\}$ were first proposed by R. Courant in 1943 [C4]. The intuitive reasoning behind these methods is as follows. Suppose that we wish to minimize $f^0(z)$ subject to $r(z) = 0$, with $z \in \mathbb{R}^n$, and $f^0(\cdot)$ and $r(\cdot)$ continuously differentiable. Now consider the problem

$$1. \qquad \text{minimize } \gamma_i(z) \triangleq f^0(z) + \lambda_i \|r(z)\|^2, \qquad i = 0, 1, 2, \ldots$$

with $0 < \lambda_0 < \lambda_1 < \lambda_2, \ldots$ If $\lambda_i > 0$ is very large, the cost in not satisfying $r(z) = 0$ becomes very high in (1) and hence one may expect the solutions $\hat{z}_i$ of (1) to lie in, or close to, the set $\Omega \triangleq \{z \mid r(z) = 0\}$. We also note that $\gamma_i(z) = f^0(z)$ for all $z \in \Omega$, and hence $\gamma_i(\hat{z}_i) \leq \min \{f^0(z) \mid r(z) = 0\}$. If $\lambda_i > 0$ is allowed to grow infinity, one may therefore expect that the values $\gamma_i(\hat{z}_i)$ will grow monotonically to the optimal value, $\min \{f^0(z) \mid r(z) = 0\}$, and that the $\hat{z}_i$ will converge to a $\hat{z} \in \Omega$ which is optimal for (1).

There are two separate reasons for wishing to consider sequences of problems such as (1), rather than to solve (1) for a single preassigned value of $\lambda_i > 0$, to obtain an approximation to $\min \{f^0(z) \mid r(z) = 0\}$. The first is that one really does not know how to pick such a λ_i and hence one prefers to observe the growth of the values $\gamma_i(\hat{z}_i)$ and to stop when this growth becomes negligible. The second reason is that if one started with a rough guess z_0 and tried to minimize $\gamma_i(z)$ for $\lambda_i > 0$ very large, the term $\lambda_i \|r(z)\|^2$ would be extremely large in comparison with $f(z)$, thus swamping it. In addition, computer overflow would be likely to occur. Thus, one would tend to increase the λ_i gradually, using $\hat{z}_i$,

which minimizes $\gamma_i(z)$, as the starting point in minimizing $\gamma_{i+1}(z)$. We shall return to computational specifics of penalty function methods after establishing two of the better-known methods in this class. (The first is due to Zangwill [Z2], the second to Fiacco and McCormack [F1], [F2].)

Exterior Penalty Functions.

Suppose that we wish to solve the following problem

2. $$P: \quad \min \{f^0(z) \mid z \in \Omega\}$$

f^0 is a continuous function from R^n into R^1, and Ω is a nonempty, closed subset of R^n.

3. Definition: A sequence $\{p_i(\cdot)\}_{i=0}^{\infty}$, of continuous real-valued functions defined on R^n, is called a sequence of (exterior) penalty functions for the set Ω if for every $i = 0, 1, 2, \ldots,$

4. $$p_i(z) = 0 \quad \text{if and only if} \quad z \in \Omega$$

and

$$p_i(z) > 0 \quad \text{for every } z \notin \Omega$$

5. $$p_{i+1}(z) > p_i(z) \quad \text{for every} \quad z \notin \Omega$$

6. $$p_i(z) \to +\infty \text{ as } i \to +\infty, \text{ for every fixed } z \notin \Omega .$$

Now consider the sequence of problems

7. $$P_i: \quad \min \{f^0(z) + p_i(z) \mid z \in R^n\}, \quad i = 0, 1, 2, \ldots,$$

where the $p_i(\cdot)$ are exterior penalty functions for Ω.

Let

8. $$b = \min \{f^0(z) \mid z \in \Omega\}$$

9. $$b_i = \min \{f^0(z) + p_i(z) \mid z \in R^n\}, \quad i = 0, 1, 2, \ldots .$$

If we assume that b and the b_i exist and that they are finite, then we have the following result.

10. Lemma: The sequence $\{b_i\}_{i=0}^{\infty}$ satisfies $b_0 \leq b_1 \leq b_2 \leq \ldots \leq b_i \ldots \leq b$.

Proof: For $i = 0, 1, 2, \ldots$, let $z_i \in \mathbb{R}^n$ be such that

10a. $$b_i = f^0(z_i) + p_i(z_i)$$

Using (5), we obtain

10b. $$b_i \leq f^0(z_{i+1}) + p_i(z_{i+1}) \leq f^0(z_{i+1} + p_{i+1}(z_{i+1}) = b_{i+1} .$$

Now, using (4) and (10a) we get

10c. $$b_i \leq f^0(z) + p_i(z) = f^0(z) \text{ for all } z \subset \Omega,$$

i.e.,

$$b_i \leq \min \{f^0(z) \mid z \in \Omega\} = b ,$$

which completes our proof.

11. Lemma: Let $\{p_i(\cdot)\}_{i=0}^{\infty}$ be a sequence of penalty functions for the constraint set Ω, and let $\{z_i\}_{i=1}^{\infty}$ be a sequence in $\mathbb{R}^n$. If $\{z_i\}_{i=1}^{\infty}$ converges to a point z^*; $z_i \notin \Omega$ for $i = 1, 2, \ldots$, and the sequence $\{p_i(z_i)\}_{i=1}^{\infty}$ is bounded, then $z^* \in \Omega$.

Proof: We shall prove this lemma by contradiction. Suppose that z^* is not in Ω, and let $M > 0$, be the bound on $p_i(z_i)$, i.e., $0 < p_i(z_i) \leq M$ for $i = 1, 2, 3, \ldots$ Since $z^* \notin \Omega$, and, by (6), $p_i(z^*) \to +\infty$, there exists an integer ν' such that $p_{\nu'}(z^*) > 2M$. Now, since $p_{\nu'}(\cdot)$ is continuous, there exists a ball B with center

z^* such that for all $z \in B$

11a. $$p_{\kappa'}(z) \geq \frac{3M}{2}$$

Note that since $p_{\kappa'}(z) = 0$ for $z \in \Omega$, $B \cap \Omega = \phi$, the empty set. Now, $z_i \to z^*$ and hence there is an integer κ'' such that $z_i \in B$ for all $i \geq \kappa''$. Let $\kappa = \max\{\kappa', \kappa''\}$, then for all $i \geq \kappa$, $z_i \in B$ and, by (5) and (11a)

11b. $$p_i(z_i) \geq p_\kappa(z_i) \geq \frac{3M}{2} ,$$

which is a contradiction, since $p_i(z_i) \leq M$. Hence $z^* \in \Omega$.

12. Theorem: Suppose that for $i = 1, 2, \ldots$, the problem p_i defined in (7) has a solution z_i. Then any accumulation point of the sequence $\{z_i\}_{i=1}^{\infty}$ is optimal for the problem p defined in (2).†

Proof: Without any loss in generality, we may assume that $z_i \to z^*$. First, if for any integer j, $p_j(z_j) = 0$, then $z_j \in \Omega$, and z_j also solves the problem P (since $f^0(z) \equiv f^0(z) + p_j(z)$ for $z \in \Omega$). Consequently, by Lemma (10),

$b_i = b$ for every $i \geq j$ and hence $p_i(z_i) = 0$ for all $i \geq j$, since $p_i(z) > 0$ for all $z \notin \Omega$. Therefore for all $i \geq j$, $z_i \in \Omega$ and is also an optimal solution to problem P. Since Ω is closed, $z^* \in \Omega$, and, since $f^0(\cdot)$ is continuous, $f^0(z^*) = b$, i.e., z^* is an optimal solution to the problem P.

Now suppose that $z_i \notin \Omega$ for $i = 0, 1, 2, \ldots$ Since $z_i \to z^*$ and $f^0(\cdot)$ is continuous, $f^0(z_i) \to f^0(z^*)$, and hence there exists a positive number $M < \infty$, such

---

†If the set $\{z \mid f^0(z) + p_0(z) \leq b\}$ is bounded, then, since this set contains the entire sequence $\{z_i\}$, $\{z_i\}$ has accumulation points.

that

13. $$|f^0(z_i)| \leq M \quad \text{for} \quad i = 0, 1, 2, \ldots$$

Consequently, since $f^0(z_i) + p_i(z_i) \leq b \quad$ for $\quad i = 0, 1, 2, \ldots,$

14. $$p_i(z_i) \leq b + M$$

and therefore, the sequence $\{p_i(z_i)\}_{i=0}^{\infty}$ is bounded. It now follows from Lemma (11) that, $z^* \in \Omega$ and so, by the definition of b,

15. $$b \leq f^0(z^*)$$

But, for $i = 0, 1, 2, \ldots,$ $b \geq f^0(z_i) + p_i(z_i)$, and $p_i(z_i) > 0$. Hence $b \geq f(z^*)$ and therefore we must have $b = f(z^*)$. Since $z^* \in \Omega$ and $f(z^*) = b$, z^* is an optimal point for P.

We shall now give some examples of penalty functions which satisfy the properties stipulated in Definition (3).

16. Proposition: Let $f^i: \mathbb{R}^n \to \mathbb{R}^1$, $i = 1, 2, \ldots, m$, be continuous functions and let

17. $$\Omega \triangleq \{z | f^i(z) \leq 0, \ i = 1, 2, \ldots, m\} \ .$$

For each $i = 0, 1, 2, \ldots,$ let $p_i: \mathbb{R}^n \to \mathbb{R}^1$ be defined by

18. $$p_i(z) = \lambda_i \sum_{j=1}^{m} [\max \{f^j(z), 0\}]^{\alpha}$$

where λ_i and α are scalars satisfying $\lambda_i > 0$ and $\alpha \geq 1$. If $\lambda_{i+1} > \lambda_i$ for $i = 0, 1, 2, \ldots,$ and $\lambda_i \to +\infty$ as $i \to \infty$, then $\{p_i(\cdot)\}_{i=0}^{\infty}$ is a sequence of penalty functions for the set Ω.

Proof: First note that Ω is closed since the functions $f^i(\cdot)$, $i = 0, 1, 2, \ldots, m$, are continuous. Next, since $f^j(\cdot)$, $j = 1, 2, \ldots, m$, is continuous, $q^j(\cdot)$ is also continuous, where

18a. $$q^j(z) \triangleq [\max \{f^j(z), 0\}]$$

Finally, since $[q^j(\cdot)]^\alpha$ is continuous, $p_i(\cdot)$ is continuous.

Since $\lambda_i > 0$, $p_i(z) = 0$ if and only if $z \in \Omega$ and $p_i(z) > 0$ for all $z \notin \Omega$. Since $\lambda_{i+1} > \lambda_i$, $i = 0, 1, 2, \ldots$, $p_{i+1}(z) > p_i(z)$ for every $z \notin \Omega$. Since $\lambda_i \to +\infty$ as $i \to \infty$ for every $z \notin \Omega$. Therefore, by Definition (3), $\{p_i(\cdot)\}_{i=0}^{\infty}$ is a sequence of penalty functions for Ω.

18c. Proposition: For every $i = 0, 1, 2, \ldots$, the function $p_i(\cdot)$ defined in (18) is continuously differentiable on $\mathbb{R}^n$ if the functions $f^i(\cdot)$, $i = 1, 2, \ldots, k$, are continuously differentiable on $\mathbb{R}^n$ and $\alpha \geq 2$.

19. Proposition: Let $r: \mathbb{R}^n \to \mathbb{R}^m$ be a continuous function on $\mathbb{R}^n$ and let $\Omega = \{z: r(z) = 0\}$. For each $i = 0, 1, 2, \ldots$, let the map $p_i: \mathbb{R}^n \to \mathbb{R}^1$ be defined by

20. $$p_i(z) = \lambda_i \|r(z)\|^\alpha$$

where λ_i and α are scalars with $\lambda_i > 0$ and $\alpha \geq 1$. If $\lambda_{i+1} > \lambda_i$ for $i = 1, 2, \ldots$, and $\lambda_i \to +\infty$ as $i \to \infty$, then $\{p_i(\cdot)\}_{i=0}^{\infty}$ is a sequence of penalty functions for Ω. Also, $p_i(\cdot)$ is continuously differentiable on $\mathbb{R}^n$ if $r(\cdot)$ is continuously differentiable on $\mathbb{R}^n$ and $\alpha \geq 2$.

20a. Proposition: Suppose that the functions $f^i(\cdot)$, introduced in (16), are convex, and that the function $r(\cdot)$, defined in (19) is affine. Then, for $i = 1, 2, \ldots$, the functions $p_i(\cdot)$, defined in (18), and the functions $p_i(\cdot)$, defined in (20), are convex.

20b. Proposition: If $\{p_i'(\cdot)\}_{i=0}^{\infty}$ is a sequence of penalty functions for the set Ω_1, and $\{p_i''(\cdot)\}_{i=0}^{\infty}$ is a sequence of penalty functions for the set Ω_2, then $\{p_i'(\cdot)+p_i''(\cdot)\}_{i=0}^{\infty}$ is a sequence of penalty functions for $\Omega_1 \cap \Omega_2$. Also, $\min \{p_i'(\cdot), p_i''(\cdot)\}_{i=0}^{\infty}$ is a sequence of penalty functions for $\Omega_1 \cup \Omega_2$.

21. Remark: Exterior penalty functions can be used not only to transform a constrained optimization problem into a sequence of unconstrained minimization problems, but also into a more tractable sequence of constrained minimization problems. For example, suppose that we wish to minimize $f^0(z)$ subject to $r(z) = 0$, $f(z) \leqq 0$ ($r\colon \mathbb{R}^n \to \mathbb{R}^\ell$, $f\colon \mathbb{R}^n \to \mathbb{R}^m$) and the function $r(\cdot)$ is not affine. Then we cannot use any of the methods of feasible directions to be described later. Now, suppose that $\{p_i(\cdot)\}_{i=0}^{\infty}$ is a sequence of penalty functions for the set $\{z: r(z) = 0\}$, then, under suitable assumptions, we can use a feasible directions method to solve the sequence of problems: $\min \{f^0(z) + p_i(z) \mid f(z) \leqq 0\}$ to obtain a solution of the original problem.

22. Proposition: Consider the problem $\min \{f^0(z) \mid r(z) = 0, f(z) \leqq 0\}$ where $f^0\colon \mathbb{R}^n \to \mathbb{R}^1$, $r\colon \mathbb{R}^n \to \mathbb{R}^m$, $f\colon \mathbb{R}^n \to \mathbb{R}^\ell$. Let $\{p_i(\cdot)\}_{i=0}^{\infty}$ be any sequence of penalty functions for the set $\{z: r(z) = 0\}$, satisfying the Definition (3) and let z_i be optimal for the problem: $\min \{f^0(z) + p_i(z) \mid f(z) \leqq 0\}$. Then any accumulation point z^* of $\{z_i\}_{i=0}^{\infty}$ is optimal for the original problem.

Interior Penalty Functions.

We shall now consider a different type of penalty functions for solving the problem:

23. $$P: \quad \min \{f^0(z) \mid z \in \Omega\}$$

where f^0: $\mathbb{R}^n \to \mathbb{R}^1$ is a continuous function and Ω is a nonempty subset of $\mathbb{R}^n$.

We shall assume that $\overline{\overset{o}{\Omega}} = \Omega \neq \emptyset$, i.e., that Ω equals the closure of its interior, and that for every α real, the set $\{z: f^0(z) \leq \alpha\}$ is bounded. We now define a sequence of penalty functions for Ω.

24. Definition: A sequence $\{p_i(\cdot)\}_{i=0}^{\infty}$ of continuous real-valued functions defined on $\overset{o}{\Omega}$ (the interior of Ω) is called a sequence of interior penalty functions for Ω, if for $i = 0, 1, 2, \ldots,$

25. $$0 < p_{i+1}(z) < p_i(z) \qquad \text{for all} \quad z \in \overset{o}{\Omega},$$

26. $$p_i(z) \to 0 \quad \text{as} \quad i \to \infty$$

27. $$\text{if } z_j \in \overset{o}{\Omega} \text{ for } j = 0, 1, 2, \; z_j \to z^* \in \partial\Omega, \text{ then } / \; p_i(z_j) \to +\infty \text{ as } j \to \infty.$$

Now consider the problems P_i defined below

28. $$P_i: \quad \min \{f^0(z) + p_i(z) \mid z \in \overset{o}{\Omega}\}, \quad i = 0, 1, 2, \ldots$$

29. Theorem: Let z_i be optimal for P_i, $i = 0, 1, 2, \ldots,$ defined in (28)†. Then every accumulation point of the sequence $\{z_i\}_{i=0}^{\infty}$ is optimal for the problem P defined in (23).

Proof: If for $i = 0, 1, 2, \ldots,$ we let

30. $$b_i \triangleq \min \{f^0(z) + p_i(z) \mid z \in \overset{o}{\Omega}\}$$

---

†Let $z_0 \subset \overset{o}{\Omega}$. Then $\tilde{\Omega}_i \triangleq \{z \in \overset{o}{\Omega} \mid f^0(z) + p_i(z) \leq \alpha_0 \triangleq f^0(z_0) + p_0(z_0)\} \subset \{z \mid f^0(z) \leq \alpha_0\}$. Hence $\tilde{\Omega}_i$ is a compact subset of $\overset{o}{\Omega}$, and $\inf \{f^0(z) + p_i(z) \mid z \in \overset{o}{\Omega}\} = \inf \{f^0(z) + p_i(z) \mid z \in \tilde{\Omega}\} = \min \{f^0(z) + p_i(z) \mid z \in \tilde{\Omega}\}$, since $f^0(\cdot)$ and $p_i(\cdot)$ are continuous.

and set

30a. $$b = \min \{f^0(z) \mid z \subset \Omega\}$$

then we have

31. $$b_0 \geq b_1 \geq \cdots \geq b_i \geq b_{i+1} \cdots \geq b \quad .$$

Since the b_i form a bounded, monotonically decreasing sequence, they must converge, i.e., $b_i \to b^*$. Now, suppose that $b^* \neq b$, then from (31), $b^* > b$. Since $f^0(\cdot)$ is continuous, there exists a ball B, with center $\hat{z}$, the optimal point for P, such that for all $z \in B$

32. $$f^0(z) \leq b^* - \frac{1}{2}(b^* - b) \quad .$$

Now take any $z' \in B \cap \overset{o}{\Omega}$, then, since $p_i(z) \to 0$ as $i \to \infty$, there exists an integer k such that for all $i \geq k$,

33. $$p_i(z') < \frac{1}{4}(b^* - b)$$

and hence, for all $i \geq k$

34. $$b_i = f^0(z_i) + p_i(z_i) \leq f^0(z') + p_i(z') < b^* - \frac{1}{4}(b^* - b) \ ,$$

which contradicts our assumption that $b_i \to b^*$. Therefore, $b^* = b$.
Now for $i = 0, 1, 2, \ldots$ $f^0(z_i) + p_i(z_i) \leq f^0(z_0) + p_0(z_0)$, and hence $z_i \in \{z \mid f^0(z) \leq f^0(z_0) + p_0(z_0)\}$ which is bounded by assumption.
Let $\{z_j\}$, $j \in K \subset \{0, 1, 2, \ldots\}$ be any convergent subsequence of $\{z_i\}_{i=0}^{\infty}$, with limit point $z^* \in \Omega$ and suppose that z^* is not an optimal point for P. Then

$f(z^*) > b$, and the sequence $\{(f(z_j) - b) + p_j(z_j)\}$, $j \in K$, cannot converge to zero, which contradicts the fact that $(b_i - b) \to 0$. Thus, all accumulation points of $\{z_i\}$ are optimal for P. This completes the proof

of the theorem. (Note also, that $p_j(z_j) \to 0$, since $f(z_j) - b + p_j(z_j) \to 0$ and $z_j \to \hat{z}$.)

35. Remark: To utilize penalty functions of the above type, we must have an initial feasible solution in the interior of Ω as a starting point for the unconstrained optimization algorithm to be used for the solution of the P_i. Since $f_i(z) + p_i(z) \to +\infty$ as z approaches the boundary of Ω, the unconstrained optimization algorithm will then generate a sequence of points z_j, $j = 1, 2, \ldots,$ which will all be in the interior of Ω.

36. Remark: Suppose that $\Omega = \{z \mid f^i(z) \leq 0, i = 1, 2, \ldots, m\}$, where the $f^i(\cdot)$ are continuous functions such that

(i) $f^i(z) \neq 0$ for all $z \in \overset{\circ}{\Omega}$,

(ii) the set $\{z \mid f^i(z) = 0, i = 1, 2, \ldots, m\}$ is contained in the closure of $\overset{\circ}{\Omega}$,

(iii) for $i = 0, 1, 2, \ldots, \lambda_i > \lambda_{i+1} > 0$, and $\lambda_i \to 0$ as $i \to \infty$, then

$p_i(z) = -\lambda_i \sum_{i=1}^{m} \frac{1}{f^i(z)}$ are penalty functions for Ω.

37. Remark: Suppose that we wish to minimize $f^0(z)$ subject to $z \in \bigcap_{i \in I} \Omega_i$ and suppose that we have a $z_0 \in \bigcap_{i \in J} \overset{o}{\Omega}_i$, $J \subset I$. Then we may use interior penalty functions for the sets Ω_i, $i \in J$, and exterior penalty functions for the remaining Ω_i. (See [3]).

Computational Aspects†

The use of penalty functions requires us to minimize a sequence of functions of the form $f^0(z) + p_i(z)$, $z \in \mathbb{R}^n$. However, as we have seen in the preceding sections, the various unconstrained minimization methods which are available to us, usually compute only local minima, and, in addition, take an infinite number of steps to compute these local minima. If we insisted on using penalty function methods in a literal sense, therefore, we could not even get past minimizing $f^0(z) + p_0(z)$, $z \in \mathbb{R}^n$, in finite time. One must therefore use truncation procedures in approximating the minima of $f^0(z) + p_i(z)$. Also, if one uses a method on $f^0(z) + p_i(z)$, which can only compute points $\hat{z}_i$ such that $\nabla f^0(\hat{z}_i) = 0$, then one must also wonder as to the nature of the accumulation points of the z sequence of $\hat{z}_i$. We shall now propose a truncation procedure for use with penalty functions and shall establish its properties for a few special cases.

Similar procedures can also be developed for use with interior penalty functions.

Thus, let us consider again our original problem,

38. $$\min \{ f^0(z) \mid z \in \Omega \subset \mathbb{R}^n \}$$

---

†The remaining results in this section do not appear to have been published before.

Let $\frac{1}{\epsilon_i} p(\cdot)$, $i = 0, 1, 2, \ldots,$ be a sequence of penalty functions (inner or outer) for Ω which we assume to be closed. We suppose that $f^0(\cdot)$ and $p(\cdot)$ are continuously differentiable and that $\epsilon_i = \frac{\epsilon}{2^i}$, $\epsilon > 0$. We now state a "first order" type algorithm for "solving" (38).

39. Algorithm: Choose $\epsilon > 0$, $z \in \mathbb{R}^n$.

Step 1: Compute

40. $$h_\epsilon(z) \triangleq - [\nabla f^0(z) + \frac{1}{\epsilon} \nabla p(z)]$$

Step 2: If $\|h_\epsilon(z)\| > \epsilon$, go to Step 3.

If $\|h_\epsilon(z)\| \leq \epsilon$, set $\epsilon = \frac{\epsilon}{2}$ and go to Step 1.

