AD744313

TOPICS IN OPTIMIZATION

BY

MICHAEL OSBORNE

STAN-CS-72-279

APRIL 1972

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

Approved for public release;
Distribution Unlimited

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

DDC
RECEIVED
JUN 3 1972
A

146

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Stanford University	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

Topics in Optimization

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Technical Report -

5. AUTHOR(S) *(First name, middle initial, last name)*

Michael Osborne

6. REPORT DATE April, 1972	7a. TOTAL NO. OF PAGES 141	7b. NO. OF REFS 148
8a. CONTRACT OR GRANT NO ONR N00014-67-A-0112-00029 NR 044-211 b. PROJECT NO c. NSF GJ 29988X d.	9a. ORIGINATOR'S REPORT NUMBER(S) STAN-CS-72- 279 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Office of Naval Research Pasadena Branch

13. ABSTRACT

These notes are based on a course of lectures given at Stanford, and cover three major topics relevant to optimization theory. First an introduction is given to those results in mathematical programming which appear to be most important for the development and analysis of practical algorithms. Next unconstrained optimization problems are considered. The main emphasis is on that subclass of descent methods which (a) requires the evaluation of first derivatives of the objective function, and (b) has a family connection with the conjugate direction methods. Numerical results obtained using a program based on this material are discussed in an Appendix. In the third section, penalty and barrier function methods for mathematical programming problems are studied in some detail, and possible methods for accelerating their convergence indicated.

I

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

Unclassified

Security Classification

Unclassified

Security Classification

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
Mathematical Programming, unconstrained optimization, descent methods, penalty functions barrier functions, rate of convergence						

II

DD FORM 1 NOV 65 1473 (BACK)

(PAGE 2)

Unclassified

Security Classification

TOPICS IN OPTIMIZATION

BY

Michael Osborne

STAN-CS-72-279
April 1972

Computer Science Department
School of Humanities and Sciences
Stanford University

III

TOPICS IN OPTIMIZATION

Michael Osborne
Computer Science Department
Stanford University

and

Australian National University
Canberra, Australia

Abstract

These notes are based on a course of lectures given at Stanford, and cover three major topics relevant to optimization theory. First an introduction is given to those results in mathematical programming which appear to be most important for the development and analysis of practical algorithms. Next unconstrained optimization problems are considered. The main emphasis is on that subclass of descent methods which (a) requires the evaluation of first derivatives of the objective function, and (b) has a family connection with the conjugate direction methods. Numerical results obtained using a program based on this material are discussed in an Appendix. In the third section, penalty and barrier function methods for mathematical programming problems are studied in some detail, and possible methods for accelerating their convergence indicated.

This research was supported in part by the National Science Foundation under grant number 29988X, and the Office of Naval Research under contract number N-00014-67-A-0112-00029 NR 044-211 . Reproduction in whole or in part is permitted for any purpose of the United States Government.

Introduction

These notes were prepared for a course on optimization given in the Computer Science Department at Stanford University during the fall quarter of 1971. In part they are based on lectures given during the year of study in numerical analysis funded by the United Kingdon Science Research Council at the University of Dundee, and on courses given at the Australian National University.

The choice of material has been regulated by limitations of time as well as by personal preference. Also, much material appropriate to the development of algorithms for linearly constrained optimization problems was covered in the parallel course on numerical linear algebra given by Professor Golub. Thus, despite some ambition to cover a larger range, the course eventually consisted of three main sections. These notes cover these sections and have been supplemented by brief additional comments and a list of references. A more extensive bibliography is also included. This is an amended version of a bibliography prepared by my former student Dr. D. M. Ryan.

The first section is intended to provide a solid introduction to the main results in mathematical programming (or at least to those results which appear to be the most important for the development and analysis of practical algorithms). The main aim has been to characterize local extrema, so that convexity and duality theory are not treated in any great detail. However, the material given is more than adequate for the purposes of the remaining sections. Opportunity has been taken to prevent the recent results of Gould and Tolle which provide an accessible and rather complete description of the first order conditions for an extremum. The second order conditions are also considered in detail.

The second section on unconstrained optimization is largely restricted to that subclass of descent methods which (a) requires the evaluation of first derivatives of the objective function, and (b) has some family connection with the so-called conjugate direction methods. This is an area in which there has been considerable recent activity, and here an attempt is made both to summarize significant recent developments and to indicate their algorithmic possibilities. An appendix (prepared with the help of M. A. Saunders) summarizes numerical results obtained with a program based on this material. One significant omission from this section is any detailed discussion of convergence. However, the convergence of certain algorithms (those that reset the Hessian estimate periodically or according to appropriate criteria) is an easy consequence of the material given.

In the third section, penalty and barrier function methods for non-linear programming are considered. This turns out to be a very nice application, in particular, of the results of the first section. These methods have advantages of robustness and simplicity but carry a definite cost penalty. However, attempts to remedy this situation show some promise. The material presented in this section has important connections with other areas: for example, with the method of regularization for the approximate solution of improperly posed problems.

Acknowledgments

The material presented here has benefited greatly from discussions with Roger Fletcher, Gene Golub, John Greenstadt, Shmuel Oren, Mike Powell, and Mike Saunders. The presentation of the course was shaped more by the

determination of the lecturer to cover as large a field as possible than by any consideration for the audience. In spite of this the level of continued interest was most gratifying, and it is a pleasure to single out the credit students Linda Kaufman, John Lewis and Margaret Wright in this respect.

A special vote of thanks is required for George Forsythe who was responsible for the invitation to Stanford, and for completing the subsequent arrangements despite the efforts of the Australian and U.K. Post Offices. In spite of this he was still prepared to combine with John Herriot in keeping the lecturer in reasonable check, and to provide the incentive to produce these notes as a CS report.