Step 3: Compute $\mu(z) \geq 0$ to be the smallest possible number such that

41. $$f^0(z + \mu(z)h_\epsilon(z)) + \frac{1}{\epsilon} p(z + \mu(z)h_\epsilon(z)) \leq f^0(z + \mu h_\epsilon(z)) + \frac{1}{\epsilon} p(z + \mu h_\epsilon(z)) \text{ for all } \mu \geq 0.$$

(or else use the method for choosing $\mu(z)$ described in (II.1.18) or (II.1.32), i.e., Set $\mu(z) = \lambda_i$).

Step 4: Set $z = z + \mu(z)h_\epsilon(z)$ and go to Step 1.

We shall now show that in a number of important cases, this algorithm will compute points $\hat{z}$ which satisfy necessary conditions of optimality for (38).

Case 1: Suppose that $\Omega = \{z \mid r(z) = 0\}$, where $r: \mathbb{R}^n \to \mathbb{R}^\ell$ is continuously differentiable and the Jacobian matrix $\frac{\partial r(z)}{\partial z}$ has maximum rank for all $z \in \mathbb{R}^n$.†

Consider the sequence of points z_i constructed by the algorithm (39), starting at an initial point z_0 with $p(z) = \frac{1}{2}\|r(z)\|^2$. Within this sequence, we

---

†Actually, it is enough to make the weaker assumption that $\frac{\partial r(z)}{\partial z}$ has maximum rank in an open set containing the sequence z_i which the algorithm (39) generates.

single out a subsequence of points z_j at which the algorithm reduced, in Step 2, the current value of $\epsilon = \epsilon_j$ to a new (and smaller) value ϵ_{j+1}. For this subsequence $\{z_j\}$ assuming that it is infinite, we find that

42. $$\|h_{\epsilon_j}(z_j)\| \leq \epsilon_j$$

and $\epsilon_j \to 0$ as $j \to \infty$, i.e., we find that

43. $$\|h_{\epsilon_j}(z_j)\| \to 0 .$$

Now, with $p(z) = \frac{1}{2}\|r(z)\|^2$,

44. $$h_{\epsilon_j}(z_j) = -\left[\nabla f^0(z_j) + \frac{1}{\epsilon_j}\left(\frac{\partial r(z_j)}{\partial z}\right)^T r(z_j)\right]$$

Suppose now that $z_j \to z^*$, then, since $\frac{\partial r(z)}{\partial z}$ has maximum rank for all $z \in \mathbb{R}^n$, we conclude that $r(z^*) = 0$, i.e., $z^* \in \Omega$. Next,

45. $$\frac{1}{\epsilon_j} r(z_j) = -\left(\frac{\partial r(z_j)}{\partial z}\left(\frac{\partial r(z_j)}{\partial z}\right)^T\right)^{-1} \frac{\partial r(z_j)}{\partial z}\left[h_{\epsilon_j} + \nabla f^0(z_j)\right]$$

We therefore conclude from the fact that $h_{\epsilon_j}(z_j) \to 0$ and from the continuity of $\frac{\partial r(z)}{\partial z}$ and $\nabla f^0(z)$, that

46. $$\lim_{\epsilon_j \to 0} \frac{1}{\epsilon_j} r(z_j) = -\left(\frac{\partial r(z^*)}{\partial z}\left(\frac{\partial r(z^*)}{\partial z}\right)^T\right)^{-1} \frac{\partial r(z^*)}{\partial z} \nabla f^0(z^*) \triangleq \psi$$

Thus, in the limit, (44) gives

47. $$\nabla f^0(z^*) + \left(\frac{\partial r(z^*)}{\partial z}\right)^T \psi = 0 ,$$

i.e., $z^* \in \Omega$ and satisfies the necessary condition of optimality (I.2.1). Therefore, if the sequence $\{z_i\}$ constructed by the Algorithm (39) for Case 1 remains in a bound set, it will have subsequences of points $\{z_j\}$ which converge to points

$z^* \in \Omega$ satisfying (47). This will always be the case if the set $\Omega = \{z \mid r(z) = 0\}$ itself is bounded. Generally, in order to ensure that the sequence $\{z_i\}$ stays bounded requires some additional assumptions on $f^0(\cdot)$ and $r(\cdot)$.

Case 2: Suppose that $\Omega = \{z \mid f(z) \leq 0\}$ where $f: \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable, and the Jacobian matrix $\frac{\partial f(z)}{\partial z}$ has maximum rank for all $z \in \mathbb{R}^n$.

Let us set $p(z) = \frac{1}{2} \sum_{i=1}^{m} (\max \{0, f^i(z)\})^2$, or, setting $J(z) = \{j \mid f^j(z) \geq 0, j \in \{1, 2, \ldots, m\}\}$, we may write $p(z) = \frac{1}{2} \sum_{i \in J(z)} (f^i(z))^2$. Now, for the algorithm (39)

48. $$h_{\epsilon_j}(z_j) = -\nabla f^0(z_j) + \sum_{i \in J(z_j)} \frac{f^i(z_i)}{\epsilon_j} \nabla f^i(z_j)$$

Again consider the subsequence $\{z_j\}$, $j \in K \subset \{0, 1, 2, \ldots\}$, of $\{z_i\}$ at which ϵ_j is reduced to the new value ϵ_{j+1}. Then

49. $$h_{\epsilon_j}(z_j) \to 0, \quad j \in K.$$

Suppose now that $z_j \to z^*$, $j \in K$, then from (48) and (49) it follows that $f^i(z^*) \leq 0$ for all $i \in \{1, 2, \ldots, m\}$ since otherwise $h_{\epsilon_j}(z_j) \to 0$, $j \in K$, is impossible. Also, since the $\nabla f^i(z)$ are linearly independent for all $z \in \mathbb{R}^n$ by assumption, we must have

50. $$\nabla f^0(z^*) + \sum_{i \in J(z^*)} \mu^i \nabla f^i(z^*) = 0$$

where $\mu^i = \lim_{\substack{j \to \infty \\ i \in J(z^*)}} \frac{f^i(z_j)}{\epsilon_j}$, $j \in K$, exists and satisfies $\mu^i \geq 0$. By inspection of (50),

we see that the point $z^* \in \Omega$ satisfies the necessary conditions of optimality (I.2.1).

Exercise: Find conditions which ensure that the Algorithm (39) will construct compact sequences only.

Exercise: Under what conditions could the modified Newton-Raphson method be utilized in an algorithm of the type (39)?

2. A Method of Centers

The method of centers, which was introduced by Huard [H3], [B5] bridges the gap between the penalty function methods presented in the preceeding section and the methods of feasible directions to be presented in the next section. Depending on one's point of view, the particular version of the method of centers which we are about to present can be considered to be either a parameter free, interior penalty function method, or else a parameter free feasible directions method.

Consider again the problem

$$1. \qquad \min \{f^0(z) \mid f^i(z) \leq 0, \quad i = 1, 2, \ldots, m\}$$

where $f^i: \mathbb{R}^n \to \mathbb{R}^1$, $i = 0, 1, \ldots, m$ are continuously differentiable functions. Suppose that we have a $z_0 \in \Omega \triangleq \{z \mid f^i(z) \leq 0, i = 1, 2, \ldots, m\}$ and that the set $\tilde{\Omega}(z_0) \triangleq \{z \mid f^0(z) - f^0(z_0) \leq 0; f^i(z) \leq 0, i = 1, 2, \ldots, m\}$ is compact and has an interior. The gyst of the method of centers is to pick z_1, the successor of z_0, to be a point well in the interior of $\tilde{\Omega}(z_0)$ (i.e., in the "center" of $\tilde{\Omega}(z_0)$) and then repeat the construction. When the "centering" of z_1 is defined in terms of giving a minimum to a suitably defined distance function, convergence to a stationary point can be established [B5].

Here we present a version of the method of centers which is considered to be the most successful one so far, and which can be established by means of the convergence theory presented in Theorem (I.3.1). Unfortunately, the original heuristic ideas involved in the methods of centers will be lost in the process.

First, note that if $\hat{z}$ is optimal for (1), then, by a trivial extension of Corollary (I.2.6),

$$2a. \qquad \min_{h \in S} (\max \{\langle \nabla f^0(\hat{z}), h\rangle, f^i(\hat{z}) + \langle \nabla f^i(\hat{z}), h\rangle, \quad i = 1, 2, \ldots, m\}) = 0,$$

where S is any set containing the origin in its interior.

Let $S = \{h \in \mathbb{R}^n \mid |h^i| \leq 1\}$ and let $\varphi\colon \mathbb{R}^n \to \mathbb{R}^1$ be defined by

2b. $$\varphi(z) = \min_{h \in S} (\max \{\langle \nabla f^0(z), h\rangle,\ f^i(z) + \langle \nabla f^i(z), h\rangle,\ i = 1, 2, \ldots, m\})$$

Thus, if z^* is optimal, then $\varphi(z^*) = 0$.

3. Remark: Note that $\varphi(z)$ can be calculated by solving $\min \sigma$ subject to

4. $$\begin{cases} \sigma - \langle \nabla f^0(z), h\rangle \geq 0\ , \\ \sigma - f^i(z) - \langle \nabla f^i(z), h\rangle \geq 0,\ i = 1, 2, \ldots, m \\ |h^i| \leq 1\ , \end{cases}$$

which is a linear programming problem. Let $(\sigma(z), h(z))$ denote a solution of (4), then $\varphi(z) = \sigma(z)$. Whenever $\sigma(z) = 0$, and the optimal h is not unique, we shall set $h(z) = 0$.

Also note that $\varphi(\cdot)$ is continuous because both $f^i(\cdot)$ and $\nabla f^i(\cdot)$ are continuous by assumption, for $i = 0, 1, 2, \ldots, m$.

5. Algorithm: Suppose we are given a $z_0 \in \Omega$.†

Step 0: Set $z = z_0$

Step 1: Solve (4) to obtain a vector $h(z)$. If $h(z) = 0$, stop, otherwise go to Step 2.

Step 2: Compute $\mu(z)$ to be the smallest positive scalar such that

6. $$d(z + \mu(z)\ h(z),\ z) = \min_{\mu \geq 0} d(z + \mu h(z),\ z)\ ,$$

where

7. $$d(z', z) \triangleq \max \{f^0(z') - f^0(z),\ f^i(z'),\ i = 1, 2, \ldots, m\}\ .$$

---

† To compute a $z_0 \in \Omega$ we apply Algorithm (5) to the problem $\min\{\sigma \mid \sigma - f^i(z) \geq 0,\ i = 1, 2, \ldots, m\}$ for which we construct an initial feasible solution $(\sigma_0, \tilde{z}_0)$ by taking $\tilde{z}_0$ arbitrary and $\tilde{\sigma}_0 = \max\{f^i(\tilde{z}_0) \mid i \in \{1, 2, \ldots, m\}\}$. Since Ω has an interior, there will be a finite integer k such that $\sigma_k < 0$ and $z_k \in \Omega$, where, for $j = 1, 2, \ldots$ (σ_j, z_j) are the successive pairs constructed by (5).

Step 3: Set $z = z + \mu(z)\, h(z)$ and go to Step 1.

Note that the function $d(\cdot,\cdot)$ acts as a "distance" function and by minimizing it we choose a point on the ray $\{z' \mid z' = z + \mu h(z),\ \mu \geq 0\}$ which is "well centered" in the set $\tilde{\Omega}(z)$.

Theorem: Let $z_0, z_1, z_2, \ldots$ be a sequence generated by the algorithm (5), i.e., for $i = 1, 2, \ldots$ z_i is the i^{th} value assigned to z in Step 3. Then either the sequence is finite, ending at z_k and $\varphi(z_k) = 0$, or else $\{z_i\}$ is infinite and every accumulation point z^* of $\{z_i\}$ satisfies $\varphi(z^*) = 0$.

Proof: It is easy to see that for any $z \in \Omega$, if $h(z)$ as computed in (4) is not zero, then $\varphi(z) < 0$ and there is a $z' = z + \mu h(z)$ in the interior of $\tilde{\Omega}(z)$ such that $d(z',z) < 0$. Hence, the sequence construction can only stop in Step 1, and this can happen at $z_0 = z_k$ if and only if $\varphi(z_k) = 0$. Thus, the case of $\{z_i\}$ finite is trivial.

Now let us consider the case when $\{z_i\}$ is infinite. To prove this part, we shall show that the assumptions of Theorem (I.3.1) are satisfied with $T \triangleq \Omega$, $c(\cdot) \triangleq - f^0(\cdot)$, $a(\cdot)$ defined by the Algorithm (5), and $z \in \Omega$ defined to be desirable if and only if $\varphi(z) = 0$. In fact, given a $z^* \in \Omega$ such that $\varphi(z^*) < 0$, we only need to show that there exist an $\epsilon^* > 0$ and a $\delta^* > 0$ such that for all $z \in \Omega$ satisfying $\|z - z^*\| \leq \epsilon^*$,

8. $$- f^0(a(z)) + f^0(z) \geq \delta^* \quad .$$

Let $\varphi^* = \varphi(z^*) < 0$ and let $h(z^*)$ be computed as in (4), with $z = z^*$. Since $\varphi(\cdot)$ is continuous, there exists an $\epsilon^* > 0$ such that

9. $$\varphi(z) \leq \varphi^*/2 \quad \text{for all} \quad z \in B(z^*, 2\epsilon^*),$$

where $B(z^*, 2\epsilon^*) \triangleq \Omega \cap \{z \mid \|z - z^*\| \leq 2\epsilon^*\}$. Now, for all $z \in B(z^*, \epsilon^*)$, and $\lambda \geq 0$,

10. $$f^0(z + \lambda h(z)) = f^0(z) + \lambda \langle \nabla f^0(z + \xi h(z)),\ h(z) \rangle$$

where $\xi \in [0,\lambda]$.

Since $\langle \nabla f^0(z), h(z) \rangle \leq \varphi^*/2$ for all $z \in B(z^*, 2\epsilon^*)$ and since $\langle \nabla f^0(\cdot), \cdot \rangle$ is uniformly continuous on $B(z^*, 2\epsilon^*) \times S$, and S is compact, there exists a $\lambda^0 > 0$ such that (see I.4.7)

11. $$f^0(z + \lambda h(z)) - f^0(z) \leq \lambda\varphi^*/4$$

for all $z \in B(z^*, \epsilon^*)$ and for all $\lambda \in [0,\lambda^0]$. Now, since $\varphi(z) \leq \varphi^*/2$, we must have, for all $z \in B(z^*, 2\epsilon^*)$ and $i = 1, 2, \ldots, m$,

12. $$f^i(z) + \langle \nabla f^i(z), h(z) \rangle \leq \varphi^*/2 \quad .$$

Also, since the $f^i(\cdot)$, $i = 1, 2, \ldots, m$, are uniformly continuous on $B(z^*, 2\epsilon^*)$ and S is compact, there exists a $\lambda^1 > 0$ such that (see I.4.1)

13. $$|f^i(z + \lambda h(z)) - f^i(z)| \leq |\varphi^*/8|, \; i = 1, 2, \ldots, m,$$

for all $\lambda \in [0,\lambda^1]$ and $z \in B(z^*, \epsilon^*)$. Since the $\langle \nabla f^i(\cdot), \cdot \rangle$ are uniformly continuous on $B(z^*, 2\epsilon^*) \times S$ there exists a $\lambda^2 > 0$ such that (see I.4.7)

14. $$|\langle \nabla f^i(z + \lambda h(z)), h(z) \rangle - \langle \nabla f^i(z), h(z) \rangle| \leq |\varphi^*/8|, \; i = 1, 2, \ldots, m,$$

for all $z \in B(z^*, \epsilon^*)$ and for all $\lambda \in [0,\lambda^2]$. Finally,

15. $$f^i(z + \lambda h(z)) = f^i(z) + \lambda \langle \nabla f^i(z + \zeta h(z), h(z) \rangle, \; i = 1, 2, \ldots, m, \; \zeta \subset [0,\lambda].$$

Now let $\lambda_m = \min \{\lambda^0, \lambda^1, \lambda^2\}$. Then, for any vector $z + \lambda_m h(z)$, $z \subset B(z^*, \epsilon^*)$, we have,

16. $$f^0(z + \lambda_m h(z)) - f^0(z) \leq \lambda_m \varphi^*/8$$

and

17. $$f^i(z + \lambda_m h(z)) \leq \max \{\varphi^*/8, \lambda_m \varphi^*/8\}, \; i = 1, 2, \ldots, m,$$

either because $f^i(z) \leq \varphi^*/4$ and because of (13); or because $f^i(z) > \varphi^*/4$, and then by reason of (12), (14) and (15). Consequently, for all $z \in B(z^*, \varepsilon^*)$,

18. $$f^0(a(z)) - f^0(z) \leq d(a(z),z) \leq \max\{\frac{\lambda_m \varphi^*}{8}, \frac{\varphi^*}{8}\} < 0$$

We now set $\delta^* = -\max\{\lambda_m \frac{\varphi^*}{8}, \frac{\varphi^*}{8}\}$ and the proof is completed.

19. Remark: When $\mu(z)$ is chosen so as to minimize $f^0(z + \mu h(z)) - f^0(z)$ subject to $\mu \geq 0$ and $(z + \mu h(z)) \in \Omega$, the algorithm convergence properties remain the same. It was stated in this form by Topkis and Veinott [T1] and is then a "feasible directions" algorithm of the type to be discussed in the next section.

The application of this algorithm to optimal control problems is essentially the same as of the methods of feasible directions and will be discussed towards the end of the next section.

3. Methods of Feasible Directions

In the particular version of the method of centers presented in the preceeding section, we had to solve the linear programming problem (2.4) in order to find a half line, $\{z' \mid z' = z + \mu h(z),\ \mu > 0\}$, on which the next point was going to lie. This linear programming problem always had (m+1) linear inequality constraints in addition to the constraints $|h^i| \leq 1$. We shall now consider a class of methods which were first introduced by Zontendijk [Z4], as well as some new modifications. The major difference between these methods and the method of centers, presented in the preceeding section, lies in fact that only a small part of the constraints used in (2.4) are now required for solving a problem of the form (2.4) to find a half line on which the next point will lie. However, since only part of the constraints are used in the computation of this half line, it becomes necessary to use "an ϵ-proceedure" so as to keep tab on the constraints which are not included. This ϵ-procedure is known by the names of "antizigzagging precaution" or "antijamming precaution".

Consider again the problem

1. $$\min \{f^0(z) \mid f^i(z) \leq 0,\ \ i = 1, 2, \ldots, m\}$$

where for $i = 0, 1, \ldots, m$ f^i: $\mathbb{R}^n \to \mathbb{R}^1$ is continuously differentiable. We shall assume that (1) has a solution. As we have already indicated, this can be ensured by requiring that for every $\alpha \in \mathbb{R}^1$, the set $\{z \mid f^0(z) \leq \alpha\}$ be compact, or else by requiring that the set $\Omega \triangleq \{z \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$ be compact, or that the set $\{z \mid f^0(z) - f^0(z_0) \leq 0,\ f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$ is compact for a given $z_0 \in \Omega$ which is then used as a starting point.

We recall that in (I.2.8), the set $J_\epsilon(z)$ was defined for any $\epsilon \geq 0$ and $z \in \Omega$, by

2. $$J_\varepsilon(z) = \{0\} \bigcup \{i \mid f^i(z) + \varepsilon \geq 0,\ i \in \{1, 2, \ldots, m\}\} ,$$

and that by (I.2.6), if $\hat{z}$ is optimal for (1), then

3. $$\min_{h \in S} \max_{i \in J_0(\hat{z})} \langle \nabla f^i(\hat{z}), h \rangle = 0 ,$$

where S is any subset of $\mathbb{R}^n$ containing the origin in its interior.

4. Definition: Let S be some compact subset of $\mathbb{R}^n$ containing the origin in its interior. For $\varepsilon \geq 0$, we define $\varphi_\varepsilon : \Omega \to \mathbb{R}^1$ as

$$\bar{\varphi}_\varepsilon(z) = \min_{h \in S} \max_{i \in J_\varepsilon(z)} \langle \nabla f^i(z), h \rangle$$

5. Remark: It is not difficult to see that $J_\varepsilon(z)$ and $\bar{\varphi}_\varepsilon(z)$ have the following properties: Suppose $z \in \Omega$ is given. Then,

5a. For any $\varepsilon > \varepsilon'$, $J_\varepsilon(z) \supset J_{\varepsilon'}(z)$ and hence $\bar{\varphi}_\varepsilon(z) \geq \bar{\varphi}_{\varepsilon'}(z)$;

5b. For any $\varepsilon > 0$, there exists a $\rho > 0$ such that $J_{\varepsilon+\rho}(z) = J_\varepsilon(z)$;

5c. For any $\varepsilon > 0$, there exists a $\rho > 0$ such that $J_\varepsilon(z') \subset J_\varepsilon(z)$ for all $z' \in B(z,\rho) \triangleq \{z' \in \Omega \mid \|z'-z\| \leq \rho\}$.

To compute $\bar{\varphi}_\varepsilon(z)$ we solve the problem

6. $$\min \{\sigma \mid \sigma - \langle \nabla f^i(z), h \rangle \geq 0 \quad \text{for} \quad i \in J_\varepsilon(z);\ h \in S\} .$$

The optimal pair $\sigma_\varepsilon(z)$, $h_\varepsilon(z)$ for (6) satisfies $\bar{\varphi}_\varepsilon(z) = \sigma_\varepsilon(z) = \max_{i \in J_\varepsilon(z)} \langle \nabla f^i(z), h_\varepsilon(z) \rangle$. In solving (6), we shall always set $h_\varepsilon(z) = 0$ whenever $\sigma_\varepsilon(z) = 0$. Note that a sensible choice for S would be $S = \{h \mid |h^i| \leq 1\}$, or $S = \{h \mid \|h\| \leq 1\}$.

The algorithms we are about to present in the form of an idealized computer program will find points $\hat{z} \in \Omega$ such that $\bar{\varphi}_0(z) = 0$. Note that these algorithms are parameterized by the particular choice for the set S, i.e., for each choice of S we get a different algorithm.

The procedure (7), below, is a minor variation of a method given by Zontendijk [Z4].

7. Algorithm: Suppose that a $z_0 \in \Omega$[†] and an $\bar{\epsilon}_0 > \epsilon' > 0$ are given.

(See next page)

[†]To find a $z_0 \in \Omega$, solve, using the Algorithm (7), the problem $\min \{\sigma \mid f^i(z) - \sigma \leq 0,\ i = 1, 2, \ldots, m\}$, with initial feasible point (z', σ') where z' is arbitrary and $\sigma' = \max \{f^i(z') \mid i = 1, 2, \ldots, m\}$. Since the optimal value $\hat{\sigma}$ for this problem satisfies $\hat{\sigma} < 0$,

Step 0: Set $z = z_0$

Step 1: Set $\epsilon(z) = \bar{\epsilon}_0$ (We shall use the abbreviated notation $\epsilon = \epsilon(z)$.)

Step 2: Compute $\bar{\varphi}_\epsilon(z)$ and $h_\epsilon(z)$ by solving (6)

Step 3: If $\bar{\varphi}_\epsilon(z) \leq -\epsilon$, set $h(z) = h_\epsilon(z)$ and go to Step 4.

If $\bar{\varphi}_\epsilon(z) > -\epsilon$ and $\epsilon \leq \epsilon'$, compute $\bar{\varphi}_0(z)$.

If $\bar{\varphi}_0(z) = 0$, set $z = z$ and Stop.

If $\bar{\varphi}_0(z) < 0$, set $\epsilon = \epsilon/2$ and go to Step 2.

If $\bar{\varphi}_\epsilon(z) > -\epsilon$ and $\epsilon > \epsilon'$, set $\epsilon = \epsilon/2$ and go to Step 2.

Step 4: Compute $\lambda(z) \geq 0$ such that

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

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

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

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

10. Theorem: Let $z_0, z_1, z_2, \ldots$, be a sequence in Ω constructed by the algorithm (7), i.e., $z_1, z_2, \ldots$, are the consecutive values assigned to z in Step 3 or Step 6. Then, either the sequence $\{z_i\}$ is finite and its last element, say z_k, satisfies $\bar{\varphi}_0(z_k) = 0$ or else $\{z_i\}$ is infinite and every accumulation point $\hat{z}$ in $\{z_i\}$ satisfies $\bar{\varphi}_0(\hat{z}) = 0$.

Proof: Obviously, the algorithm (7) defines a map $a: \Omega \to \Omega$. We shall show that this map together with the map $-f^0(\cdot)$ ($-f^0(\cdot)$ taking the place of $c(\cdot)$ and Ω the

place of T) satisfy the assumptions of Theorem (I.3.1). For the purpose of applying Theorem (I.3.1) we shall agree to call a point $\hat{z} \in \Omega$ desirable if $\bar{\varphi}_0(\hat{z}) = 0$.

First we must show that the characterization (I.3.2) is satisfied. Thus, suppose that $z_0 \in \Omega$ satisfies $\bar{\varphi}_0(z_0) = 0$. Then, since for all $\epsilon_0 > 0$, $J_{\epsilon_0}(z_0) \supset J_0(z_0)$, we must have $-\epsilon_0 < \bar{\varphi}_0(z_0) \leq \bar{\varphi}_{\epsilon_0}(z_0)$. Hence, after a finite number of halvings of ϵ_0 in Step 3, the algorithm will find that $\bar{\varphi}_0(z_0) = 0$ and will set $z_0 = z_0$, i.e., $a(z_0) = z_0$. This is in agreement with (I.3.2).

Now, given a point $z_0 \in \Omega$, the algorithm can only construct a new point z_1 such that $f^0(z_1) \leq f^0(z_0)$. Hence, suppose that the algorithm sets $z_1 = z_0$ (i.e., $z_0 = z_0$ in Step 3 or Step 6). If z_0 was reset to z_0 in Step 3, $\bar{\varphi}_0(z_0) = 0$. Suppose z_0 was reset to z_0 in Step 6, i.e., $\mu(z_0)h(z_0) = 0$. Then this implies that $h(z_0) = 0$[†] and hence that $\bar{\varphi}_{\epsilon_0}(z_0) = 0$, i.e., that $\bar{\varphi}_{\epsilon_0}(z_0) > -\epsilon_0$: a condition in Step 3 which does not permit a continuation to Step 6. Thus z_0 can only be reset to the value z_0 in Step 3 and then it satisfies $\bar{\varphi}_0(z_0) = 0$.

We shall now show that Condition (I.3.3) is satisfied. Let $z_0 \in \Omega$ be any point such that $\bar{\varphi}_0(z_0) < 0$. Then, from (9) and (I.4.7),

11. $$f^0(z_0 + \mu(z_0)h(z_0)) - f^0(z_0) \triangleq -\delta_0 < 0 .$$

It now follows from (5c) that there must exist a $\rho' > 0$ such that

12. $$J_{\epsilon_0}(z) \subset J_{\epsilon_0}(z_0) \quad \text{for all } z \in B(z_0, \rho') ,$$

where $B(z_0, \rho') \triangleq \{z \mid z \in \Omega, \|z-z_0\| \leq \rho'\}$ and ϵ_0 is the value of ϵ used in Step 2 in computing the $h(z_0)$ which is then used in Steps 4, 5, and 6 (i.e., it is the last value of ϵ used in conjunction with the given $z = z_0$). Let $\bar{m}: \mathbb{R}^n \to \mathbb{R}^1$ be defined by

---

[†]If $h(z_0) \neq 0$, then by construction $\bar{\varphi}_{\epsilon_0}(z_0) < -\epsilon_0 < 0$, (see (5)). It now follows from (I.4.1) and (I.4.7) that $\mu(z_0) \neq 0$.

Copyright © 2013, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

13. $$\bar{m}(z) = \min_{h \in S} \max_{i \in J_{\varepsilon_0}(z_0)} \langle \nabla f^i(z), h \rangle$$

Then $\bar{m}(\cdot)$ is continuous (see (I.4.18)) and there is a $\rho'' > 0$ such that

14. $$|\bar{m}(z) - \bar{\varphi}_{\varepsilon_0}(z_0)| \leq \varepsilon_0/2 \quad \text{for all } z \in B(z_0, \rho'').$$

Let $\rho = \min\{\rho', \rho''\}$, then, because of (12) and (14) and the fact that $\bar{\varphi}_{\varepsilon_0}(z_0) \leq -\varepsilon_0$, we have, for all $z \in B(z_0, \rho)$, that

15. $$\bar{\varphi}_{\varepsilon_0}(z) \leq \bar{m}(z) \leq -\varepsilon_0/2 .$$

But $J_{\varepsilon_0/2}(z) \subset J_{\varepsilon_0}(z)$, and hence, for all $z \in B(z_0, \rho)$, we have

16. $$\bar{\varphi}_{\varepsilon_0/2}(z) \leq \bar{\varphi}_{\varepsilon_0}(z) \leq -\varepsilon_0/2 .$$

We therefore conclude that for all $z \in B(z_0, \rho)$ the algorithm (7) will use a value $\varepsilon(z) \geq \varepsilon_0/2$ in computing the $h(z)$ in Step 2 for use in Steps 4, 5 and 6, i.e., for all $z \in B(z_0, \rho)$ and for all $i \in J_{\varepsilon(z)}(z)$, $\langle \nabla f^i(z), h(z) \rangle \leq -\varepsilon_0/2$.

Now, for any $z \in B(z_0, \rho)$ and $i = 0, 1, 2, \ldots, m$, we have, by the mean value theorem, that

17. $$f^i(z + \lambda h(z)) = f^i(z) + \lambda \langle f^i z + \zeta h(z), h(z) \rangle ,$$

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 the compact set $B(z_0, \rho) \times S$, there exists a $\lambda^1 > 0$ (see (I.4.7)) such that for all $z \in B(z_0, \rho/2)$, and for all $i \in \{0, 1, 2, \ldots, m\}$,

18. $$|\langle \nabla f^i(z + \zeta h(z)), h(z) \rangle - \langle \nabla f^i(z), h(z) \rangle| \leq \varepsilon_0/4 ,$$

for all $\zeta \in [0, \lambda^1]$. Similarly, since the functions $f^i(\cdot)$ are uniformly continuous on $B(z_0, \rho)$ and since S is compact, there exists a $\lambda^2 > 0$ (see (I.4.1)) such that for all $z \in B(z_0, \rho/2)$ and for all $i \in \{1, 2, \ldots, m\}$,

19. $$|f^i(z + \zeta h(z)) - f^i(z)| \leq \varepsilon_0/2 ,$$

for all $\zeta \in [0,\lambda^2]$. Now, for each $z \in B(z_0, \rho/2)$ and for each $i \in J_{\varepsilon(z)}(z)$, $\langle \nabla f^i(z), h(z) \rangle \leq -\varepsilon_0/2$, and for each $z \in B(z_0, \rho/2)$ and for each $i \in \bar{J}_{\varepsilon(z)}(z)$[†], $f^i(z) \leq -\varepsilon_0/2$. Hence, setting $\lambda_m = \min \{\lambda^1, \lambda^2\}$, we have, for any $z \in B(z_0, \rho/2)$

30a. $$f^i(z + \lambda_m h(z)) - f^i(z) \leq \lambda_m \varepsilon_0/4 \quad \text{for all } i \in J_{\varepsilon(z)}(z) ;$$

30b. $$f^i(z + \lambda_m h(z)) \leq 0 \quad \text{for all } i \in \bar{J}_{\varepsilon(z)}(z) .$$

Since for all $z \in B(z_0, \rho/2)$ we must have $\mu(z) \geq \lambda_m$, we are led to the conclusion that

21. $$-f^0(z + \mu(z)h(z)) - (-f^0(z)) \geq \lambda_m \varepsilon/4, \quad \text{for all } z \in B(z_0, \rho/2) ,$$

i.e., that condition (I.3.3) is satisfied. This completes our proof.

We have already observed that by setting $S = \{h \in \mathbb{R}^n | \ |h^i| \leq 1\}$, we can compute $\bar{\varphi}_{\varepsilon(z)}(z)$ and $h(z)$ by solving a linear programming problem, i.e., these quantities are obtainable by finite step procedures. Thus, the weak link in the algorithm (7) seems to be the requirement of solving exactly equations of the form $f^i(z_0 + \lambda h(z_0)) = 0$ for $\lambda(z_0)$ and of minimizing the function $f^0(\cdot)$ along the linear segment $\{z | \ z = z_0 + \mu h(z_0), \ \mu \in [0, \lambda(z_0)]\}$. Neither of these operations can be performed in a finite number of steps. The following propositions are obvious in the light of Theorem (I.3.16) and show to what extent these operations may be approximated without affecting the convergence properties of the algorithm (7). The reader should have no difficulty in adapting them also for algorithm (2.5) in which the requirement of minimizing $d(z',z)$ exactly / results in the same sort

---

[†] $\bar{J}_{\varepsilon(z)}(z)$ denotes the complement of $J_{\varepsilon(z)}(z)$ in $\{0, 1, 2, \ldots, m\}$.

of difficulty.

22. Proposition: Suppose that in Step 6 of the Algorithm (7) z_0 is reset to $z_0 + \mu_0 h(z_0)$, where, for a fixed $\beta \in (0,1]$, μ_0 satisfies

23. $$(f^0(z_0) - f^0(z_0 + \mu_0 h(z_0))) \geq \beta(f^0(z_0) - f^0(z_0 + \mu(z_0)h(z_0))) \quad ,$$

where $\mu(z_0)$ is defined as in (9). Then Theorem (10) remains valid. (c.f. (II.1.15)).

24. Proposition: Suppose that the functions $f^i(\cdot)$ are convex, and that the sets $\{z \mid f^i(z) \leq 0\}$ are compact for $i = 0, 1, \ldots, m$, and that Steps 4 and 5 of the Algorithm (7) are replaced by the Steps 4', 5' below.† Then Theorem (10) still remains valid.

Step 4': Assume an $\alpha \in (0,1/2)$ is given. Compute $\lambda^i > 0$, $i = 0, 1, \ldots, m$, to satisfy

25a. $$(1-\alpha)\lambda^0 \langle \nabla f^0(z), h(z) \rangle \leq f^0(z + \lambda^0 h(z)) - f^0(z) \leq \alpha \lambda^0 \langle \nabla f^0(z), h(z) \rangle \; ;$$

25b. $$\lambda^i(\alpha) \langle \nabla f^i(z), h(z) \rangle \leq f^i(z + \lambda^i h(z)) - f^i(z) \leq -f^i(z) \quad \text{for } i \neq 0,\ i \in J_\varepsilon(z),$$

25c. $$-\lambda^i \alpha \leq f^i(z + \lambda^i h(z)) \leq 0 \quad , \text{ for } i \in \bar{J}_\varepsilon(z) \quad .$$

Step 5': Set $\mu(z) = \min \{\lambda^0, \lambda^1, \ldots, \lambda^m\}$.

26. Proposition: Suppose that a $\rho > 0$, a $\beta \in (0,1)$ and an $\alpha \in (0,1)$ are given, and suppose that Steps 4 and 5 of Algorithm (7) are replaced by the Steps 4", 5" below. Then Theorem (10) remains valid.

Step 4": Compute the smallest integer $k \geq 0$ such that

27a. $$f^0(z + \beta^k \rho h(z)) - f^0(z) - \beta^k \rho \, \alpha \langle \nabla f^0(z), h(z) \rangle \leq 0^{\dagger\dagger}$$

27b. $$f^i(z + \beta^k \rho h(z)) \leq 0 \quad \text{for } i = 1, 2, \ldots, m \quad .$$

Step 5": Set $\mu(z) = \beta^k \rho$.

---

†The reader should compare this procedure with (II.1.18).

††β^k is β to the power k.

The introduction of ε into the Algorithm (7) ensures that for each non-optimal $z_0 \in \Omega$, there exists a $\rho > 0$ and a $\lambda_m > 0$ such that for all $z \in \Omega$, $\|z-z_0\| \leq \rho$, we have $z + \lambda h(z) \in \Omega$ for all $\lambda \in [0,\lambda_m]$, i.e., it ensures a minimal step size about each non-optimal $z_0 \in \Omega$. This effect was used in the proof of Theorem (10).

A second important, but not entirely independent, effect of using ε in (7) is to ensure that we do not solve systems of simultaneous equations of the form $f^i(z) = 0$, $i \in I$, 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 $z_0, z_1, z_2, \ldots$, constructed by an algorithm not using an ε-procedure. Thus, an algorithm would jam (or zig-zag) without "the antijamming precautions" defined by the use of ε in the algorithm (7).

Equality Constraints.

We shall now indicate how the exterior penalty function method (1.39) can be combined with the method of feasible directions to solve problems of the form

28. $$\min \{f^0(z) \mid f(z) \leqq 0, \quad r(z) = 0\}$$

where $f^0: \mathbb{R}^n \to \mathbb{R}^1$, $f: \mathbb{R}^n \to \mathbb{R}^m$, $r: \mathbb{R}^n \to \mathbb{R}^\ell$ are continuously differentiable. We shall assume that the matrices $\frac{\partial f(z)}{\partial z}$ and $\frac{\partial r(z)}{\partial z}$ are of maximum rank for all z in a "sufficiently large" open set containing the set $\{z \mid f(z) \leqq 0, r(z) = 0\}$. We shall also assume that the set $\{z \mid f(z) \leqq 0\}$ has an interior.

To solve (28), we can apply Algorithm (7) (or one of its modifications) to the problem

29. $$\min \{f^0(z) + \frac{1}{2\varepsilon''} \|r(z)\|^2 \mid f^i(z) \leq 0, \ i = 1, 2, \ldots, m\} \ .$$

where $\varepsilon'' > 0$ and is driven to zero by modifying the Algorithm (7) as follows.

We assume that we start with $\varepsilon'' > \varepsilon_0 > \varepsilon'$, with ε'' fairly large, and that we have a $z_0 \in \Omega \triangleq \{z \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$. We also introduce scale factors $\beta', \beta'' \subset (0,1)$

30. Algorithm:†

Step 0: Set $z = z_0$

Step 1: Set $\varepsilon(z) = \varepsilon_0$ $(\varepsilon(z) = \varepsilon)$.

Step 2: Compute $\bar{\varphi}_\varepsilon(z)$, $h_\varepsilon(z)$ by solving (6) with

$$f^0(z) + \frac{1}{2\varepsilon''}\|r(z)\|^2 \quad \text{taking the place of } f^0(z).$$

Step 3: If $\bar{\varphi}_\varepsilon(z) \leq -\varepsilon$, set $h(z) = h_\varepsilon(z)$ and go to Step 4.

If $\bar{\varphi}_\varepsilon(z) > -\varepsilon$, and $\varepsilon \leq \varepsilon'$, set $\varepsilon' = \beta'\varepsilon'$, $\varepsilon'' = \beta''\varepsilon''$ and go to Step 1.

If $\bar{\varphi}_\varepsilon(z) > -\varepsilon$, and $\varepsilon > \varepsilon'$, set $\varepsilon = \varepsilon/2$ and go to Step 2.

Step 4: Compute $\lambda(z) \geq 0$ such that

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

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

32. $$f^0_{\varepsilon''}(z + \mu(z)h(z)) = \min\{f^0_{\varepsilon''}(z + \mu h(z)),\ \mu \in [0,\lambda(z)]\},$$

where

33. $$f^0_{\varepsilon''}(z) \triangleq f^0(z) + \frac{1}{2\varepsilon''}\|r(z)\|^2$$

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

We may proceed somewhat as for the Algorithm (1.39) to establish that the sequence of points $\{z_i\}$, computed by the above algorithm, may have at least one subsequence which converges to a point $\hat{z}$ satisfying the necessary condition of optimality (see (I.2.3)),

---

†This method does not seem to have been published before.

34. $$-\nabla f^0(\hat{z}) + \frac{\partial r(\hat{z})}{\partial z}^T \psi + \left(\frac{\partial f(\hat{z})}{\partial z}\right)^T \mu = 0$$

35. $$\mu \leqq 0, \quad \langle \mu, f(\hat{z}) \rangle = 0 .$$

It is also possible to introduce some elements of the modified Newton-Raphson method into feasible directions algorithms. The manner in which this can be done will be sketched out towards the end of the next section, and will then be discussed in detail in Section 5.

4. Further Applications to Optimal Control

We shall now show, by means of two examples, how the Algorithms (2.5), (3.7) and (3.30) (and their modifications) can be applied to certain classes of optimal control problems.

Case 1: Consider the optimal control problem,

1. $$\text{minimize} \sum_{i=0}^{k-1} f_i^0(x_i,u_i)$$

subject to

2. $$x_{i+1} - x_i = f_i(x_i,u_i), \quad i = 0, 1,\ldots, k-1,$$

with $x_i \in R^\nu$, $u_i \in R^1$ restricted as follows

3a. $$x_0 = \hat{x}_0,\ q^j(x_k) \le 0 \quad \text{for} \quad j = 1, 2,\ldots, m,$$

3b. $$|u_i| \le 1, \quad i = 0, 1,\ldots, k-1 .$$

Setting $z = (u_0, u_1,\ldots, u_{k-1})$, this problem becomes

4. $$\min \{f^0(z) \mid f^i(z) \le 0, \quad i = 1, 2,\ldots, m ;$$
$$|u_j| \le 1, \quad j = 0, 1,\ldots, k-1\} ,$$

where

5. $$f^0(z) = \sum_{i=0}^{k-1} f_i^0(x_i(z),u_i) , \quad i = 1, 2,\ldots, m$$

6. $$f^i(z) = q^i(x_k(z)) ,$$

and $x_i(z)$, $i = 0, 1, 2,\ldots,$ are determined by $x_0(z) = \hat{x}_0$ and

7. $$x_{i+1}(z) - x_i(z) = f_i(x_i(z),u_i), \quad i = 0, 1, 2,\ldots, k-1 .$$

The problem (4) is in standard form for the Algorithm (3.7), since $|u_j| \leq 1$ is equivalent to the pair of inequalities $u_j - 1 \leq 0$, $-1 - u_j \leq 0$, provided we assume that $\Omega \triangleq \{z \mid f^i(z) \leq 0 \quad i = 1, 2, \ldots, m;\ |u_j| \leq 1,\ j = 0,1,\ldots,k-1\}$ has an interior.
Thus, the only thing that remains to be done is to see how to utilize the dynamical structure of (7) in the calculation of the various required derivatives for (3.7). We note that we have already developed a procedure for calculating $\nabla f^0(z)$ in Section II.5 to which the reader is referred. To calculate $\nabla f^i(z)$, $i \in \{1, 2, \ldots, m\}$, we note that

$$8. \qquad \nabla f^i(z) = \left(\frac{\partial x_k(z)}{\partial z}\right)^T \nabla q^i(x_k(z))$$

Hence, for $i = 1, 2, \ldots, m$,

$$9. \qquad \frac{\partial f^i(z)}{\partial u_j} = \langle \frac{\partial x_k(z)}{\partial u_j}, \nabla q^i(x_k(z)) \rangle, \quad j = 0, 1, \ldots, k-1$$

But by (II.5.8),

$$10. \qquad \frac{\partial x_k(z)}{\partial u_j} = \Phi_{k,j+1} \frac{\partial f_j(x_j(z),u_j)}{\partial u_j}$$

where $\Phi_{k,j+1}$ is a $\nu \times \nu$ matrix calculated from

$$11. \qquad \Phi_{i+1,j+1} - \Phi_{i,j+1} = \frac{\partial f_i(x_i(z),u_i)}{\partial x_i} \Phi_{i,j+1}, \Phi_{j+1,j+1} = I \quad \text{(the identity matrix)}, \quad i = j+1, j+2, \ldots, k,$$

Thus, from (9) and (10), for $i = 1, 2, \ldots, m$, and $j = 0, 1, 2, \ldots, k-1$,

$$12. \qquad \frac{\partial f^i(z)}{\partial u_j} = \langle \frac{\partial f_j(x_j(z),u_j)}{\partial u_j}, \Phi^T_{k,j+1} \nabla q^i(x_k(z)) \rangle$$

Referring to the development in Section II.5 we now see that the $\frac{\partial f^i(z)}{\partial u_j}$ can be calculated as follows for a given z. First calculate the $x_j(z)$, $j = 0, 1, 2, \ldots, k$, using (7), with $x_0(z) = \hat{x}_0$. Next, for $i = 1, 2, \ldots, m$, calculate the vectors $p_{i,i}, p_{i+1,i}, \ldots, p_{k,i}$, in $\mathbb{R}^\nu$, from

13. $$p_{ji} - p_{(j+1)i} = \left(\frac{\partial f_j(x_j(z),\, u_j)}{\partial x_j}\right)^T p_{(j+1)i};\ p_{ki} = \nabla q^i(x_k(z)) .$$

This yields $p_{ji} = \Phi_{k,j}^T \nabla q^i(x_k(z))$ and hence, for $i = 1, 2, \ldots, m$, and $j = 0, 1, 2, \ldots, k-1$,

14. $$\frac{\partial f^i(z)}{\partial u_j} = \left\langle \frac{\partial f_j(x(z), u_j)}{\partial u_j}\, ,\ p_{(j+1)\ i} \right\rangle .$$

Thus, at no time do we really need to manipulate particularly large arrays in calculating the desired derivatives. Next we note that the linearity of the inequalities $|u_j| \leq 1$, which must be satisfied by the u_i, can also be exploited. Thus, to calculate $\bar{\varphi}_\varepsilon(z)$, we must solve by (3.6) (with $S = \{h \mid |h^i| \leq 1\}$), i.e., we must solve

15. $$\text{minimize } \sigma$$

subject to

15a. $$\sigma - \langle \nabla f^i(z), h \rangle \geq 0, \text{ for } i = 0 \text{ and all } i \in \{1, 2, \ldots, m\} / \text{ such that } f^i(z) + \varepsilon \geq 0;^\dagger$$

15b. $$\sigma - h^j \geq 0 \text{ for all } j \text{ such that } u_j - 1 + \varepsilon \geq 0$$

15c. $$\sigma + h^j \geq 0 \text{ for all } j \text{ such that } -u_j - 1 + \varepsilon \geq 0$$

15d. $$|h^j| \leq 1, \quad j = 0, 1, \ldots, k-1.$$

Algorithm (2.5) may be used in a similar manner.
To solve problem (15) efficiently, one should use generalized upper boundary techniques (Sec. [IV]). With these techniques one would have to invert matrices whose dimension is governed by the number of inequalities in (15a) only. Alternatively, one may compute a feasible direction h at z by solving

16. Minimize σ subject to

16a. $\sigma - \langle \nabla f^i(z), h \rangle \geq 0$ for $i = 0$ and all $i \in \{1,\ldots,m\}$ such that $f^i(z) + \epsilon \geq 0$

16b. $-h^j \geq 0$ for all j such that $u_j - 1 + \epsilon \geq 0$

16c. $h^j \geq 0$ for all j such that $-u_j - 1 + \epsilon \geq 0$

16d. $|h^j| \leq 1$, $j = 0, 1, \ldots, k-1$

The justification for (16) can be obtained as follows: Consider for the moment problem (3.1) and let $J_\epsilon^A(z) \subset J_\epsilon(z)$ and $J_\epsilon^N(z) \subset J_\epsilon(z)$ be such that $J_\epsilon^A(z) \cup J_\epsilon^N(z) = J_\epsilon(z)$ and for all $i \in J_\epsilon^A(z)$, $f^i(z)$ be affine.

17. Definition: For $\epsilon > 0$, let $\eta_\epsilon : \Omega \to R^1$ defined by

$$\eta_\epsilon(z) = \text{Min}\{\sigma \mid \sigma - \langle \nabla f^i(z), h \rangle \geq 0,\ i \in J_\epsilon^N(z);\ -\langle \nabla f^i(z), h \rangle \geq 0,\ i \in J_\epsilon^A(z);\ h \in S\}.$$

It is now easy to show that

(a) $\eta_\epsilon(z) \leq \bar{\varphi}_\epsilon(z) \leq 0$ for all $\epsilon \geq 0$

(b) If z is optimal for (3.1), then $\eta_0(z) = 0$.

(c) Can use $\eta_\epsilon(\cdot)$ instead of $\bar{\varphi}_\epsilon(\cdot)$ in Algorithm (3.7) without upsetting the convergence properties of the algorithm.

Case 2: Consider the following simple problem:

18. $$\text{minimize } \frac{1}{2} \sum_{i=0}^{k-1} (\|x_i - x_i^*\|^2 + u_i^2)$$

---

†We number the components of h in the same manner as the components of the control sequence z, i.e., $h = (h^0, h^1, \ldots, h^{k-1})$.

subject to

18d. $$x_{i+1} = Ax_i + bu_i, \quad i = 0, 1, \ldots, k-1, \quad x_i \in R^{\nu}, \; u_i \in \mathbb{R}^1,$$

with the boundary conditions $x_0 = x_0^*$, $x_k = x_k^*$, and the $u_i \in [-1,+1]$ for $i = 0, 1, 2, \ldots, k-1$.

Assuming that this problem is "nondegenerate," i.e., that $p^0 = -1$, the necessary conditions of optimality stated in (I.2.14) become for this problem: if $\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{k-1}$ and $\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_{k-1}$ are optimal for (18), (18a), then,

19. $$\hat{x}_{i+1} = A\hat{x}_i + b\hat{u}_i, \quad i = 0, 1, \ldots, k-1 ,$$

So, $$\hat{x}_0 = x_0^*, \; \hat{x}_k = x_k^*$$

and there exist multiplier vectors $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k$ in $\mathbb{R}^{\nu}$, satisfying

21. $$\hat{p}_i = A^T \hat{p}_{i+1} - (\hat{x}_i - x_i^*), \quad i = 0, 1, 2, \ldots, k-1 .$$

such that (from (I.2.17)),

22. $$\hat{u}_i = \text{sat} \langle \hat{p}_{i+1}, b \rangle, \quad i = 0, 1, \ldots, k-1 .$$

Substituting from (22) into (19), we obtain

23. $$\hat{x}_{i+1} = A\hat{x}_i + b \; \text{sat} \langle \hat{p}_{i+1}, b \rangle$$

Thus, to find the optimal control sequence, $\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{k-1}$, together with optimal trajectory, $\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_{k-1}$, we must solve (23) and (21) (and use (22)) with the mixed boundary conditions $x_0 = x_0^*$, $x_k = x_k^*$. Now, in (23), sat $\langle p_{i+1}, b \rangle$ is not continuously differentiable in p_{i+1} and hence we cannot apply the modified Newton-Raphson method (II.3.2) to this problem (c.f. Section II.5).

However, we can apply Algorithm (3.30), as follows. We construct a penalized cost function, with $z = (u_0, u_1, \ldots, u_{k-1})$, and $\varepsilon'' > 0$ large,

24. $$f_\varepsilon^0{}''(z) \triangleq \frac{1}{2}\sum_{i=1}^{k-1}{}' \|x_i(z) - x_i^*\|^2 + \sum_{i=0}^{k-1}{}' u_i^2 + \frac{1}{2\varepsilon''}\|x_k(z) - x_k^*\|^2$$

Then, we apply Algorithm (3.30) to min $\{f_\varepsilon^0{}''(z) \mid |u_i| \leq 1,\ i = 0, 1, 2, \ldots, k-1\}$. To evaluate $\bar{\varphi}_\varepsilon(z)$, for a given z, we solve the linear program,

25. minimize σ

subject to

25a. $$\sigma - \langle \nabla f_\varepsilon^0{}''(z), h\rangle \geq 0$$

25b. $$\sigma - h^j \geq 0 \quad \text{for all } j \text{ such that } u_j - 1 + \varepsilon \geq 0$$

25c. $$\sigma + h^j \geq 0 \quad \text{for all } j \text{ such that } -u_j - 1 + \varepsilon \geq 0$$

25d. $$|h^j| \leq 1,\ j = 0, 1, 2, \ldots, k-1$$

where we calculate $\nabla f_\varepsilon^0{}''(z)$ as indicated in (II.5).

An Extension of Algorithm (3.7)†

In computing $\bar{\varphi}_\varepsilon(z)$ by (15) or by (25) we may often find that the inequalities (15b), (15c) (or (25b) and (25c)) can be quite numerous and may therefore have an appreciable effect on the computation time, even if special linear programming codes are used (such as described in [V1, V2].

We shall now present a few modifications of the Algorithm (3.7) which are more suitable for use with optimal control problems. Consider again the problem

26. $$\min\{f^0(z) \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$$

where the $f^i(\cdot)$, $i = 0, 1, \ldots, m$, are continuously differentiable functions from $\mathbb{R}^n$ into $\mathbb{R}^1$.

27. Definition: For $\epsilon \geq 0$, let $\hat{\varphi}_\epsilon$: $\Omega \to \mathbb{R}^1$ ($\Omega \triangleq \{z \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$) be defined by

28. $$\hat{\varphi}_\epsilon(z) = \min \{\langle \nabla f^0(z), h\rangle \mid \langle \nabla f^i(z), h\rangle + \epsilon \leq 0 \text{ for } i \neq 0,\ i \in J_\epsilon(z);\ h \in S\}$$

where S is any compact set containing the origin in its interior. (This function was first used in [Z5].)

29. Proposition: Suppose that for every $z \in \Omega$ there exists a vector $h \in \mathbb{R}^n$ such that $\langle \nabla f^i(z), h\rangle < 0$ for all $i \neq 0$, $i \in J_0(z)$. Then, for any $z \in \Omega$, $\hat{\varphi}_0(z) = 0$ if and only if $\bar{\varphi}_0(z) = 0$.

Proof: We give a proof by contraposition.

$\Longrightarrow$ Suppose that for some $z \in \Omega$, $\bar{\varphi}_0(z) < 0$, then by inspection, $\hat{\varphi}_0(z) < 0$, i.e., $\hat{\varphi}_0(z) = 0 \Longrightarrow \bar{\varphi}_0(z) = 0$.

$\Longleftarrow$ Now suppose that for some $z \in \Omega$, $\hat{\varphi}_0(z) < 0$, with $\hat{\varphi}_0(z) = \langle \nabla f^0(z), \hat{h}\rangle$. Then $\langle \nabla f^0(z), \hat{h}\rangle < 0$, $\langle \nabla f^i(z), \hat{h}\rangle \leq 0$, $i \in J_0(z)$, $i \neq 0$ and $\hat{h} \in S$. Let $h \in \mathbb{R}^n$ be such that $\langle \nabla f^i(z), h\rangle < 0$. Then, since S has an interior, there exist a $\lambda^1 > 0$ such that $\tilde{h} \triangleq \lambda^1(\hat{h} + \lambda^2 h) \in S$ and $\langle \nabla f^i(z), \tilde{h}\rangle < 0$ for $i \in J_0(z)$, i.e., $\hat{\varphi}_0(z) < 0 \Longrightarrow \bar{\varphi}_0(z) < 0$, which is equivalent to the statement that $\bar{\varphi}_0(z) = 0 \Longrightarrow \hat{\varphi}_0(z) = 0$. This completes our proof.

Thus, under the assumptions stated in (29), finding points $z \in \Omega$ which satisfy $\hat{\varphi}_0(z) = 0$ seems to be about as good an idea as finding points $z \in \Omega$ which satisfy $\bar{\varphi}_0(z) = 0$. However, for optimal control problems such as the ones we have examined in this section, $\hat{\varphi}_\epsilon(z)$ is much easier to compute than $\bar{\varphi}_\epsilon(z)$. For example, for the problem considered in Case I, to compute $\hat{\varphi}_\epsilon(z)$ we solve

30. $$\min \langle \nabla f^0(z), h\rangle$$

subject to

30a. $\langle \nabla f^i(z), h\rangle + \varepsilon \leq 0$ for all $i \in \{1, 2, \ldots, m\}$

such that $f^i(z) + \varepsilon \geq 0$;

30b. $-h^j + \varepsilon \leq 0$ for all $j \in \{0, 1, 2, \ldots, k-1\}$ such that $-u_j - 1 + \varepsilon \geq 0$;

30c. $h^i + \varepsilon \leq 0$ for all $j \in \{0, 1, \ldots, k-1\}$ such that $u_j - 1 + \varepsilon \geq 0$;

30d. $|h^j| \leq 1, \quad j = 0, 1, 2, \ldots, k-1$.

Now the only inequalities which determine the dimension of the matrices to be inverted when solving (30) by means of the simpler algorithm are (30a), which usually are few in number. Hence $\hat{\varphi}_\varepsilon(z)$ is easier to compute for this case than $\bar{\varphi}_\varepsilon(z)$.

31. Theorem: Consider the problem (26). Suppose that for every $z \in \Omega$ there exists a vector $h \in \mathbb{R}^n$ such that $\langle \nabla f^i(z), h\rangle < 0$ for all $i \neq 0$, $i \in J_0(z)$. Then the function $\hat{\varphi}_\varepsilon(\cdot)$ in Algorithm (3.7) without effecting the convergence properties of (3.7), i.e., Theorem (3.10) remains valid.

We leave the proof of this theorem as an exercise for the reader who will find in the second half of the next section a few helpful results.

We observe that we can also modify Algorithm (2.5) to make its application to optimal control problems easier, as follows.

32. Definition: Consider the problem (26). Let $\hat{\varphi}: \Omega \to \mathbb{R}^1$ be defined by

$$\hat{\varphi}(z) = \min \{\langle \nabla f^0(z), h\rangle \mid f^i(z) + \langle \nabla f^i(z), h\rangle \leq 0 \quad i = 1, 2, \ldots, m;\ h \in S\} ;$$

where S is any compact set containing the origin in its interior.

33. Proposition: Suppose that for every $z \in \Omega$ there exists an $h \in S$ such that $\langle \nabla f^i(z), h \rangle < 0$ for all $i \neq 0$, $i \in J_0(z)$. Then, for any $z \subset \Omega$, $\hat{\varphi}(z) = 0$ if and only if $\varphi(z) = 0$, where $\varphi(\cdot)$ was defined in (2.2b).

The proof of this proposition is similar to that of (29) and will therefore be omitted.

34. Theorem: Suppose that for every $z \in \Omega$ there exists an $h \in S$ such that $\langle \nabla f^i(z), h \rangle < 0$ for all $i \neq 0$, $i \in J_0(z)$. Then the function $\hat{\varphi}(\cdot)$ in Algorithm (2.5) without affecting its convergence properties, i.e., Theorem (2.8) remains valid.

We again leave the proof of this theorem as an exericse for the reader and again suggest that he read the next section before attempting to carry out the proof.

For the problem considered in Case I $\hat{\varphi}(z)$ is computed by solving

35. $$\min \langle \nabla f^0(z), h \rangle$$

subject to

35a. $$f^i(z) + \langle \nabla f^i(z), h \rangle \leq 0, \quad i = 1, 2, \ldots, m$$

35b. $$u_j - 1 + h^j \leq 0, \quad j = 0, 1, \ldots, k-1$$

35c. $$-u_j - 1 - h^j \leq 0, \quad j = 0, 1, \ldots, k-1$$

35d. $$|h^j| \leq 1, \quad j = 0, 1, \ldots, k-1$$

Again it can be seen that this is easier to solve by means of the simplex algorithm then the problem whose solution yields $\varphi(z)$.

A "Second Order" Extension of Algorithm (3.7).

It is interesting to observe that some ideas of the modified Newton-Raphson method can also be injected into the method of feasible directions. Consider again the problem min $\{f^0(z) \mid z \in \mathbb{R}^n\}$. Then, the direction $h(z_0)$ given by the Newton-Raphson method at z_0 is the one which is obtained by minimizing the quadratic approximation

36. $$\langle \nabla f^0(z_0), h\rangle + \frac{1}{2} \langle h \frac{\partial^2 f^0(z_0)}{\partial z^2} h\rangle$$

to $f^0(z) - f^0(z_0)$ (with $h = z - z_0$) at z_0. Indeed, taking the gradient of (36) with respect to h and setting it equal to zero, we obtain

36. $$\nabla f^0(z_0) + \frac{\partial^2 f^0(z_0)}{\partial z^2} h = 0$$

so that $h(z_0) = -\left(\frac{\partial^2 f^0(z_0)}{\partial z^2}\right)^{-1} \nabla f^0(z_0)$ solves (36). Let us now return to the problem (3.1), and assume that $f^0(\cdot)$ is convex, and that $\left\|\frac{\partial^2 f^0(z)}{\partial z^2}\right\| \geq \lambda > 0$ for all z in a "sufficiently large" open set in $\mathbb{R}^n$.

38. Definition: Let H(z) be an n x n positive definite matrix whose elements are continuous functions of z. For $\epsilon > 0$, we define the function $\tilde{\varphi}_\epsilon : \Omega \to R^1$ as

$$\tilde{\varphi}_\epsilon(z) = \min \{\langle \nabla f^0(z), h\rangle + \langle h, H(z)h\rangle \,\Big|\, h \in \mathbb{R}^n, \langle \nabla f^i(z), h\rangle + \epsilon \leq 0 \text{ for } i \neq 0, i \in J_\epsilon(z)\}$$

39. Proposition: Suppose that for every $z \in \Omega$ there exists a vector $h \in \mathbb{R}^n$ such that $\langle \nabla f^i(z), h\rangle < 0$ for all $i \neq 0$, $i \in J_0(z)$. Then $\tilde{\varphi}_0(z) = 0$ if and only if $\bar{\varphi}_0(z) = 0$, for every $z \in \Omega$.

The proof of this proposition will be given in the next section.

40. Theorem: Suppose that for every $z \in \Omega$ there exists a vector $h \in \mathbb{R}^n$ such that $\langle \nabla f^i(z), h\rangle < 0$ for all $i \neq 0$, $i \in J_0(z)$. If $\frac{\partial^2 f^0(z)}{\partial z^2} > 0$ and is continuous for all $z \in \mathbb{R}^n$ (or a "sufficiently large" open subset of $\mathbb{R}^n$), and $H(z) = \frac{\partial^2 f^0(z)}{\partial z^2}$ in(38) , then the convergence properties of the Algorithm (3.7) are preserved when the function $\tilde{\varphi}_\epsilon(\cdot)$ is used instead of $\bar{\varphi}_\epsilon(\cdot)$, i.e., Theorem (3.10) remains valid for Algorithm (3.7) modified by the substitution of $\tilde{\varphi}_\epsilon(\cdot)$ for $\bar{\varphi}_\epsilon(\cdot)$. (This theorem can be proved by following the steps in the proof of Theorem (3.10) and making use of the lower semicontinuity of $\tilde{\varphi}_\epsilon(\cdot)$. That $\tilde{\varphi}_\epsilon(\cdot)$ is lower semicontinuous will be established in the next section).

It may be expected that when Algorithm (3.7) uses $\tilde{\varphi}_\epsilon(\cdot)$ it will converge faster than when it uses $\bar{\varphi}_\epsilon(\cdot)$ or $\hat{\varphi}_\epsilon(\cdot)$ as far as the number of iterations is concerned. However, each iteration takes more effort to perform, since second partial derivatives must be calculated and since a quadratic programming problem is harder to solve than a linear programming problem.

In application to control problems, the above indicated modification of Algorithm (3.7) may be particularly attractive for use on minimum energy problems of the type,

41. $$\min \frac{1}{2} \sum_{i=0}^{k-1} (u_i)^2$$

subject to

41a. $$x_{i+1} - x_i = f_i(x_i, u_i), \quad x_i \in \mathbb{R}^\nu, \quad u_i \in \mathbb{R}^k$$

41b. $$|u_i^j| \leq 1, \quad i = 0, 1, \ldots, k-1,$$

41c. $$q^j(x_k) \leq 0, \quad j = 1, 2, \ldots, m \ .$$

In this case, with $z = (u_0, u_1, \ldots, u_{k-1})$, we see that $f^0(z) = \frac{1}{2}\langle z,z\rangle$ and (38) gives $\tilde{\varphi}(z) \overset{\Delta}{=} \min\{\langle z,h\rangle + \frac{1}{2}\langle h,h\rangle \mid \langle \nabla f^i(z),h\rangle + \varepsilon \leq 0$ for all $i \in \{1, 2, \ldots, m\}$ such that $f^i(z) + \varepsilon \geq 0$; $h^j + \varepsilon \leq 0$ for all $j \in \{0, 1, \ldots, k-1\}$ such that $u_j - 1 + \varepsilon \geq 0$; and $-h^j + \varepsilon \leq 0$ for all $j \in \{0, 1, \ldots, k-1\}$ such that $-u_j - 1 + \varepsilon \geq 0\}$.

5. A Second Look At Feasible Directions Algorithms †

In the last three sections, a method of centers as well as a class of methods of feasible directions were presented as algorithms for finding a zero of the functions $\varphi(\cdot)$, $\bar{\varphi}_0(\cdot)$ $\hat{\varphi}(\cdot)$, $\hat{\varphi}_0(\cdot)$, $\tilde{\varphi}_0(\cdot)$ which are zero at all optimal points of the problem

1. $$\min \{f^0(z) \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\} ,$$

where the f^i: $\mathbb{R}^n \to \mathbb{R}^1$, $i = 0, 1, \ldots, m$, are continuously differentiable functions. We shall now show both how the family of such functions and of the resulting minimization algorithms can be extended further. Since we are about to run out of bars and tildas which we have been placing on φ, we shall change our notation slightly.

2. Theorem: For $\epsilon \geq 0$ and $z \in \Omega \triangleq \{z \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$, let $J_\epsilon(z)$ be defined as in (3.2) (i.e., $J_\epsilon(z) = \{0\} \cup \{i \mid f^i(z) + \epsilon \geq 0,\ i \in \{1, 2, \ldots, m\}\}$); let $H_0, H_1, \ldots, H_m$ be arbitrary $n \times n$ positive semidefinite matrices; let S be a compact subset of $\mathbb{R}^n$ containing the origin in its interior, and let $\varphi^1(\cdot)$, $\varphi^2_\epsilon(\cdot)$, $\varphi^3(\cdot)$ and $\varphi^4_\epsilon(\cdot)$, mapping $\mathbb{R}^n$ into $\mathbb{R}^1$, be defined as follows:

3. $$\varphi^1(z) = \min_{h \in S} \left(\max \{\langle \nabla f^0(z), h\rangle;\ f^i(z) + \langle \nabla f^i(z), h\rangle,\ i = 1, 2, \ldots, m\}\right)$$

4. $$\varphi^2_\epsilon(z) = \min_{h \in S} \max_{i \in J_\epsilon(z)} \langle \nabla f^i(z), h\rangle ;$$

5. $$\varphi^3(z) = \min_{h \in S} \left(\max \{\langle \nabla f^0(z), h\rangle + \langle h, H_0 h\rangle ;\ f^i(z) + \langle \nabla f^i(z), h\rangle + \langle h, H_i h\rangle,\ i = 1, 2, \ldots, m\}\right)$$

6. $$\varphi^4_\epsilon(z) = \min_{h \in S} \max_{i \in J_\epsilon(z)} \left(\langle \nabla f^i(z), h\rangle + \langle h, H_i h\rangle\right) .$$

---

† Except as stated, the results in this section appear to be new.

Then, for any $z \in \Omega$, $\varphi^1(z) = 0 \Longleftrightarrow \varphi_0^2(z) = 0 \Longleftrightarrow \varphi^3(z) = 0 \Longleftrightarrow \varphi_0^4(z) = 0$.† (Note that by (I.3.6) $\varphi_0^2(\hat{z}) = 0$ for all $\hat{z}$ which are optimal for (1). Hence, by this theorem we have that if $\hat{z}$ is optimal for (1), then $\varphi^1(\hat{z}) = \varphi_0^2(\hat{z}) = \varphi^3(\hat{z}) = \varphi_0^4(\hat{z}) = 0$).

Proof: (We give a proof by contraposition)

(i) $\varphi^1(z) = 0 \Longleftrightarrow \varphi_0^2(z) = 0$.

$\Longrightarrow$ Suppose that for some $z^* \in \Omega$, $\varphi_0^2(z^*) = \max_{i \in J_0(z^*)} \langle \nabla f^i(z^*), h^* \rangle < 0$. Then, since the origin is in the interior of S, there exists a $\lambda^* > 0$ such that for all $\lambda \in (0, \lambda^*]$ $\lambda h^* \in S$ and, in addition, $\max \{\lambda \langle \nabla f^0(z^*), h^* \rangle; f^i(z^*) + \lambda \langle \nabla f^i(z^*), h^* \rangle, i = 1, 2, \ldots, m\} < 0$, which implies that $\varphi^1(z^*) < 0$, i.e., $\varphi^1(z^*) = 0 \Longrightarrow \varphi_0^2(z^*) = 0$.

$\Longleftarrow$ Suppose that $\varphi^1(z^*) = \max \{\langle \nabla f^i(z^*), h^* \rangle; f^i(z^*) + \langle \nabla f^i(z^*), h^* \rangle, i = 1, 2, \ldots, m\} < 0$. Then $\varphi_0^2(z^*) \leq \max_{i \in J_0(z^*)} \langle \nabla f^i(z^*, h^* \rangle < 0$, i.e., $\varphi_0^2(z^*) = 0 \Longrightarrow \varphi^1(z^*) = 0$.

(ii) $\varphi^1(z) = 0 \Longleftrightarrow \varphi^3(z) = 0$. $\Longrightarrow$ Suppose that for some $z^* \in \Omega$, $\varphi^3(z^*) < 0$, then, by inspection, $\varphi^1(z^*) < 0$, i.e., $\varphi^1(z^*) = 0 \Longrightarrow \varphi^3(z^*) = 0$.

$\Longleftarrow$ Now suppose that $\varphi^1(z^*) = \max \{\langle \nabla f^0(z^*), h^* \rangle; f^i(z^*) + \langle \nabla f^i(z^*), h^* \rangle, i = 1, 2, \ldots, m\} < 0$. Then, because the origin is in the interior of S, there exists a $\lambda^* > 0$ such that for all $\lambda \in (0, \lambda^*]$, $\lambda h^* \in S$ and, in addition, (because when λ is very small the linear terms dominate the quadratic ones), $\varphi^3(z^*) \leq \max \{\lambda \langle \nabla f^0(z^*), h^* \rangle + \lambda^2 \langle h^*, H_0 h^* \rangle; f^i(z^*) + \lambda \langle \nabla f^i(z^*), h^* \rangle + \lambda^2 \langle h^*, H_i h^* \rangle, i = 1, 2, \ldots, m\} < 0$, i.e., $\varphi^3(z^*) = 0 \Longrightarrow \varphi^1(z^*) = 0$.

(iii) To complete the proof we must show that $\varphi_0^2(z) = 0 \Longleftrightarrow \varphi_0^4(z) = 0$. We omit this part of the proof since it is essentially the same as (ii) above.

7. Theorem: Suppose that the functions $f^i(\cdot)$ in (1) are twice continuously differentiable. For $i = 0, 1, 2, \ldots, m$, let $H_i = \frac{1}{2} \frac{\partial^2 f^i(z)}{\partial z^2}$ if $\frac{\partial^2 f^i(z)}{\partial z^2} \geq 0$ for all $z \in \Omega$, and let $H_i = 0$ (the zero matrix) otherwise. Then the function $\varphi^3(\cdot)$ may be used instead of the function $\varphi^1(\cdot)$ in Algorithm (2.5) and the function $\varphi_\epsilon^4(\cdot)$ may be used instead of the function $\varphi_\epsilon^2(\cdot)$ in Algorithm

†$A \Longrightarrow B$ denotes "A implies B". $A \Longleftrightarrow B$ denotes "A implies B and B implies A".

(3.7) without affecting the convergence properties of these algorithms, i.e., Theorems (2.8) and (3.10) remain valid with these substitutions. (At this point the reader is advised to re-read carefully (2.5) and (3.7) as well as the proofs of (2.8) and (3.10)).

Proof: To establish this theorem we only need to observe two facts.

(i) By Section I.4, the functions $\varphi^3(\cdot)$ and $\tilde{m}(\cdot)$, with

$\tilde{m}(z) = \min_{h \in S} \max_{i \in J_\varepsilon(z^*)} \{\langle \nabla f^i(z), h\rangle + \langle h, H_i h\rangle\}$, $\varepsilon \in [0, \varepsilon_0]$ (where $\varepsilon_0 > 0$ is as given in (3.7)), $z^* \in \Omega$, are continuous;

(ii) Since $H_i \geq 0$ for $i = 0, 1, \ldots, m$, for every $z \in \Omega$, and every $h \in S$ $\langle \nabla f^i(z), h\rangle \leq \langle \nabla f^i(z), h\rangle + \langle h, H_i h\rangle$. The reader can now complete the proof by using these facts to modify slightly the proofs of Theorems (2.8) and (3.10), respectively.

The use of the functions $\varphi^3(\cdot)$ and of $\varphi^4_\varepsilon(\cdot)$ in a feasible directions algorithm introduces information about the second-order properties of the functions $f^i(\cdot)$, $i = 0, 1, 2, \ldots, m$, and may therefore be expected to result in accelerated computation, as far as the number of iterations is concerned. However, this advantage is off set (if not totally obliterated) by the fact that to compute $\varphi^3(z)$ or $\varphi^4_\varepsilon(z)$ one must solve a minimization problem with linear cost and quadrative constraints, which is not amenable to finite step procedures. Thus we are led to two other accelerated versions of Algorithms (2.5) and (3.7) (one of which was already sketched out in Section 4) which only require us to solve quadratic programming problems that are amenable to finite step solution. We begin with a few preliminaries.

8. Definition: Let H_0 be a positive definite $n \times n$ matrix. Then, for every $z \in \Omega$ we define $\varphi^5 : \Omega \to \mathbb{R}^1$ by

9. $$\varphi^5(z) = \min \{\langle \nabla f^0(z), h\rangle + \langle h, H_0 h\rangle \mid f^i(z) + \langle \nabla f^i(z), h\rangle \leq 0, \quad i = 1, 2, \ldots, m\}^\dagger$$

---

†A feasible directions algorithm based on this function was presented by Topkis and Veinott in [T1], without proof of convergence.

Also, for every $z \in \Omega$ and every $\varepsilon > 0$, we define

10. $$\varphi_\varepsilon^6(z) = \min \{\langle \nabla f^0(z), h\rangle + \langle h, H_0 h\rangle \mid \langle \nabla f^i(z), h\rangle + \varepsilon \leq 0 \text{ for } i \in J_\varepsilon(z),\ i \neq 0\}$$

11. Theorem: Suppose that for every $z' \in \Omega$ there is a vector $h' \in \mathbb{R}^n$ such that $\langle \nabla f^i(z'), h'\rangle < 0$ for all $i \in J_0(z')$, $i \neq 0$. Then, for every $z \in \Omega$, $\varphi^5(z) = 0 \Longleftrightarrow \varphi_0^6(z) = 0 \Longleftrightarrow \varphi^1(z) = 0$.

Proof: (i) $\varphi^5(z) = 0 \Longleftrightarrow \varphi^1(z) = 0$. $\Longrightarrow$ Suppose that for some $z^* \in \Omega$, $\varphi^1(z^*) < 0$, then by Theorem (2), $\varphi^3(z^*) < 0$, with H_0 as above and all other $H_i = 0$. Hence $\varphi^5(z^*) < 0$, i.e., $\varphi^5(z^*) = 0 \Longrightarrow \varphi^1(z^*) = 0$.

$\Longleftarrow$ Now suppose that for some $z^* \in \Omega$ $\varphi^5(z^*) = \langle \nabla f^0(z^*), h^*\rangle + \langle h^*, H_0 h^*\rangle < 0$. (where h^* satisfies the constraints in (9)). Then there exists a $\lambda' > 0$ and a $\lambda_m > 0$ such that $\lambda(h^* + \lambda' h') \in S$ for all $\lambda \in [0, \lambda_m]$, and $\lambda\langle \nabla f^0(z^*), (h^* + \lambda' h')\rangle + \lambda^2\langle (h^* + \lambda' h'), H_0(h^* + \lambda' h')\rangle < 0$, $f^i(z^*) + \lambda \langle \nabla f^i(z^*), (h^* + \lambda' h')\rangle < 0$, $i = 1, 2, \ldots, m$, for all $\lambda \in (0, \lambda_m]$, and hence $\varphi^1(z^*) < 0$, i.e., $\varphi^1(z^*) = 0 \Longrightarrow \varphi^5(z^*) = 0$.

(ii) $\varphi_0^6(z) = 0 \Longleftrightarrow \varphi^1(z) = 0$. By Theorem (2) we may prove instead that $\varphi_0^6(z) = 0 \Longleftrightarrow \varphi_0^4(z) = 0$, with $H_i = 0$ for $i = 1, 2, \ldots, m$. $\Longrightarrow$ Suppose that for some $z^* \in \Omega$, $\varphi_0^4(z^*) < 0$, then, by inspection, $\varphi_0^6(z^*) < 0$, i.e., $\varphi_0^6(z^*) = 0 \Longrightarrow \varphi_0^4(z^*) = 0$.

$\Longleftarrow$ Now suppose that $\varphi_0^6(z^*) = \langle \nabla f^0(z^*), h^*\rangle + \langle h^*, H_0 h^*\rangle < 0$. Then there exist a $\lambda' > 0$ and a $\lambda_m > 0$ such that $\lambda(h^* + \lambda' h') \in S$ for all $\lambda \in$ $\lambda_m]$, and $\lambda\langle \nabla f^0(z^*), (h^* + \lambda' h')\rangle + \lambda^2\langle h^* + \lambda' h', H_0(h^* + \lambda' h')\rangle < 0$, $\lambda\langle \nabla f^i(z^*), (h^* + \lambda' h')\rangle < 0$ for $i \in J_0(z^*)$ and therefore $\varphi_0^4(z^*) < 0$, i.e., $\varphi_0^4(z^*) = 0 \Longrightarrow \varphi_0^6(z^*) = 0$.

Thus, under the assumptions stated, finding a zero of $\varphi^5(\cdot)$ or of $\varphi_0^6(\cdot)$ seems to be about as good an idea as finding a zero of $\varphi^1(\cdot)$ or $\varphi_0^2(\cdot)$ or $\varphi^3(\cdot)$ or $\varphi_0^4(\cdot)$ and, from what has been said at the end of the preceeding section, an algorithm for finding a zero of $\varphi^5(\cdot)$ or of $\varphi_0^6(\cdot)$ may possibly be somewhat faster than the

algorithms we have already considered for finding a zero of $\varphi^1(\cdot)$ or of $\varphi_0^2(\cdot)$. However, to show that we may substitute $\varphi^5(\cdot)$ for $\varphi^1(\cdot)$ in Algorithm (2.5) or $\varphi_\epsilon^6(\cdot)$ for $\varphi_\epsilon^2(\cdot)$ in Algorithm (3.7) we must exhibit some additional properties of these two new functions.

We digress for a moment to establish a general result.

12. Theorem: Let $\psi(\cdot,\cdot)$ be a continuous function from $\mathbb{R}^n \times \mathbb{R}^n$ into $\mathbb{R}^1$. For every $z \in \mathbb{R}^n$ let $\Omega(z)$ be a subset of $\mathbb{R}^n$ and suppose that

13. $$\bar{\psi}(z) = \min\{\psi(h,z) \mid h \in \Omega(z)\}$$

is well defined. Suppose that $h^* \in \Omega(z^*)$ is arbitrary. If

(i) For every $h^* \in \Omega(z^*)$, $\overset{\circ}{B}(h^*, \epsilon) \neq \emptyset$, where for any $\epsilon > 0$, $B(h^*, \epsilon) \triangleq \{h \in \Omega(z^*) \mid \|h-h^*\| \leq \epsilon\}$;†

(ii) For every $\tilde{h} \in \overset{\circ}{\Omega}(z^*)$ there exists an $\epsilon^* > 0$ such that $\tilde{h} \in \Omega(z)$ for all $z \in \{z \mid \|z-z^*\| \leq \epsilon^*\}$. Then, for every $z^* \in \mathbb{R}^n$ and every $\delta > 0$ there exists an $\hat{\epsilon} > 0$ such that

14. $$\bar{\psi}(z) \leq \bar{\psi}(z^*) + \delta \quad \text{for all} \quad z \in \{z \mid \|z-z^*\| \leq \hat{\epsilon}\}$$

that is, $\bar{\psi}(\cdot)$ is lower semi-continuous.

Proof: Let $z^* \in \mathbb{R}^n$ be arbitrary and suppose that $\bar{\psi}(z^*) = \psi(h^*,z^*)$, $h^* \in \Omega(z^*)$. Since $\psi(\cdot,\cdot)$ is continuous, for any $\delta > 0$ there exists an $\tilde{\epsilon} > 0$ such that

15. $$\psi(h,z) \leq \bar{\psi}(z^*) + \delta$$

for all $\|z-z^*\| \leq \tilde{\epsilon}$, $\|h-h^*\| \leq \tilde{\epsilon}$. Now choose an $\tilde{h} \in \overset{\circ}{B}(h^*, \tilde{\epsilon})$, then, by hypothesis, there is an $\epsilon^* > 0$ such that $\tilde{h} \in \Omega(z)$ for all $\|z-z^*\| \leq \epsilon^*$. Let $\hat{\epsilon} = \min\{\tilde{\epsilon}, \epsilon^*\}$. Then $\tilde{h} \in \Omega(z)$ for all $\|z-z^*\| \leq \hat{\epsilon}$ and

16. $$\bar{\psi}(z) \leq \psi(z,\tilde{h}) \leq \bar{\psi}(z^*) + \delta \quad \text{for all} \quad z \in \{z \mid \|z-z^*\| \leq \hat{\epsilon}\}.$$

This completes our proof.

---

†We denote the interior of a set A by $\overset{\circ}{A}$.

We now return to our functions $\varphi^5(\cdot)$ and $\varphi^6_\varepsilon(\cdot)$.

17. Corollary: Suppose that for every $z' \in \Omega$ there exists a vector $h' \in \mathbb{R}^n$ such that $\langle \nabla f^i(z'), h' \rangle < 0$ for all $i \neq 0$, $i \in J_0(z')$. Then the convex set $\{h \mid f^i(z') + \langle \nabla f^i(z'), h \rangle \leq 0,\ i = 1, 2, \ldots, m\}$ has an interior and the function $\varphi^5(\cdot)$ is lower semi-continuous.

18. Corollary: Suppose that for some $\varepsilon_0 > 0$ and for every $z' \in \Omega$ there exists a vector $h' \in \mathbb{R}^n$ such that $\langle \nabla f^i(z'), h' \rangle < 0$ for all $i \neq 0$, $i \in J_{\varepsilon_0}(z')$. Then the convex set $\{h \mid \langle \nabla f^i(z'), h \rangle + \varepsilon \leq 0,\ i \neq 0,\ i \in J_\varepsilon(z')\}$ has an interior for all $\varepsilon \in [0, \varepsilon_0]$, and for any $z^* \in \Omega$ the function $\tilde{m}(z) \triangleq \min \{\langle \nabla f^0(z), h \rangle + \langle h, H_0 h \rangle \mid \langle \nabla f^i(z), h \rangle + \varepsilon \leq 0,\ i \neq 0,\ i \in J_\varepsilon(z^*)\}$ is lower semicontinuous for all $\varepsilon \in [0, \varepsilon_0]$.

Both of these corollaries are easy to establish by showing that the assumptions of Theorem (12) are satisfied and their proof will therefore be omitted.

In the definition of functions $\varphi^5(\cdot)$ and $\varphi^6_\varepsilon(\cdot)$, the vector h is not restricted to a compact set. We shall now show that whenever z lies in a compact set about a $z^* \in \Omega$, the minimizing h, which is used to obtain the value $\varphi^5(z)$ or $\varphi^6_\varepsilon(z)$, also lies in a compact set.

19. Theorem: Suppose that the assumptions stated in Corollary (17) are satisfied. Let $z^* \in \Omega$, let $\delta > 0$ and let $\varepsilon > 0$ be such that

$$\varphi^5(z) \leq \varphi^5(z^*) + \delta \tag{20.}$$

for all $z \in B(z^*, \varepsilon) \triangleq \{z \in \Omega \mid \|z - z^*\| \leq \varepsilon\}$. If $f^0(\cdot)$ is twice continuously differentiable, $H_0(z) = \frac{1}{2}\, \partial^2 f^0(z)/\partial z^2 > 0$ and h satisfies (9), i.e., $\varphi^5(z) = \langle \nabla f^0(z), h \rangle + \langle h, H_0(z) h \rangle$, for $z \in B(z^*, \varepsilon)$, then h is bounded.

Proof: Since $H_0(z) > 0$ and is uniformly continuous on $B(z^*, \varepsilon)$, there exists a $\lambda^1 > 0$ such that $\langle h, H_0 h \rangle \geq \lambda^1 \|h\|^2$ for all $z \in B(z^*, \varepsilon)$ and $h \in \mathbb{R}^n$. Also, since $\nabla f^0(z)$ is uniformly continuous on $B(z^*, \varepsilon)$, there exists a $\lambda^2 > 0$ such that

$|\langle \nabla f^0(z), h\rangle| \leq \lambda^2 \|h\|$ for all $z \in B(z^*, \varepsilon)$ and $h \in \mathbb{R}^n$. Hence

21. $$\bigcup_{z \in B(z^*,\ \varepsilon)} \{h \mid \langle \nabla f^0(z), h\rangle + \langle h, H_0(z)h\rangle \leq \varphi^5(z^*) + \delta\}$$

is a bounded set whose closure is compact and contains every h such that $\varphi^5(z) = \langle \nabla f^0(z), h\rangle + \langle h, H_0(z)h\rangle$, $z \in B(z^*, \varepsilon)$. This completes our proof.

We now state without proof a similar result for $\varphi^6_\varepsilon(\cdot)$.

22. Theorem: Suppose that the assumptions of Corollary (18) are satisfied and that ε_0 is defined as in (18). Let $z^* \in \Omega$, let $\delta > 0$, let $\varepsilon \in [0, \varepsilon_0]$, and let

23. $$\tilde{m}(z) \triangleq \min \{\langle \nabla f^0(z), h\rangle + \langle h, H_0 h\rangle \mid \langle \nabla f^i(z), h\rangle + \varepsilon \leq 0,\ i \neq 0,\ i \in J_\varepsilon(z^*)\}$$

If $f^0(\cdot)$ is twice continuously differentiable, $H_0 = \frac{1}{2} \frac{\partial^2 f^0(z)}{\partial z^2} > 0$ and $\tilde{\varepsilon} > 0$ is such that

24. $$\tilde{m}(z) < \tilde{m}(z^*) + \delta$$

for all $z \in B(z^*, \tilde{\varepsilon})$, then the closure of the set

25. $$\bigcup_{z \in B(z^*,\ \tilde{\varepsilon})} \{h \mid \langle \nabla f^0(z), h\rangle + \langle h, H_0(z)h\rangle \leq \bar{m}(z^*) + \delta\}$$

is compact.

With the above results established the following theorem is readily proved by essentially repeating the proofs of Theorems (2.8) and (3.10). It is therefore stated without proof.

26. Theorem: Suppose that the function $f^0(\cdot)$ is twice continuously differentiable and that $\frac{\partial^2 f^0(z)}{\partial z^2} > 0$ for all $z \in \Omega$.

(i) If the assumptions of Corollary (17) are satisfied, then $\varphi^5(\cdot)$ can be substituted for $\varphi^1(\cdot)$ in Algorithm (2.5) without affecting its convergence properties as stated in (2.8).

(ii) If the assumptions of Corollary (18) are satisfied, then $\varphi_\varepsilon^6(\cdot)$ can be substituted for $\varphi_\varepsilon^2(\cdot)$ in Algorithm (3.7) without affecting its convergence properties, as stated in (3.10).

This concludes our discussion of the methods of feasible directions.

6. Gradient Projection Methods

We conclude this chapter with several gradient projection methods. The first class of such methods to be considered consists of modifications of Rosen's gradient projection method [R1], [P1] and are designed for solving problems of the form

1. $$\text{minimize } \{f^0(z) \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$$

when the constraint set $\Omega = \{z \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$ is a convex polytope with interior and $f^0(\cdot)$ is convex. (Actually, these methods are also applicable to the case when Ω has no interior. However, this requires that one restrict oneself to the linear manifold containing Ω, and this adds to the complexity of the formulas to be derived. Since these are already quite complex, we shall leave to the reader the extension of the results presented below to the case when Ω has no interior.)

For the Rosen type methods we shall assume that the cost function $f^0(\cdot)$ is convex and that the functions $f^i(\cdot)$, $i = 1, 2, \ldots, m$, are of the form

2. $$f^i(z) = \langle f_i, z\rangle - b^i,$$

with $f_i \in \mathbb{R}^n$ and $b^i \in \mathbb{R}^1$. We shall assume that the set $\Omega \triangleq \{z \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$ has an interior.

3. Definition: For every $z \in \Omega$ and $\epsilon \geq 0$, let

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

4. Assumption: We shall suppose that there exists an $\epsilon^* > 0$ such that for every $z \in \Omega$ and $\epsilon \in [0, \epsilon^*]$ the vectors f_i, $i \in I_\epsilon(z)$ are linearly independent. (This assumption can be removed at the expense of increased complexity in the algorithms to be presented, which must then include a scan over all or most possible constructions of a new direction).

5. Definition: For every $\varepsilon \in [0, \varepsilon^*]$ and every $z \in \Omega$ let

6. $$F_{I_\varepsilon(z)} \triangleq (f_i)_{i \in I_\varepsilon(z)}$$

be a matrix whose columns are f_i, $i \in I_\varepsilon(z)$ (ordered linearly on i). Let $P_{I_\varepsilon(z)}$ be the matrix which projects $\mathbb{R}^n$ onto the subspace spanned by the vectors f_i, $i \in I_\varepsilon(z)$, and let $P^{\perp}_{I_\varepsilon(z)}$† be the matrix which projects $\mathbb{R}^n$ onto the subspace orthogonal to all the f_i, $i \in I_\varepsilon(z)$, i.e.,

7. $$P_{I_\varepsilon(z)} = F_{I_\varepsilon(z)} \left(F^T_{I_\varepsilon(z)} F_{I_\varepsilon(z)} \right)^{-1} F^T_{I_\varepsilon(z)}$$††

8. $$P^{\perp}_{I_\varepsilon(z)} = I - P_{I_\varepsilon(z)} \ .$$

(Note that matrices $P_{I_\varepsilon(z)}$, $P^{\perp}_{I_\varepsilon(z)}$ are symmetric and positive semidefinite.)

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

9. $$\nabla f^0(z) = P_{I_\varepsilon(z)} \nabla f^0(z) + P^{\perp}_{I_\varepsilon(z)} \nabla f^0(z) = F_{I_\varepsilon(z)} \xi_\varepsilon(z) + P^{\perp}_{I_\varepsilon(z)} \nabla f^0(z)$$

where

10. $$\xi_\varepsilon(z) = \left(F^T_{I_\varepsilon(z)} F_{I_\varepsilon(z)} \right)^{-1} F^T_{I_\varepsilon(z)} \nabla f^0(z) \ .$$

---

†When $I_\varepsilon(z)$ is empty, we shall assume that $P_{I_\varepsilon(z)}$ is the zero matrix and that $P^{\perp}_{I_\varepsilon(z)}$ is the identity matrix.

††Note that for $\left(F^T_{I_\varepsilon(z)} F_{I_\varepsilon(z)} \right)^{-1}$ to exist assumption (4) must be satisfied.

Thus, when (4) does not hold, one is forced to use combinatorial methods for reducing $I_\varepsilon(z)$.

It now follows directly from (I.2.1) and (I.2.13) that $\hat{z}$ is optimal if and only if

11a. $$P^{\perp}_{I_0(\hat{z})} \nabla f^0(\hat{z}) = 0 \ ,$$

and

11b. $$\xi_0(\hat{z}) \leq 0 \ .$$

We make one more observation before stating an algorithm. Consider the expansion (9) and let $j \in I_\varepsilon(z)$. Then, from (9) (since $P^{\perp}_{I_\varepsilon(z)-j} P^{\perp}_{I_\varepsilon(z)} = P^{\perp}_{I_\varepsilon(z)}$),

12. $$P^{\perp}_{I_\varepsilon(z)-j} \nabla f^0(z) = \xi^j_\varepsilon(z) \, P^{\perp}_{I_\varepsilon(z)-j} f_j + P^{\perp}_{I_\varepsilon(z)} \nabla f^0(z) \ ,$$

and, since (12) is a decomposition into orthogonal components,

13. $$\|P^{\perp}_{I_\varepsilon(z)-j} \nabla f^0(z)\|^2 = (\xi^j_\varepsilon(z))^2 \, \|P^{\perp}_{I_\varepsilon(z)-j} f_j\|^2 + \|P^{\perp}_{I_\varepsilon(z)} \nabla f^0(z)\|^2 .$$

Finally note that

14. $$\langle f_j, P^{\perp}_{I_\varepsilon(z)-j} \nabla f^0(z) \rangle = \xi^j_\varepsilon(z) \, \langle f_j, P^{\perp}_{I_\varepsilon(z)-j} f_j \rangle \ .$$

15. Algorithm: Suppose we are given an $\bar{\varepsilon}_0 \in (0, \varepsilon^*]$, with ε^* as in (4), an $\varepsilon' \in (0, \varepsilon_0)$ and a $z_0 \in \Omega$.

Step 0: Let $z = z_0$

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

Step 2: Compute

16. $$h_\varepsilon(z) = P^{\perp}_{I_\varepsilon(z)} \nabla f^0(z) \ .$$

Step 3: If $\|h_\varepsilon(z)\|^2 > \varepsilon$, set $h(z) = -h_\varepsilon(z)$ and go to Step 6.

If $\|h_\varepsilon(z)\|^2 \leq \varepsilon$ and $\varepsilon \leq \varepsilon'$, compute $h_0(z)$ (as in (16)) and $\xi_0(z)$ (as in (10)).

If $\|h_0(z)\|^2 = 0$ and $\xi_0(z) \leqq 0$, set $z = z$ and stop (z is optimal). Otherwise, set $h(z) = -h_\epsilon(z)$ and go to Step 4.

If $\|h_\epsilon(z)\|^2 \leq \epsilon$ and $\epsilon > \epsilon'$, go to Step 4.

Step 4: Compute $\xi_\epsilon(z)$ (as in (10)).

If $\xi_\epsilon(z) \leqq 0$, set $h(z) = -h_\epsilon(z)$ and go to Step 5.

If $\xi_\epsilon(z) \not\leqq 0$, compute

17. $$\bar{h}_\epsilon(z) = P^{\perp}_{I_\epsilon(z)-j}\, \nabla f^0(z)$$

such that

18. $$\|\bar{h}_\epsilon(z)\| = \max_{\substack{i \in I_\epsilon(z) \\ \xi^i_\epsilon(z) > 0}} \|P^{\perp}_{I_\epsilon(z)-i}\, \nabla f^0(z)\| .$$

Set $h(z) = -\bar{h}_\epsilon(z)$ and go to Step 5.

Step 5: If $\|h(z)\|^2 \leq \epsilon$, set $\epsilon = \epsilon/2$ and go to Step 2.

If $\|h(z)\|^2 > \epsilon$, go to Step 6.

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

19. $$f^0(z + \mu(z)\, h(z)) = \min \{f^0(z + \mu h(z)) \mid \mu \geq 0,\ (z + \mu h(z)) \in \Omega\}$$

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

20. Theorem: Let $z_0, z_1, z_2, \ldots$, be a sequence in Ω constructed by the Algorithm (15), i.e., $z_1, z_2, \ldots$, are the consecutive values assigned to z in Step 7. Then, either $\{z_i\}$ is finite and its last element is optimal, or else $\{z_i\}$ is infinite and every accumulation point of $\{z_i\}$ is optimal. (When $f^0(\cdot)$ is strictly convex, the problem has a unique optimal solution $\hat{z}$ and then $z_i \to \hat{z}$.)

Proof: We shall again make use of Theorem (I.3.1) under the assumption that $T = \Omega$, $A: \Omega \to \Omega$ is defined by the Algorithm (15), $c(\cdot) = -f^0(\cdot)$ and $\hat{z} \in \Omega$ is defined to be desirable if $P^{\perp}_{I_0(\hat{z})} \nabla f^0(\hat{z}) = 0$ and $\xi_0(\hat{z}) \leq 0$. We begin by showing that the characterization (I.3.2) is satisfied. Suppose that z_0 is optimal. Then $h_0(z_0) = 0$ and $\xi_0(z_0) \leq 0$. Now, for any $\varepsilon \in [0,\bar{\varepsilon}_0]$ / $I_\varepsilon(z_0) \supset I_0(z_0)$ and the vectors f_i, $i \in I_\varepsilon(z_0)$ are linearly independent, hence for any $\varepsilon \in [0,\bar{\varepsilon}_0]$,

21. $$\xi_\varepsilon(z_0) = \xi_0(z_0) \leq 0$$

and

22. $$\|h_\varepsilon(z_0)\| = \|h_0(z_0)\| = 0 \ .$$

Consequently, after a finite number of halvings of ε in Step 5, the algorithm will stop in Step 3, resetting z to its original value. This satisfies (I.3.2). By construction, if the algorithm stops and sets $z = z$ in Step 3 then z is optimal. This is the only possible condition for setting $z = z$, since it is not possible to have $\mu(z)\ h(z) = 0$ in Step 7 for the following reasons. First, $h(z) = 0$ cannot occur in Step 7 because of the logic in Step 5. Second, from the results in Section I.4, it follows that if $h(z) \neq 0$, then $\mu(z) \neq 0$, since for all $i \in I_\varepsilon(z)$, $\langle h(z), f_i \rangle \leq 0$ and $\langle \nabla f^0(z), h(z) \rangle = -\|h(z)\|^2 < 0$.

We must now show that (I.3.3) is satisfied, i.e., that if $z_0 \in \Omega$ is not desirable, / then there exists a $\bar{\rho} > 0$ and $\bar{\delta} > 0$ such that

23. $$-f^0(z + \mu(z)\ h(z)) + f^0(z) \geq \bar{\delta}$$

for all $z \in B(z_0, \bar{\rho}) \triangleq \{z \in \Omega \mid \|z - z_0\| \leq \bar{\rho}\}$. Let ε_0 be the last value of $\varepsilon(z_0)$ (i.e., the value of $\varepsilon(z_0)$ used in the calculation of $h(z_0)$ in Step 3 or Step 4 for $z = z_0$). Then, either

24. $$\|h_{\varepsilon_0}(z_0)\|^2 > \varepsilon_0 \ ,$$

or else

25. $$\|\bar{h}_{\varepsilon_0}(z_0)\|^2 > \varepsilon_0 \ .$$

Suppose (24) took place, i.e., that $h(z_0) = -h_{\varepsilon_0}(z_0)$. Then there exists a $\rho' > 0$ such that

26. $$\|P^{\perp}_{I_{\varepsilon_0}(z_0)} \nabla f^0(z)\|^2 \geq \varepsilon_0/2 \quad \text{for all} \quad z \in B(z_0, \rho')$$

Let $\rho'' > 0$ be such that $I_{\varepsilon_0}(z) \subset I_{\varepsilon_0}(z_0)$ for all $z \in B(z_0, \rho'')$ and let $\rho = \min\{\rho', \rho''\}$. Then, for every $z \in B(z_0, \rho)$ and every $\alpha \in [0, \varepsilon_0]$,

27. $$\|P^{\perp}_{I_\alpha(z)} \nabla f^0(z)\|^2 \geq \|P^{\perp}_{I_{\varepsilon_0}(z)} \nabla f^0(z)\|^2 \geq \|P^{\perp}_{I_{\varepsilon_0}(z_0)} \nabla f^0(z)\|^2 \geq \varepsilon_0/2 \ .$$

We therefore conclude that if (24) took place, then for all $z \in B(z_0, \rho)$, the algorithm will use a final value of $\varepsilon(z) \geq \varepsilon_0/2$.

Now suppose that (25) took place, i.e., that $h(z_0) = -\bar{h}_{\varepsilon_0}(z_0)$. Then, either $\|h_{\varepsilon_0}(z_0)\| > 0$ or $\|h_{\varepsilon_0}(z_0)\| = 0$.

Suppose that $\|h_{\varepsilon_0}(z_0)\| = \delta' > 0$. Let $\rho'' > 0$ be such that $I_{\varepsilon_0}(z) \subset I_{\varepsilon_0}(z_0)$ for all $z \in B(z_0, \rho'')$. Then there exists a $\tilde{\rho} \in (0, \rho'']$, such that for every $z \in B(z_0, \tilde{\rho})$ and for every $\alpha \in [0, \varepsilon_0]$,

28. $$\|\bar{h}_\alpha(z)\|^2 > \|P^{\perp}_{I_\alpha(z)} \nabla f^0(z)\|^2 > \|P^{\perp}_{I_{\varepsilon_0}(z)} \nabla f^0(z)\|^2$$
$$\geq \|P^{\perp}_{I_{\varepsilon_0}(z_0)} \nabla f^0(z)\|^2 \geq \delta'/2 \ ,$$

and hence for every $z \in B(z_0, \tilde{\rho})$, the algorithm will set $\varepsilon(z) \geq [\delta'/2] > 0$.†

Now suppose that $\|h_{\varepsilon_0}(z_0)\| = 0$. Then, $\nabla f^0(z_0) = \sum_{i \in I_{\varepsilon_0}(z_0)} \xi^i_{\varepsilon_0}(z_0) f_i$

---

†Suppose k is an integer such that $\bar{\varepsilon}_0/2^{k+1} \leq \delta'/2 \leq \bar{\varepsilon}_0/2^k$, then we define $[\delta'/2] = \bar{\varepsilon}_0/2^{k+1}$.

and in this representation the coefficients $\xi^i_{\varepsilon_0}$ are unique since the f_i, $i \in I_{\varepsilon_0}(z_0)$ are linearly independent. Now let

$$\delta_1 = \min \{ \|P_I^{\perp} \nabla f^0(z_0)\|^2 \mid I \subset I_{\varepsilon_0}(z_0),\ \|P_I^{\perp} \nabla f^0(z_0)\|^2 > 0 \}$$

29. and

$$\delta_2 = \min \{ \max_{\xi^i_{\varepsilon_0}(z_0) > 0} \|P_{I-i}^{\perp} \nabla f^0(z_0)\|^2 \mid I \subset I_{\varepsilon_0}(z_0),\ \|P_I^{\perp} \nabla f^0(z_0)\|^2 = 0 \} .$$

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 $I_{\varepsilon_0}(z) \subset I_{\varepsilon_0}(z_0)$ for all $z \in B(z_0, \rho'')$. Therefore by first considering all possible subsets I of $I_{\varepsilon_0}(z)$ and then all possible subsets I of $I_{\varepsilon_0}(z_0)$ such that $\xi_{\varepsilon_0}(z_0) > 0$ for $i \in I$. There exists a $\hat{\rho} \in (0, \rho'']$ such that for every $z \in B(z_0, \hat{\rho})$ and for every $\alpha \in [0, \varepsilon_0]$, either

$$\|P^{\perp}_{I_\alpha(z)} \nabla f^0(z)\|^2 \geq \delta''/2$$

30. or

$$\max_{\substack{i \in I_\alpha(z) \\ \xi^i_\alpha(z) > 0}} \|P^{\perp}_{I_\alpha(z)-i} \nabla f^0(z)\|^2 \geq \delta''/2 .$$

We therefore conclude that if (25) took place, then for all $z \in B(z_0, \hat{\rho})$, the algorithm will use a final value of $\varepsilon(z) \geq [\delta''/2] > 0$.

Now, for every $z \in B(z_0, \rho)$(or for all $z \in B(z_0, \hat{\rho})$), whichever is appropriate to consider), and for all $i \in I_{\varepsilon(z)}(z)$, we have $\langle f_i, h(z) \rangle \leq 0$ (see (14), (17)), and so, as far as these constraints are concerned, one can displace oneself an arbitrary amount in the direction h(z) from z without constraint violation. Since for every $z \in B(z_0, \rho)$ and for all $i \in I_{\varepsilon(z)}$, $\langle f_i, z \rangle + b_i - \varepsilon(z) \leq - [\delta'/2]$ (or $- [\delta''/2]$ as the case may be) we now conclude

(as in the case of the feasible directions algorithm) that there exists a $\lambda_m > 0$ such that $z + \lambda h(z)/\|h(z)\| \in \Omega$ for all $\lambda \in [0, \lambda_m]$ and $z \in B(z_0, \rho)$.

Next, we note that $\langle \nabla f^0(z), h(z) \rangle \leq -\epsilon_0/2$ (or $- [\delta'/2]$ or $- [\delta''/2]$) for all $z \in B(z_0, \rho)$ (or $z \in B(z_0, \tilde{\rho})$ or $z \in B(z_0, \hat{\rho})$) and that there exists a γ such that $\|h(z)\| \leq \gamma$ for all $z \in B(z_0, \rho)$ (or $z \in B(z_0, \tilde{\rho})$ or $z \in B(z_0, \hat{\rho})$). It now follows from the results presented in Section (I.4) that (23) is satisfied for all $z \in B(z_0, \rho/2)$ (or $z \in B(z_0, \tilde{\rho}/2)$ or $z \in B(z_0, \hat{\rho}/2)$) for some fixed $\bar{\delta} > 0$. This completes our proof.

Since $\langle \nabla f^0(z), h(z) \rangle \leq - \|h(z)\|^2$, one may wish to accelerate the Algorithm (15) by increasing $\|h(z)\|^2$ as much as possible at each step. The following acceleration procedure is very easily seen as not affecting the convergence properties of the Algorithm (15). (To account for it we need to modify the proof of Theorem (20) only very slightly).

Step 1': (Acceleration procedure, to be inserted between Step 1 and Step 2 of (15)):

Compute $\xi_\epsilon(z)$, $h_\epsilon(z)$ $\bar{h}_\epsilon(z)$, (as in (10), (16), (17)).

If $\xi_\epsilon(z) \leq 0$, go to Step 3.

If $\xi_\epsilon(z) \not\leq 0$ and $\|\bar{h}_\epsilon(z)\| \geq 2\|h_\epsilon(z_0)\|$, set $h(z) = \bar{h}_\epsilon(z)$ and go to Step 5.

If $\xi_\epsilon(z) \not\leq 0$ and $\|h_\epsilon(z)\| < 2\|h_\epsilon(z)\|$, go to Step 3.

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

We recall that in the Algorithm (3.7), to obtain a "feasible direction" $h(z)$, we had to solve a minimization problem. In the Algorithm (15) this process was replaced by the computation of a projection operator which, sometimes, may be

easier to calculate. However, algorithm (15) is only applicable to problems with linear inequality constraints. We shall now present a modification of (15) which applies to more general situations.† We shall suppose from now on that all the functions f^i, $i = 0, 1, 2, \ldots, m$ in (1) are convex and that the set $\Omega = \{z \mid f^i(z) \leq 0, i = 0, 1, 2, \ldots, m\}$ has an interior.

32. Assumption: We shall suppose that there exists an $\epsilon^* > 0$ such that for every $\epsilon \in [0, \epsilon^*]$ and $z \in \Omega$, the vectors $\nabla f^i(z)$, $i \in I_\epsilon(z)$ are linearly independent (where $I_\epsilon(z)$ was defined in (3)).

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

33. $$F_{I_\epsilon(z)} = (\nabla f^i(z))_{i \in I_\epsilon(z)}$$

be a matrix whose columns are the $\nabla f^i(z)$, $i \in I_\epsilon(z)$ (ordered linearly on i). The projection matrices $P_{I_\epsilon(z)}$, $P^{\perp}_{I_\epsilon(z)}$ will still be defined by (7) and (8), respectively, with the matrix $F_{I_\epsilon(z)}$ now defined by (33), etc.

34. Algorithm: Suppose we are given an $\bar{\epsilon}_0 \in (0, \epsilon^*]$ with ϵ^* as in (32), an $\epsilon' \in (0, \bar{\epsilon}_0)$ and a $z_0 \in \Omega$.

Step 0: Set $z = z_0$.

Step 1: Set $\epsilon(z) = \bar{\epsilon}_0$. (We shall use the abbreviated notation $\epsilon = \epsilon(z)$).

Step 2: Compute

35. $$h_\epsilon(z) = P^{\perp}_{I_\epsilon(z)} \nabla f^0(z) \quad .$$

Step 3: If $\|h_\epsilon(z)\|^2 > \epsilon$, set $h(z) = -h_\epsilon(z)$ and go to Step 6.

If $\|h_\epsilon(z)\|^2 \leq \epsilon$, and $\epsilon \leq \epsilon'$, compute $h_0(z_0)$ (with $\epsilon = 0$ as in (35)) and $\xi_0(z)$ (as in (10)).

---

†The prototype of Algorithm (34) was published without proof of convergence in [K1], while the form (34) together with the proof of convergence was published in [P1].

If $h_0(z) = 0$ and $\xi_0(z) \leqq 0$, set $z = z$ and stop. (z is optimal). Otherwise set $h(z) = -h_\varepsilon(z)$ and go to Step 4.

If $\|h_\varepsilon(z)\|^2 \leq \varepsilon$ and $\varepsilon > \varepsilon'$, go to Step 4.

Step 4: Compute $\xi_\varepsilon(z)$.

If $\xi_\varepsilon(z) \leqq 0$, set $h(z) = -h_\varepsilon(z)$ and go to Step 5.

If $\xi_\varepsilon(z) \not\leqq 0$, compute

36. $$\bar{h}_\varepsilon(z) = P^{\perp}_{I_\varepsilon(z_0)-j} \nabla f^0(z)$$

such that

37. $$\|\bar{h}_\varepsilon(z)\| = \max_{\substack{i \in I_\varepsilon(z) \\ \xi^i_\varepsilon(z) > 0}} \|P^{\perp}_{I_\varepsilon(z)-i} \nabla f^0(z)\| .$$

Set $h(z) = -\bar{h}_\varepsilon(z)$ and go to Step 5.

Step 5: If $\|h(z)\|^2 \leq \varepsilon$, set $\varepsilon = \varepsilon/2$ and go to Step 2.

If $\|h(z)\|^2 > \varepsilon$, go to Step 6.

Step 6: Set $K_\varepsilon(z) = I_\varepsilon(z)$ when $h(z) = -h_\varepsilon(z)$ and set $K_\varepsilon(z) = I_\varepsilon(z)-j^\dagger$ when $h(z) = -\bar{h}_\varepsilon(z)$. Compute

38. $$v(z) \triangleq \beta(z)\, h(z) + F_{K_\varepsilon(z)} \left[F^T_{K_\varepsilon(z)} F_{K_\varepsilon(z)}\right]^{-1} t$$

where $t = -\varepsilon(1, 1, \ldots, 1)$ and $\beta(z) \geq 1$ is the smallest positive scalar in $[1,\infty)$ such that

39. $$\langle \nabla f^k(z),\, v(z) \rangle \leq -\varepsilon$$

for $k = 0$ when $h(z) = -h_\varepsilon(z)$ and for $k = 0, j^\dagger$ when $h(z) = -\bar{h}_\varepsilon(z)$.

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

40. $$\lambda(z) = \max \{\lambda \mid f^i(z + \zeta v(z)) \leq 0,\ \zeta \in [0, \lambda],\ i = 1, 2, \ldots, m\} .$$

---

†The index j is the one which was used to construct $\bar{h}_\varepsilon(z)$, i.e., $\bar{h}_\varepsilon(z) = P^{\perp}_{I_\varepsilon(z)-j} \nabla f^0(z)$.

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

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

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

42. Remark: Note that the above algorithm differs from the Algorithm (15) only in the operations defined in Step 6.

43. Theorem: Let $z_0, z_1, z_2, \ldots,$ be a sequence in Ω constructed by the Algorithm (34), i.e., $z_1, z_2, \ldots,$ are the consecutive values assigned to z in Step 7. Then either $\{z_i\}$ is finite and its last element is optimal, or else $\{z_i\}$ is infinite and every accumulation point of $\{z_i\}$ is optimal. (When either $f^0(\cdot)$ 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 $\{z_i\}$, when infinite).

Proof: Again, we shall simply show that the assumptions of Theorem (I.3.1) are satisfied for $c(\cdot) = -f^0(\cdot)$, $T = \Omega$, $a(\cdot)$ defined by (34), and $\hat{z}$ defined to be desirable if $P^{\perp}_{I_0(\hat{z})} \nabla f^0(\hat{z}) = 0$ and $\xi(\hat{z}) \leqq 0$. We omit a demonstration that condition (I.3.2) is satisfied since in this case it is identical to the one given for Algorithm (15) in the proof of Theorem (20).

We shall now show that for every non-optimal $z_0 \in \Omega$, there exist a $\bar{\rho} > 0$ and a $\bar{\delta} > 0$ such that

44. $$- (f^0(z + \mu(z)\ v(z)) + f^0(z)) \geq \bar{\delta} \qquad \text{for all} \qquad z \in B(z_0, \bar{\rho}) .$$

First, proceeding exactly as in the proof of Theorem (20), and, in addition, using the fact that the $f^i(\cdot)$ are continuously differentiable, we can show that if $z_0 \in \Omega$ is not optimal, then there exists a $\rho > 0$ and a $\delta > 0$ such that for all $z \in B(z_0, \rho)$

45. $$\|h(z)\| \geq \delta/2 > 0 ,$$

i.e., $\varepsilon(z) \geq [\delta/2]$, for all $z \in B(z_0, \rho)$. Next, we find that, by (39), for all $z \in B(z_0, \rho)$

46. $$\langle \nabla f^0(z), v(z) \rangle \leq -\varepsilon(z) \leq -[\delta/2]$$

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

47. $$\langle \nabla f^i(z), v(z) \rangle \leq -\varepsilon(z) \leq -[\delta/2] \ .$$

Furthermore, by construction, for all $i \in K_{\varepsilon(z)}(z)$, $z \in B(z_0, \rho)$

48. $$\langle \nabla f^j(z), v(z) \rangle = -\varepsilon(z) \leq -[\delta/2] \ .$$

Finally, an inspection of (38), (39), (16) and (17) indicates that there exists a $\bar{\rho} \in (0, \rho]$ and an $M \in (0, \infty)$ such that $\|v(z)\| \leq M$ for all $z \in B(z_0, \bar{\rho})$. Making use of the results in Section I.4, in a manner similar to the one used in the proof of (3.10), we can now readily show that for all $z \in B(z_0, \bar{\rho})$ there exists a $\bar{\delta} > 0$ such that (44) is satisfied.

49. Remark: The acceleration Step 1' proposed for Algorithm (15) can also be utilized in the present algorithm.

IV. CONVEX OPTIMAL CONTROL PROBLEMS

1. A Further Extension of the Methods of Feasible Directions

This Chapter will be devoted to discrete optimal control problems which transcribe into mathematical programming problems of the form

1. $$\text{minimize } \{f^0(z) \mid f^i(z) \leqq 0,\ i = 1, 2, \ldots, m,\ Rz - b = 0\}$$

where the f^i, $i = 0, 1, \ldots, m$, are convex, continuously differentiable functions such that the set $\Omega = \{z \mid f^i(z) \leq 0,\ i = 1, 2, \ldots, m\}$ has an interior, R is an $\ell \times n$ matrix and $b \in \mathbb{R}^\ell$. Examining (I.1.13), (I.1.14), (I.1.15), we find that the above may hold, if the dynamics of the system (I.1.2) are linear, the cost functions $f_i^0(\cdot,\cdot)$ in (I.1.3) are convex both in the control and in the state variables, and all the inequality constraints are convex and all the equality constraints are affine.

The methods of feasible directions presented in Section III.3 can also be used for solving (1), after a minor modification which is the consequence of the fact that $\hat{z}$ is optimal for (1) if and only if

2. $$\min_{h \in S'} \max_{i \in J_0(\hat{z})} \langle \nabla f^i(\hat{z}), h \rangle = 0$$

where $J_0(\hat{z})$ is defined as in (I.2.8) and $S' = S \cap N$, where S is any set in $\mathbb{R}^n$ containing the origin in its interior and $N = \{z \mid Rz = 0\}$ is the null space of R. The condition (2) can be obtained as a trivial extension of the results presented in Section I.2.

Thus, all we need to do to apply the method of feasible directions (III.3.7) (or its modifications which were discussed in Section III.3)) to (1) is to change the definition of $\bar{\varphi}_\epsilon(z)$ as follows,

3. $$\bar{\varphi}_\varepsilon(z) = \min_{h \in S'} \max_{i \in J_\varepsilon(z)} \langle \nabla f^i(z), h \rangle$$

Where $J_\varepsilon(z)$ is defined as in (I.2.8). Typically, to compute $\bar{\varphi}_\varepsilon(z)$, we now solve (with $S = \{h \mid |h^i| \leq 1\}$)

4a. $$\min \sigma$$

subject to

4b. $$\langle \nabla f^i(z), h \rangle - \sigma \leq 0, \quad i \in J_\varepsilon(z) \ ,$$

4c. $$Rh = 0$$

4d. $$|h^i| \leq 1, \quad i = 1, 2, \ldots, n \ .$$

As an example of an optimal control problem in this class, consider the problem,

5a. $$\min \ \frac{1}{2} \sum_{i=0}^{k-1} (\|x_i - x_i^*\|^2 + u_i^2) \quad x_i \in \mathbb{R}^\nu, \ u_i \in \mathbb{R}^1$$

subject to

5b. $$x_{i+1} = Ax_i + bu_i, \quad i = 0, 1, \ldots, k-1$$

5c. $$x_0 = x_0^*; \quad Cx_k - d = 0; \quad q^j(x_k) \leq 0, \quad j = 1, 2, \ldots, m,$$

$$|u_i| \leq 1, \quad i = 1, 2, \ldots, k-1 \ ;$$

where the q^j: $R^n \to R^1$ are convex functions, C is a $\ell \times \nu$ matrix and $d \in \mathbb{R}^\ell$.

Setting $z = (u_0, u_1, \ldots, u_{k-1})$, we find that $x_i(z)$, the solution of (5b) for $x_0 = x_0^*$ and $(u_0, u_1, \ldots, u_{k-1}) = z$, is given by

6. $$x_i(z) = A^i x_0^* + \sum_{j=0}^{i-1} A^{i-j-1} b u_j$$

(Note that $\frac{\partial x_i(z)}{\partial u_j} = A^{i-j-1} b$). Problem (5) therefore becomes,

7a. $$\text{minimize } \frac{1}{2} \sum_{i=0}^{k-1} (\|x_i(z) - x_i^*\|^2 + u_i^2)$$

subject to

7b. $$|u_i| \leq 1, \; i = 0, 1, 2, \ldots, k-1 \; .$$

7c. $$Cx_k(z) - d = 0, \; q^j(x_k(z)) \leq 0, \; j = 1, 2, \ldots, m \; .$$

For $i = 0, 1, 2, \ldots, k-1$, let R_i be a $\nu \times k$ matrix whose j^{th} column is r_{ij}, with $r_{i(j+1)} \triangleq A^{i-j-1} b$ for $j = 0, 1, \ldots, i-1$, and $r_{ij} \triangleq 0$ for $j = i+1, i+2, \ldots, k$. Then, by (6), $x_i(z) = A^i x_0^* + R_i z$, and hence (7) can be rewritten as

8a. $$\min f^0(z) \triangleq \left(\frac{1}{2} \sum_{i=0}^{k-1} \|A^i x_0 - x_i^* + R_i z\|^2 + \frac{1}{2} \langle z, z \rangle \right)$$

subject to

8b. $$C(A^k x_0^* + R_k z) - d = 0, \; f^j(z) \triangleq q^j(A^k x_0^* + R_k z) \leq 0,$$

$$j = 1, 2, \ldots, m \; ,$$

8c. $$|u_i| \leq 1, \; i = 0, 1, \ldots, k-1 \; ,$$

i.e., $f^0(z)$ is convex, the inequality constraints $f^j(z)$ are convex, and the equality constraints are affine.

2. Decomposition Algorithms

We shall now present two dual methods for a class of optimal control problems. Dual methods differ from primal methods, such as the methods of feasible directions, in that they depend on the optimality condition (I.2.1) rather than on the equivalent form (I.2.7). Dual methods iterate not only on the vector z but also on the multiplier vectors μ and ψ in order to find a set of vectors which satisfy (I.2.1). As we shall see, the dual methods to be presented decompose an optimal control problem into a sequence (usually infinite) of considerably easier subproblems.

A typical example of an optimal control problem which is particularly suitable for solution by dual methods is the following one.

1a. $$\min \frac{1}{2} \sum_{i=0}^{k-1} (\|x_i - x_i^*\|_P^2 + u_i^2), \quad x_i \in \mathbb{R}^\nu, \; u_i \in \mathbb{R}^1$$

subject to

1b. $$x_{i+1} = Ax_i + bu_i, \quad i = 0, 1, 2, \ldots, k-1 \; ,$$

1c. $$x_0 = x_0^*, \; q(x_k) \triangleq \frac{1}{2} \|x_k - x_k^*\|_Q^2 - \gamma \leq 0, \quad \gamma > 0;$$

1d. $$|u_i| \leq 1 \quad \text{for } i = 0, 1, 2, \ldots, k-1 \; .$$

P is a $\nu \times \nu$ symmetric positive semi-definite matrix and $\|x\|_P^2 = \langle x, Px \rangle$, and Q is a $\nu \times \nu$ symmetric positive definite matrix.

Now let $z = (u_0, u_1, \ldots, u_{k-1})$ then for $x_0 = x_0^*$,

2. $$x_i = d_i + R_i z \; , \qquad i = 1, 2, \ldots, k$$

where $d_i = A^i x_0^*$ and R_i is a $\nu \times k$ matrix whose j^{th} column is r_{ij}: $r_{i(j+1)} = A^{i-j-1}b$, for $j = 0, 1, \ldots, i-1$ and $r_{ij} = 0$, for $j = i+1, i+2, \ldots, k$.

Let $C(\alpha) \subset \mathbb{R}^\nu$ be the set of states reachable by (1b) at time k, from x_0^*, with cost not exceeding α ($\alpha \geq 0$) and using controls satisfying $|u_i| \leq 1$, i.e.,

$$3. \quad C(\alpha) = \left\{ x \in \mathbb{R}^\nu \mid x = d_k + R_k z, \ |u_j| < 1; \ \frac{1}{2} \sum_{i=0}^{k-1} \left(\| d_i / {}^{-x_i^*} + R_i z \|_p^2 + u_i^2 \right) \leq \alpha \right\}$$

Since the set $\Omega \triangleq \{x \mid q(x) \leq 0\}$ is compact and strictly convex and since the set $C(\alpha)$ is also convex, we can view problem (1) as that of finding an $\hat{\alpha} \geq 0$, and a control sequence $\hat{z} = (\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{k-1})$ such that

$$4a. \quad \hat{\alpha} = \min \{\alpha \geq 0 \mid C(\alpha) \cap \Omega \neq \emptyset\},$$

$$4b. \quad \hat{x}_k \triangleq (d_k + R_k \hat{z}) \in C(\hat{\alpha}) \cap \Omega .$$

Since the set Ω is strictly convex, $C(\hat{\alpha}) \cap \Omega$ consists of exactly one point, $\hat{x}_k$.

As we shall later see, the optimal control sequence $\hat{z} = (\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{k-1})$ for the optimal control problem (1) is easy to compute if we first determine the $\hat{\alpha}$ satisfying (4a), the $\hat{x}_k$ satisfying (4b), and a unit vector $\hat{s} \in \mathbb{R}^\nu$ which is normal to a hyperplane separating Ω from $C(\hat{\alpha})$. Consequently, Problem (1) may be considered to be a particular case of the geometric problem which we shall now state.

In order to motivate the various assumptions which we are about to make, we note that for every $\alpha \geq 0$, in our example problem (1),

$$5. \quad C(\alpha) = \{x = d_k + R_k z \mid z = (u_0, u_1, \ldots, u_{k-1}), \ |u_j| \leq 1\} \cap \{x = d_k + R_k z \mid z = (u_0, u_1, \ldots, u_{k-1}), \ \frac{1}{2} \sum_{i=0}^{k-1} \| d_i / {}^{-x_i^*} + R_i z \|_p^2 + u_i^2 \leq \alpha\}$$

i.e., $C(\alpha)$ is the intersection of a fixed convex polytope K and of a hyperlipsoid $\Sigma(\alpha)$ which grows monotonically and continuously with α.

6. The Geometric Problem: We are given a map $C(\cdot)$ from $\mathbb{R}^+$ into the set of all subsets of $\mathbb{R}^\nu$ such that

(i) for every $\alpha \geq 0$, $C(\alpha)$ is a compact, convex set which has an interior for every $\alpha > 0$;

(ii) $C(\cdot)$ is continuous in the Hausdorff metric;†

(iii) For every $\alpha \geq 0$, $C(\alpha) = \Sigma(\alpha) \cap K$, where K is a convex polytope with interior, and, for every $\alpha > 0$, $\Sigma(\alpha)$ is a strictly convex set. We also assume that if $0 \leq \alpha_1 < \alpha_2$, then $\Sigma(\alpha_1)$ is contained in the interior of $\Sigma(\alpha_2)$.

We are also given a compact set Ω, which either consists of a unique point or else is strictly convex, and are required to find an $\hat{\alpha} \geq 0$, a vector $\hat{x} \in \Omega$, and a unit vector $\hat{s} \in \mathbb{R}^\nu$ such that

7a. $$\hat{\alpha} = \min \{\alpha \mid C(\alpha) \cap \Omega \neq \emptyset, \alpha \geq 0\};$$

7b. $$\{\hat{x}\} = C(\hat{\alpha}) \cap \Omega ;$$

7c. $$\langle x - \hat{x}, \hat{s} \rangle \leq 0 \quad \text{for all } x \in \Omega ;$$

7d. $$\langle x - \hat{x}, \hat{s} \rangle \geq 0 \quad \text{for all } x \in C(\hat{\alpha}) .$$

8. Assumption: To ensure the existence of a solution and to avoid having to discuss the degenerate case when $\Omega \cap K$ consists of a single point, we shall assume that for some $\alpha \in (0,\infty)$, Ω has points in $\Sigma(\alpha) \cap \overset{o}{K}$, where $\overset{o}{K}$ is the interior of the polytope K.

---

†Given two compact sets A, B in $\mathbb{R}^n$ d(A,B) the Hansdorff distance between these two sets, is defined by $d = \max \{d_1, d_2\}$, where $d_1 = \max_{x \in A} \min_{y \in B} \|x-y\|$ and $d_2 = \max_{y \in B} \min_{x \in A} \|x-y\|$.

9. Proposition: Let $\hat{\alpha}$ be defined by (1), then for every $0 \leq \alpha_1 < \alpha_2 \leq \hat{\alpha}$, $C(\alpha_1) \neq C(\alpha_2)$.

Proof: If for any $\alpha_1, \alpha_2 \in [0,\hat{\alpha}]$, $\alpha_1 \neq \alpha_2$, $C(\alpha_1) = C(\alpha_2)$, then because of (iii) in (6), $C(\alpha_2) = K$, which is impossible, since $C(\alpha_2) \subset C(\hat{\alpha})$ and $C(\hat{\alpha}) \neq K$ by (8).

10. Definition: Let $S = \{s \in \mathbb{R}^\nu \mid \|s\| = 1\}$, and let $v: S \to \Omega$ be the map defined by

11. $$\langle x - v(s), s\rangle \leq 0 \quad \text{for all } x \in \Omega$$

(Note that $v(\cdot)$ is a continuous map.)

12. Definition: Let $\hat{\alpha}$ be defined as in (7b). We shall say that a vector $\hat{s} \in S$ is optimal for the problem (6) if $\hat{x} \triangleq v(\hat{s})$ satisfies (7a), and $\hat{s}$ together with $\hat{x} \triangleq v(\hat{s})$ satisfy (7c) and (7d), i.e., $\{v(\hat{s})\} = C(\hat{\alpha}) \cap \Omega$ and

$$\langle x - v(\hat{s}), \hat{s}\rangle \leq 0 \quad \text{for all } x \in \Omega$$

$$\langle x - v(\hat{s}), \hat{s}\rangle \geq 0 \quad \text{for all } x \in C(\hat{\alpha})$$

Thus, we say that $\hat{s} \in S$ is optimal if it defines a hyperplane which separates Ω from $C(\hat{\alpha})$.

To define an algorithm, we shall need the following sets and maps.

13. Definition: For every $s \in S$ let $P(s)$ denote the hyperplane

$$P(s) = \{x \in \mathbb{R}^\nu \mid \langle x - v(s), s\rangle = 0\}$$

(Note that $P(s)$ is a support hyperplane to Ω at $v(s)$, with outward normal s.)

14. Definition: Let $T \subset S$ be defined by

$$T = \{s \in S \mid \langle x - v(s), s\rangle \geq 0 \quad \text{for all } x \in C(0)\},$$

i.e., if $s \in T$ then $P(s)$ separates $C(0)$ from Ω. It is not difficult to see that if $\hat{s} \in S$ is optimal, then $\hat{s} \in T$, and therefore we can restrict our search for an optimal $\hat{s}$ to the subset T of S.

15. Definition: Let $c: T \to \mathbb{R}^1$ be defined by

$$c(s) = \min \{\alpha \mid \mathcal{C}(\alpha) \cap P(s) \neq \emptyset, \quad \alpha \geq 0\}$$

Note that $c(\cdot)$ is well defined. For suppose that $\mathcal{C}(\alpha) \cap P(s) = \emptyset$ for all $\alpha \geq 0$. Then, since $\mathcal{C}(0) \subset \mathcal{C}(\alpha)$ for all $\alpha \geq 0$, $P(s)$ must separate $\mathcal{C}(\alpha)$, from Ω, for all $\alpha \geq 0$, in contradiction of our assumption that Ω has points in $\mathcal{C}(\alpha)$ for some $\alpha > 0$.

16. Definition: Let $w: T \to \mathbb{R}^\nu$ be defined by

$$\{w(s)\} = \mathcal{C}(c(s)) \cap P(s)$$

We have already concluded that for every $s \in T$, $c(s)$ is well defined and hence $\mathcal{C}(c(s)) \cap P(s) \neq \emptyset$. Now suppose that for some $s \in T$, $\mathcal{C}(c(s)) \cap P(s)$ contains two points $w_1 \neq w_2$. Then, since it is convex, it must also contain the line segment $\{w \mid \lambda w_1 + (1-\lambda)\, w_2, \lambda \in [0,1]\}$. But $\Sigma(c(s))$ is strictly convex and hence we conclude that $P(s)$ must be a support hyperplane to K. However, this is not possible since Ω has points in the interior of K. Therefore $w(\cdot)$ is well defined.

17. Definition: For any two vectors $x, y \in \mathbb{R}^\nu$, let $\pi(x,y)$ denote the operator which projects $\mathbb{R}^\nu$, orthogonally, onto the subspace spanned by x, y. Let $p: \mathbb{R}^\nu \times \mathbb{R}^\nu \to \mathbb{R}^1$ be defined by

17a. $$p(x,y) = \min \{\alpha \mid \pi(x,y)(\mathcal{C}(\alpha)) \cap \pi(x,y)(\Omega) \neq \emptyset, \alpha \geq 0\}^\dagger$$

18. Definition: Let $A(\cdot)$ be a search function from T into the set of all subsets of T, defined by

18a. $$A(s) \subset \sigma(s) \triangleq \{s' \in T \mid s' = \lambda s + \mu(w(s) - v(s)), \lambda, \mu \in (-\infty,+\infty)\}$$

---

†It is not difficult to see that for any two $x,y \in \mathbb{R}^\nu$ which are linearly independent $p(x,y) = \max \{c(s) \mid s \in \tilde{\sigma}(x,y) \triangleq \{s \in T \mid s = \lambda x + \mu y, \lambda, \mu \in (-\infty,+\infty)\}\}$.

18b. $$c(s') = p(s, w(s) - v(s)) .$$

The various functions defined above are illustrated in Fig. 2.

19. A Decomposition Algorithm:†

Step 1: Find a point $s_0 \in T$.

Step 2: Compute a point s' in $A(s_0)$.

Step 3: If $c(s') = c(s_0)$, stop; s_0 is optimal. Otherwise set

$$s_0 = s' \text{ and go to Step 2.}$$

20. Theorem: Let $s_0, s_1, s_2, \ldots,$ be any sequence in T constructed by the Algorithm (19) (i.e., $s_1, s_2, \ldots,$ are the consecutive values assigned to s_0 in Step 3). Then either the sequence $\{s_i\}$ is finite and its last element is optimal, or it is infinite and every cluster point $\hat{s}$ in $\{s_i\}$ is optimal.

Proof: To prove Theorem (20), we shall simply show that the assumptions of Theorem (I.3.16) are satisfied by the maps $A(\cdot)$ and $c(\cdot)$, above, with $s \in T$ defined to be desirable if $c(s') \leq c(s)$ for all $s' \in A(s)$. First, note that (I.3.17) is satisfied by the maps $c(\cdot)$ and $A(\cdot)$, defined in (15), and (18), respectively. Hence, if the sequence $\{s_i\}$ generated by the Algorithm (19) is finite, its last element must be optimal.

We shall now show that the maps $c(\cdot)$ and $A(\cdot)$, under discussion, satisfy (I.3.18). Clearly, to show this it will suffice to show that the maps $c(\cdot)$ and $\bar{c}(\cdot)$ are continuous at all nonoptimal $s \in T$, where $\bar{c}: T \to \mathbb{R}^1$ is defined by

21. $$\bar{c}(s) = p(s, w(s) - v(s))$$

Continuity of $c(\cdot)$: Let s be any point in the interior of T and let δ be any number in $[0, c(s)]$. Then the sets $C(c(s)-\delta)$ and $P(s)$ are strictly

---

†This algorithm has evolved from the work of Krassovskii [K3], Neustadt [N1], Eaton [E1] and Polak and Deparis [P3]. The above version was presented by Polak in [P2].

separated. Let $w' \in P(s)$ and $w'' \in \mathcal{C}(c(s)-\delta)$ be such that

22. $$\|w'-w''\| = \min \{\|x-y\| \mid x \in P(s), \quad y \in \mathcal{C}(c(s)-\delta)\}$$

Let $w = (w' + w'')/2$. Then, by uniform continuity of $\langle \cdot - w, \cdot \rangle$ on $\Omega \times S$, it follows that there exists an $\epsilon' > 0$ such that for all $s' \in T$, satisfying $\|s'-s\| \leq \epsilon'$, the hyperplane $P(w,s') \triangleq \{x \in \mathbb{R}^\nu \mid \langle x - w, s' \rangle = 0\}$ separates $\mathcal{C}(c(s)-\delta)$ from Ω, and hence

23. $$c(s') \geq c(s) - \delta \quad \text{for all} \quad s' \in T, \quad \|s'-s\| \leq \epsilon'$$

Similarly, we can show that there exists an $\epsilon'' > 0$ such that

24. $$c(s') \leq c(s) + \delta \quad \text{for all} \quad s' \in T, \quad \|s'-s\| \leq \epsilon''$$

Let $\epsilon = \min \{\epsilon', \epsilon''\}$, then

25. $$|c(s') - c(s)| \leq \delta \quad \text{for all } s' \in T, \|s'-s\| \leq \epsilon$$

which proves the continuity of $c(\cdot)$ at all points in the interior of T. Since an accumulation point of $\{s_i\}$ cannot be on ∂T, the boundary of T, because $c(s) = 0$ for all $s \in \partial T$ and $s_i \in \{s \in T \mid c(s) > c(s_1) > 0\}$, we need not consider the behavior of $c(\cdot)$ on the boundary of T.

Continuity of $\bar{c}(\cdot)$: First, by an argument similar to the one above, it can be shown that the map $p(\cdot,\cdot)$ defined by (17a) is continuous at every pair of linearly independent vectors (x,y). Now, whenever s is not optimal, the vector $w(s) - v(s) \neq 0$ and is orthogonal to s. Hence, $\bar{c}(\cdot)$ is continuous at every non-optimal $s \in T$ if $w(\cdot)$ is continuous at every non-optimal $s \in T$ (recall that $v(\cdot)$ is continuous on S).

Let $s^* \in T$ be non-optimal and let $\{s_i\}$ be any sequence in T converging to s^*. Then, setting $c_i = c(s_i)$, we have that $c_i \to c^* \triangleq c(s^*)$ and $\mathcal{C}(c_i) \to \mathcal{C}(c^*)$, by continuity of $c(\cdot)$ and of $\mathcal{C}(\cdot)$. Now, let w^* be an accumulation point of

$\{w(s_i)\}$, i.e., $w(s_i) \to w^*$ for $i \in K \subset \{0, 1, 2, \ldots\}$. Then $w(s_i) \in C(c_i)$, and therefore $w^* \in C(c^*)$. Also, since $w(s_i) \in P(s_i)$,

26. $$\langle w(s_i) - v(s_i), s_i \rangle = 0 \quad \text{for} \quad i = 0, 1, 2, \ldots$$

Consequently, since $s_i \to s^*$, $v(s_i) \to v(s^*)$, and $w(s_i) \to w^*$ for $i \in K$, we must have $\langle w^* - v(s^*), s^* \rangle = 0$, i.e., $w^* \in P(s^*)$. Thus,

27. $$w^* \in C(c(s^*)) \cap P(s^*)$$

But $C(c^*) \cap P(s^*)$ consists of only one point $w(s^*)$. Consequently, $w^* = w(s^*)$ and $w(\cdot)$ is continuous at s^*. This completes our proof.

We shall now see what is involved in applying Algorithm (19) to the problem (1).† First, given a vector $s \in S$, we compute $v(s)$ from the fact that

28. $$\nabla q(v(s)) = \lambda s, \quad \lambda > 0 .$$

Thus,

29. $$\nabla q(v(s)) = Q(v(s) - x_k^*) = \lambda s, \quad \lambda > 0$$

Hence $(v(s) - x_k^*) = \lambda Q^{-1} s$ and we therefore compute $\lambda > 0$ from

30. $$\frac{1}{2} \langle \lambda Q^{-1} s, \lambda s \rangle - \gamma = 0 ,$$

i.e., $\lambda = +(2\gamma / \langle s, Q^{-1} s \rangle)^{1/2}$. Thus

$$v(s) = x_k^* + (2\gamma / \langle s, Q^{-1} s \rangle)^{1/2} Q^{-1} s ,$$

which presents no serious problems in computing.

Next, to compute a point $s_0 \in T$ we may proceed as follows. From (5), $\{d_k\} = C(0)$ and from (1c) $x_k^* \in \Omega$. Now, let us compute a $\bar{\lambda} \in [0,1]$ such that

---

†Note that when Ω consists of one point only, i.e., $\Omega = \{\hat{x}_k\}$, $v(s) = \hat{x}_k$ for all $s \in T$ and hence presents no difficulties in evaluation.

$q(\bar{\lambda}\, x_k^* + (1-\bar{\lambda})\, d_k) = 0$ by solving the quadratic equation

32. $$\frac{1}{2}\langle (1-\lambda)x_k^* - (1-\lambda)d_k,\ Q((1-\lambda)x_k^* - (1-\lambda)d_k)\rangle - \gamma = 0 \ ,$$

i.e.,

33. $$(1-\lambda)^2 = 2\gamma/x_k^* - d_k\|_Q^2$$

Then $\bar{x} = (\bar{\lambda}\, x_k^* + (1-\bar{\lambda})d_k) \in \partial\Omega$, the boundary of Ω, and $\frac{\nabla q(\bar{x})}{\|\nabla q(\bar{x})\|} \in T$, hence, set $s_0 = \frac{\nabla q(\bar{x})}{\|\nabla q(\bar{x})\|}$.

Now, to compute $c(s)$ and $w(s)$ for a given $s \in T$, we must solve (with $z = (u_0, u_1, \ldots, u_{k-1})$)

34. $$\min \sum_{i=0}^{k-1} \|d_i + R_i z\|_P^2 + u_i^2$$

subject to

34a. $$\langle (d_k + R_k\ z) - v(s),\ s\rangle = 0$$

34b. $$|u_i| \le 1, \quad i = 0, 1, 2, \ldots, k-1.$$

We note that (34) is a quadratic programming problem solvable by finite step procedures (such as Wolfe's [W1]). We also note that a special case of (34) arises when the matrix $P = 0$ in (34). In this case the necessary and sufficient conditions developed in Section I.2 show that the optimal $\hat{z} = (\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{k-1})$ is given by

35. $$\hat{u}_i = \text{sat}\ \langle r_{k(i+1)}, \hat{\lambda} s\rangle \quad i = 0, 1, \ldots, k-1$$

where $r_{k(i+1)} = A^{k-i-1} b$ for $i = 0, 1, \ldots, k-1$, and $\hat{\lambda} > 0$ is easily determined from the following piecewise linear equation.

36. $$\langle d_k + \sum_{i=1}^{k} \text{sat}(\hat{\lambda}\ \langle r_{ki}, s\rangle\ r_{ki} - v(s),\ s\rangle = 0$$

which is obtained from the fact that

37. $$w(s) = \left(d_k + \sum_{i=0}^{k-1} \operatorname{sat}(\hat{\lambda} \langle r_{k(i+1)}, s \rangle)\, r_{k(i+1)} \right) \in P(s) \quad .$$

Then $c(s) = \frac{1}{2} \sum \hat{u}_i^2$.

Thus, so far, we have encountered no difficulties. Now, while it is quite clear that a point $s' \in A(s)$ cannot be computed exactly, it is clear from Theorem (I.3.16) and the discussion which preceeds it, that if we carry out no more than a very coarse search along the arc $\sigma(s)$ for a point which maximizes $c(s')$, $s' \in \sigma(s)$, we should, with reasonable certainity, find a point $s' \in \sigma(s)$ satisfying $c(s') - c(s) \geq \beta(\bar{c}(s) - c(s))$ for some fixed but very small $\beta > 0$, and hence obtain convergence. Experiments carried out by the author bear out this heuristic conclusion.

The problem becomes considerably less tractable when the set Ω is described by several inequalities, since now we can compute neither $v(s)$ nor a point $s' \in A(s)$, in a finite number of steps. Theorem (I.3.21), however, leads us to the following heuristic development which is one of several that are possible. First we must introduce a set to approximate $v(s)$ for $s \in T$. Thus, suppose that

38. $$\Omega = \{x \in \mathbb{R}^\nu \mid q^i(x) \leq 0,\ i = 1, 2, \ldots, m\}$$

where the q^i: $R^\nu \to R^1$ are continuously differentiable, strictly convex functions, and by our assumption Ω is compact and strictly convex (or else $\Omega = \{\hat{x}_k\}$ is a point in which case $v(s) = \hat{x}_k$ for all $s \in T$).

39. Definition: Let $\bar{p}$: $\mathbb{R}^\nu \to \mathbb{R}^1$ be defined by

39a. $$\bar{p}(x) = \sum_{i=0}^{m} (\max\{0, q^i(x)\})^2$$

Let $\alpha \geq 0$, $\beta > 0$ be given scale factors, let $\varepsilon > 0$ and let

39b. $$V_\varepsilon(s) = \{x \in \mathbb{R}^\nu \mid \| \frac{\partial}{\partial x} (\langle x,s\rangle + \frac{1}{\beta\varepsilon} \overline{p}(s)) \|^2 \leq \alpha\varepsilon \}$$

Note that if α is chosen to be zero, then $V_\varepsilon(s)$ contains exactly one point $x_\varepsilon(s)$, which minimizes the strictly convex function $\langle x,s\rangle + \frac{1}{\beta\varepsilon} p(x)$ over $\mathbb{R}^\nu$. From (III.1.10) we see that $x_\varepsilon(s)$ must satisfy $x_\varepsilon(s) \to v(s)$ as $\varepsilon \to 0$ and

40. $$-\langle x_\varepsilon(s),s\rangle \leq -\langle v(s),s\rangle$$

i.e.,

41. $$\langle x_\varepsilon(s) - v(s),s\rangle \geq 0 \quad .$$

Thus, $x_\varepsilon(s)$ is separated from Ω by the hyperplane $P(s)$ passing through $v(s)$ and therefore, if we define the hyperplane,

42. $$P(x,s) = \{x' \mid \langle x'-x,s\rangle = 0\}$$

then Ω lies to one side of $P(x_\varepsilon(s),s)$, i.e., either $\Omega \cap P(x_\varepsilon(s),s) = \emptyset$ or $\Omega \cap P(x_\varepsilon(s),s) = \{v(s)\}$ for all $\varepsilon > 0$, and $s \in S$.

We now extend our functions $c(\cdot)$ and $w(\cdot)$.

43. Definition: Let $U' \subset \mathbb{R}^\nu \times S$ be such that for every $(s,x) \in U'$ there exists an $\alpha \geq 0$ such that $\mathcal{C}(\alpha) \cap P(x,s) \neq \emptyset$. We define the map $\tilde{c}$: $U' \to \mathbb{R}^1$ as

43a. $$\tilde{c}(x,s) = \min \{\alpha \mid \mathcal{C}(\alpha) \cap P(x,s) \neq \emptyset\}$$

44. Definition: Let $U \subset U'$ be such that for every $(x,s) \in U$, $\mathcal{C}(\tilde{c}(x,s)) \cap P(x,s)$ consists of a unique point. Then we define $\tilde{w}$: $U \to \mathbb{R}^\nu$ as

44a. $$\{\tilde{w}(x,s)\} = \mathcal{C}(\tilde{c}(x,s)) \cap P(x,s) \quad .$$

45. Decomposition Algorithm with ϵ-Procedure

Suppose that an $\bar{\epsilon}_0 > 0$, and scale factors $\alpha \geq 0$, $\beta > 0$ for (39b) and an $s_0 \in T$ are given.

Step 0: Set $s = s_0$.

Step 1: Set $\epsilon = \bar{\epsilon}_0$.

Step 2: Compute a point $v_\epsilon(s) \in V_\epsilon(s)$ (as defined in 39b).

Step 3: Compute $\tilde{c}(v_\epsilon(s),s)$, $\tilde{w}(v_\epsilon(s),s)$ and the curve $\tilde{\sigma}(v_\epsilon(s),s)$ which is the intersection of T with the two-dimensional subspace spanned by s and $\tilde{w}(v_\epsilon(s),s) - v_\epsilon(s)$.

Step 4: For each $s' \in \tilde{\sigma}(v_\epsilon(s),s)$, compute a vector $v_\epsilon(s')$ in $V_\epsilon(s')$ and then find a vector $s_1 \in \tilde{\sigma}(v_\epsilon(s),s)$ such that

46. $$\tilde{c}(v_\epsilon(s_1),s_1) = \max \{\tilde{c}(v_\epsilon(s'),s') \mid s' \in \tilde{\sigma}(v_\epsilon(s),s)\}$$

Step 5: If $\tilde{c}(v_\epsilon(s_1),s_1) - \tilde{c}(v_\epsilon(s),s) \geq \epsilon$, set $s = s_1$ and go to Step 1.
If $\tilde{c}(v_\epsilon(s_1),s_1) - \tilde{c}(v_\epsilon(s),s) < \epsilon$, set $\epsilon = \epsilon/2$ and go to Step 2.

47. Theorem: Suppose that $\alpha = 0$. Let $\{s_i\}$ be any infinite sequence in T constructed by the Algorithm (45) (i.e., s_1, s_2,..., are the consecutive values assigned to s_0 in Step 5), then any cluster point of $\{s_i\}$ is optimal.

We omit a proof of this theorem since it can easily be established by using Theorem (I.3.21).

In practice, Algorithm (45) cannot be applied with $\alpha = 0$, since the computation of $v_\epsilon(s) \triangleq x_\epsilon(s)$ cannot, in general, be performed by a finite step procedure. Nor can the point s_1, defined in Step 4, be computed by a finite step procedure. Thus, in practice, one must choose $\alpha > 0$ and use some finite search over the curve $\tilde{\sigma}(v_\epsilon,s)$ for a point s_1. For example, one may examine the points, $(s + \frac{i}{j}\xi)/\|s + \frac{i}{j}\xi\|$, where j is some positive integer, i = 0, 1,..., j, and

$\xi = \rho(\tilde{w}(v_\varepsilon(s),s) - v_\varepsilon(s))$, with $\rho \in (0,1]$. Although with $\alpha > 0$ and an approximate evaluation of s_1 in Step 4, it does not seem possible to establish mathematically the convergence of Algorithm (45), there is a certain amount of experimental evidence to support a claim that the convergence of Algorithm (45) is usually not affected by these approximations.

To conclude this section we shall describe one more algorithm for solving the Geometric Problem (6). This algorithm was first introduced by Frank and Wolfe [F5] due to Barr and Gilbert [B2], [B3] and used as a subprocedure algorithm and rediscovered independently by Gilbert [G1] in conjunction with the solution of optimal control problems.

48. Definition: For every $s \in T$, let $y(s) \in C(c(s))$ be such that

$$\|y(s) - v(s)\| = \min \{\|y - v(s)\| \mid y \in C(c(s))\}$$

49. Proposition: The map function $y: T \to \mathbb{R}^\nu$ defined by (48) is continuous.

50. Barr-Gilbert Algorithm [B2]. Suppose $\Omega = \{v\}$ (i.e., Ω consists of a unique point) and suppose that an $s_0 \in T$ is given.

Step 1: Set $s = s_0$.

Step 2: Compute $c(s)$, $y(s)$.

Step 3: Set $s = (y(s) - v)/\|(y(s) - v)\|$

Since $c(\cdot)$ and $y(\cdot)$ are continuous, it is easy to establish by means of Theorem (I.3.1) that if $\{s_i\}$ is a sequence constructed by the Algorithm (50), then every cluster point of $\{s_i\}$ is optimal, and that $w(s_i) \to v$ and $c(s_i) \to c^*$, with $c^* = \min \{\alpha \mid v \in C(\alpha), \alpha \geq 0\}$.

In order to apply this algorithm to a control problem with Ω bigger than one point, Gilbert introduces new sets $\tilde{C}(\alpha) = C(\alpha) - \Omega$, which can be used instead of $C(\alpha)$, and let $v = 0$ in (50). To see the details of how this is done, as well as how one may approximate the computation of $y(s)$, the reader should look up [G1], [B2], [B3].

This concludes our discussion of decomposition algorithms for optimal control problems.

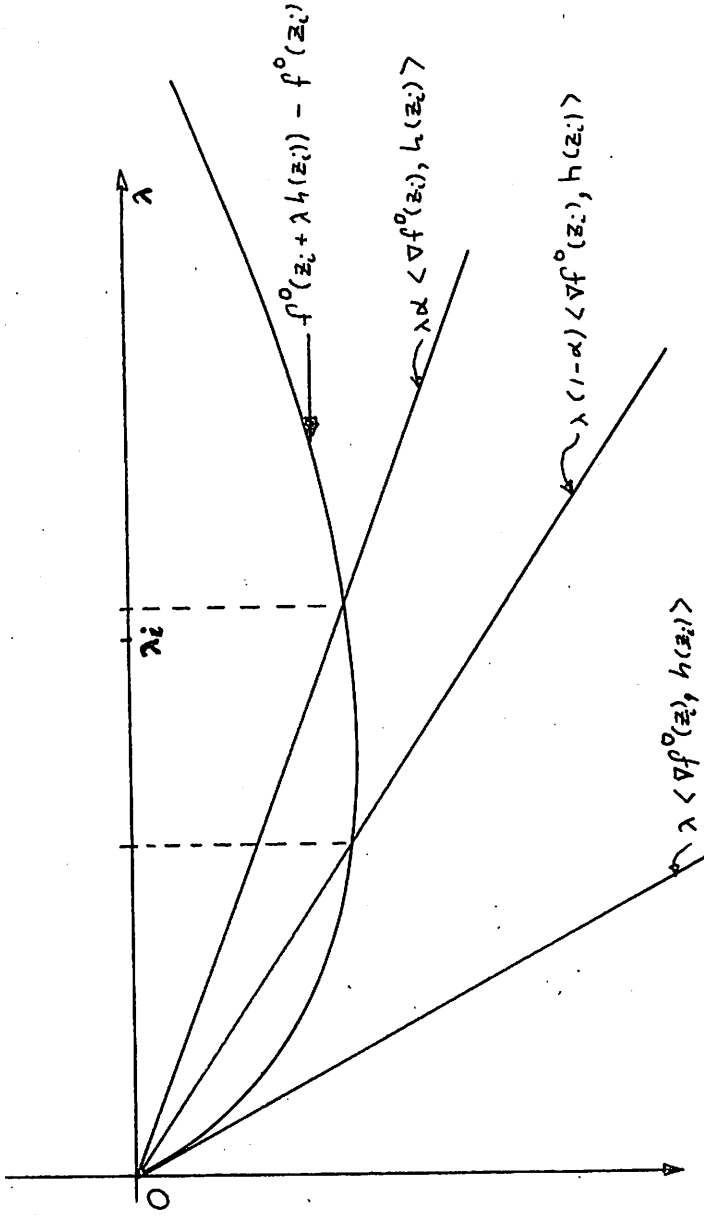

FIG 1. Calculation of Step Size

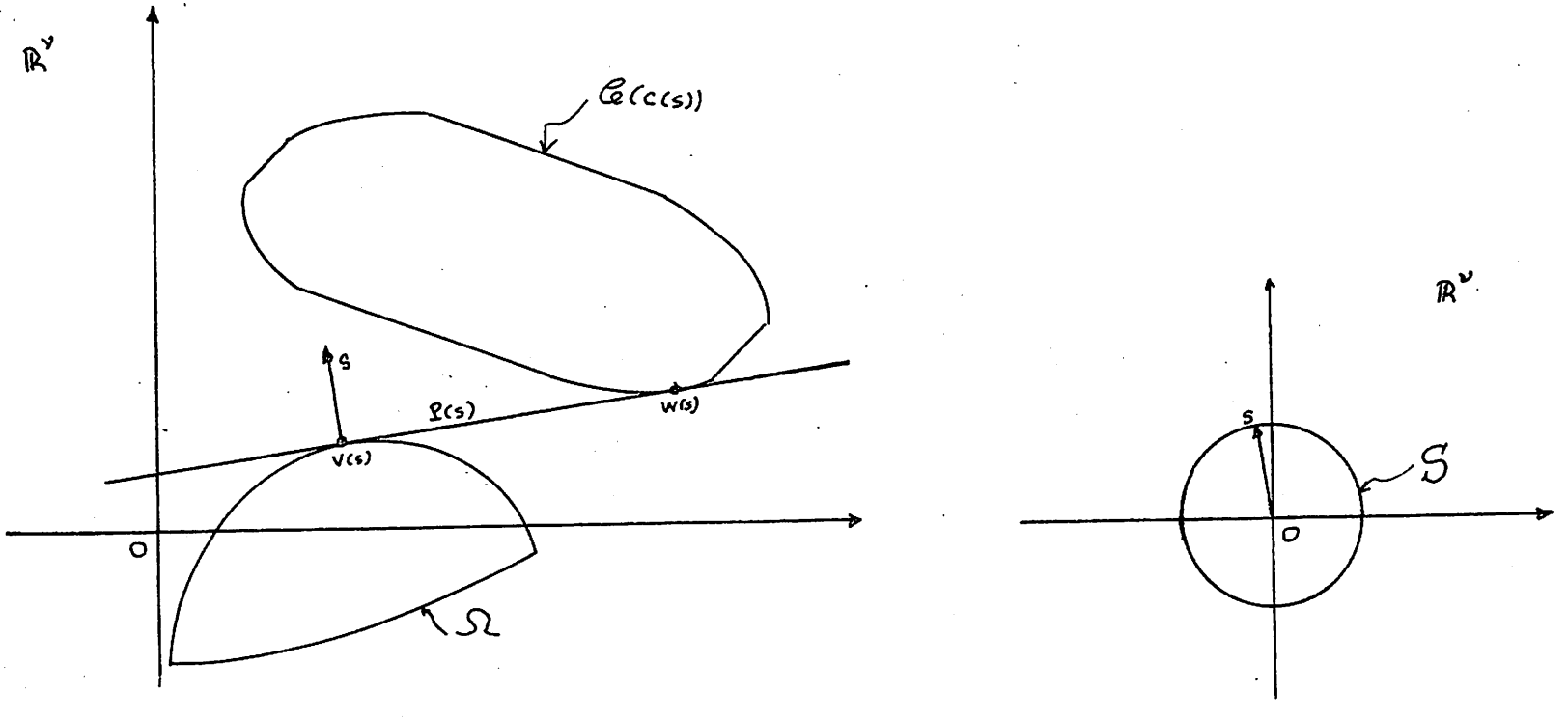

FIG 2. The Geometry of the Problem.

REFERENCES

B1 N. V. Banitchouk, V. M. Petrov and F. L. Chernousko, "Numerical Solution of Problems With Variational Limits by the Method of Local Variations", Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, Vol. 6, No. 6, 1966, pp. 947-961.

B2 R. O. Barr and E. G. Gilbert, "Some Iterative Procedures for Computing Optimal Controls", Proceedings of Third Congress of the International Federation of Automatic Control, London, June 20-25, 1966, Paper No. 24.D.

B3 R. O. Barr, "Computation of Optimal Controls by Quadratic Programming on Convex Reachable Sets", Ph.D. Dissertation, University of Michigan, 1966.

B4 R. Bellman and R. Kalaba, "Quasilinearization and Boundary Value Problems", American Elsevier Publishing Co., New York, 1965.

B5 Bui-Trong-Lieu and P. Huard, "La Methode des Centres Dans un Espace Topologique", Numerische Matematik, Vol. 8, 1966, pp. 56-67.

C1 M. Canon, C. Cullum and E. Polak, "Theory of Optimal Control and Mathematical Programming", McGraw-Hill, 1969, Chapters I-IV.

C2 A. L. Cauchy, "Methode Generale Pour la Resolution des Systemes d'Equations Simultanees", Comptes Rendus, Acad. de Scie. Paris, Vol. XXV, 1847, pp. 536-38.

C3 Y. Cherruault, "Une Methode Directe de Minimization et Applications, Revue Francaise d'Informatique et de Recherche Operationelle, No. 10, July, 1968.

C4 R. Courant, "Variational Methods for the Solution of Problems of Equilibrium and Vibrations", Bull. Amer. Math. Soc., Vol. 49, 1943, pp. 1-23.

C5 H. Curry, "The Method of Steepest Descent for Nonlinear Minimization Problems", Quarterly of Applied Math., II, 1944, pp. 258-61.

D1 J. W. Daniel, "The Conjugate Gradient Method for Linear and Nonlinear Operator Equations", SIAM J. Num. Anal. Vol. 4, No. 1, 1967.

D2 W. C. Davidon, "Variable Metric Methods for Minimization", A.E.C. Research and Development Report ANL 5990 (Rev), 1959.

E1 J. H. Eaton, "An Iterative Solution to Time-Optimal Control", J. Math. Anal. and Appl., Vol. 5, No. 2, 1962, pp. 329-44.

F1 A. V. Fiacco and G. P. McCormack, "The Sequential Unconstrained Minimization Technique for Nonlinear Programming, A Primal-Dual Method", Management Science, Vol. 10, No. 2, 1964, pp. 360-66.

F2 A. V. Fiacco and G. P. McCormack, "Computational Algorithm for the Sequential Unconstrained Minimization Technique for Nonlinear Programming", Management Science, Vol. 10, No. 2, pp. 601-17.

F3 R. Fletcher and M. J. D. Powell, "A Rapidly Convergent Descent Method For Minimization", The Computer Journal, Vol. 6, 1963, p. 163.

F4 R. Fletcher and C. M. Reeves, "Function Minimization by Conjugate Gradients", The Computer Journal, Vol. 7, No. 2, 1964, pp. 149-154.

F5 M. Frank and P. Wolfe, "An Algorithm for Quadratic Programming", Naval Research Logistics Quarterly, Vol. 3, 1956, pp. 95-110.

G1 E. G. Gilbert, "An Iterative Procedure for Computing the Minimum of a Quadratic Form on a Convex Set", SIAM Control, Vol. 4, No. 1, 1966, pp. 61-80.

G2 A. A. Goldstein and J. F. Price, "An Effective Algorithm for Minimization", Numerische Mathematik 10, 1967, pp. 184-189.

H1 M. R. Hestenes and E. Stiefel, "Methods of Conjugate Gradients for Solving Linear Equations", J. of Research of Nat. Bureau of Standards, Vol. 49, 1952, p. 409.

H2 M. R. Hestenes, "The Conjugate Gradient Method for Solving Linear Systems", Proceedings of Symposia on Applied Math. - Vol. VI Numerical Analysis, McGraw-Hill, New York, 1956, pp. 83-102.

H3 P. Huard, "Programmation Matematique Convexe", Revue Francaise d'Informatique et de Recherche Operationelle, No. 7, 1968, pp. 43-59.

J1 F. John, "Extremum Problems With Inequalities as Side Conditions", Courant Anniversary Volume, K. O. Friedrichs, O. E. Neugebauer and J. J. Stoker eds., Interscience, New York, 1948, pp. 187-204.

K1 P. Kalfond, G. Ribiere, J. C. Sogno, "A Method of Feasible Directions Using Projection Operators", Proc. IFIP Congress 68, Edinburgh, August, 1968.

K2 L. V. Kantorovich and G. P. Akilov, "Functional Analysis in Normed Spaces", The MacMillan Co., Publishers, New York, 1964, Chapter 15.

K3 N. N. Krasovskii, "On an Optimal Control Problem", Priklad. Mat. i Mekh., Vol. 21, No. 5, 1957, pp. 670-77.

K4 H. W. Kuhn and A. W. Tucker, "Nonlinear Programming", Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1951, pp. 481-492.

L1 G. S. Lee, "Quasilinearization and Invariant Imbedding, With Applications to Chemical Engineering and Adaptive Control", Academic Press, New York, 1968.

M1 G. P. McCormack and W. I. Zangwill, "A Technique for Calculating Second-Order Optima", Research Analysis Corp. McLean, Virginia (mimeo), 1967.

M2 D. Mayne, "A Second-Order Gradient Method for Determining Optimal Control Trajectories of Nonlinear Discrete-Time Systems", Int. J. Control, Vol. 3, No. 1, 1966, pp. 85-95.

N1 L. W. Neustadt, "Synthesis of Time Optimal Control Systems", J. Math. Anal. and Appl., Vol. 1, 1960, pp. 484-492.

P1 E. Polak, "On The Convergence of Optimization Algorithms", Revue Francaise d'Informatique et de Recherche Operationelle, Serie Rough, No. 16, 1969, pp. 17-34.

P2 E. Polak, "On Primal and Dual Methods for Solving Discrete Optimal Control Problems", Proceedings, Second International Conference on Computing Methods in Optimization Problems, San Remo, Italy, Sept. 9-13, 1968, Academic Press, 1969.

P3 E. Polak and M. Deparis, "An Algorithm for Minimum Energy Control", Univ. of California, Electronics Res. Lab., Berkeley, Memorandum M225, Nov., 1967.

P4 E. Polak and G. Ribiere, "Note sur la Convergence de Methodes de Directions Conjuguees", Revue Fracaise d'Informatique et de Recherche Operationelle, No. 16, 1969.

P5 M. J. D. Powell, "A Survey of Numerical Methods for Unconstrained Optimization", Presented at the SIAM 1968 National Meeting.

R1 J. B. Rosen, "The Gradient Projection Method for Nonlinear Programming, Part I. Linear Constraints", J. SIAM, Vol. 8, No. 1, 1960, pp. 181-217.

T1 D. M. Topkis and A. Veinott, Jr., "On The Convergence of Some Feasible Directions Algorithms for Nonlinear Programming", J. SIAM Control, Vol. 5, No. 2, 1967, pp. 268-79.

Z1 W. I. Zangwill, "Nonlinear Programming: A Unified Approach", Prentice-Hall Inc., Englewood Cliffs, N. J., 1969.

Z2 W. I. Zangwill, "Nonlinear Programming via Penalty Functions", Management Science-A, Vol. 13, No. 5, 1967, pp. 344-58.

Z3 G. Zoutendijk, "Methods of Feasible Directions", Elsevier Publ. Co., Amsterdam, 1960, p. 23.