Contents

I. Introduction to Mathematical Programming 5

1. Minimum of a constrained function 6

2. Some properties of linear inequalities 12

3. Multiplier relations 14

4. Second order conditions 20

5. Convex programming problems 25

II. Descent Methods for Unconstrained Optimization 36

1. General properties of descent methods 37

2. Methods based on conjugate directions 44

Appendix. Numerical questions relating to Fletcher's algorithm . . 68

III. Barrier and Penalty Function Methods 78

1. Basic properties of barrier functions 79

2. Multiplier relations (first order analysis) 83

3. Second order conditions 86

4. Rate of convergence results 91

5. Analysis of penalty function methods 111

6. Accelerated penalty and barrier methods 117

Note. Equations, Theorems, etc. are numbered in sequence according to subsection. Cross references between sections are indicated explicitly. For example, equation 3 of subsection 4 in the Mathematical Programming section is referenced as MP equation 4.3 in other sections.

I. Introduction to Mathematical Programming

1. Minimum of a constrained function.

Consider a function $f(\underset{\sim}{x})$ on $S \subseteq E_n \to E_1$ where S is a given point set.

Definition: $\underset{\sim}{x}^*$ is the global minimum of f on S if

$$f(\underset{\sim}{x}^*) \leq f(\underset{\sim}{x}) \quad \forall \underset{\sim}{x} \in S \quad . \tag{1.1}$$

Remark: $\underset{\sim}{x}^*$ exists, for example, if S is finite, or if S is compact and $f(\underset{\sim}{x})$ continuous on S.

Definition: $\underset{\sim}{x}^*$ is a local minimum of f on S if $\exists\, \delta > 0$ such that

$$f(\underset{\sim}{x}^*) \leq f(\underset{\sim}{x}) \quad \forall \underset{\sim}{x} \in N(\underset{\sim}{x}^*, \delta) \tag{1.2}$$

where

$$N(\underset{\sim}{x}, \delta) = \{\underset{\sim}{t} \,;\, S \cap \{\underset{\sim}{t} \,;\, \|\underset{\sim}{t} - \underset{\sim}{x}\|^{+/} \leq \delta\}\} \quad . \tag{1.3}$$

If strict inequality holds in either (1.1) or (1.2) whenever $\underset{\sim}{x} \neq \underset{\sim}{x}^*$ then the minimum is said to be isolated.

Definition: S is convex if $\underset{\sim}{x}_1, \underset{\sim}{x}_2 \in S \Rightarrow \theta \underset{\sim}{x}_1 + (1-\theta)\underset{\sim}{x}_2 \in S$ for $0 \leq \theta \leq 1$.

Example: If S is convex all finite combinations of points in S is again in S. That is $\sum_{i=1}^{m} \lambda_i \underset{\sim}{x}_i \in S$ where $\underset{\sim}{x}_i \in S$, $\sum_{i=1}^{m} \lambda_i = 1$, $\lambda_i \geq 0$, $1 \leq m < \infty$.

Definition: $f(\underset{\sim}{x})$ is a convex function on the convex set S if

$$f(\theta \underset{\sim}{x}_1 + (1-\theta)\underset{\sim}{x}_2) \leq \theta f(\underset{\sim}{x}_1) + (1-\theta) f(\underset{\sim}{x}_2) \quad , \quad 0 \leq \theta \leq 1 \quad . \tag{1.4}$$

+/ $\|\underset{\sim}{t}\| = \{\sum_{i=1}^{n} t_i^2\}^{1/2}$, the euclidean vector norm of $\underset{\sim}{t}$.

If strict inequality holds when $0 < \theta < 1$ then f is <u>strictly convex</u>. Say $g(\underset{\sim}{x})$ is <u>concave</u> (strictly concave) if $-g$ is convex (strictly convex).

<u>Lemma 1.1</u>: If $f(\underset{\sim}{x})$ is a convex function on the convex set S then a local minimum of f is the global minimum. If f is strictly convex then the minimum is unique.

<u>Proof</u>: It is necessary to consider only the case f bounded below. If $\underset{\sim}{x}^*$ is a local minimum but not the global minimum $\exists\ \underset{\sim}{x}^{**}$ such that $f(\underset{\sim}{x}^{**}) < f(\underset{\sim}{x}^*)$. Now, by assumption, $\exists\ \delta > 0$ such that $f(\underset{\sim}{x}) \geq f(\underset{\sim}{x}^*)$ for $\underset{\sim}{x} \in N(\underset{\sim}{x}^*,\delta)$. Choose $\theta > 0$ sufficiently small for $\theta\underset{\sim}{x}^{**} + (1-\theta)\underset{\sim}{x}^* \in N(\underset{\sim}{x}^*,\delta)$ then

(i) $f(\underset{\sim}{x}^*) \leq f(\theta\underset{\sim}{x}^{**} + (1-\theta)\underset{\sim}{x}^*)$ as $\underset{\sim}{x}^*$ is a local minimum, and

(ii) $f(\theta\underset{\sim}{x}^{**} + (1-\theta)\underset{\sim}{x}^*) \leq \theta f(\underset{\sim}{x}^{**}) + (1-\theta)f(\underset{\sim}{x}^*)$ by convexity

$$< f(\underset{\sim}{x}^*) \quad \text{unless} \quad f(\underset{\sim}{x}^*) = f(\underset{\sim}{x}^{**}) .$$

Now assume $\underset{\sim}{x}^*$, $\underset{\sim}{x}^{**}$ both are global minima and that f is strictly convex. Then

$$f(\theta\underset{\sim}{x}^* + (1-\theta)\underset{\sim}{x}^{**}) < \theta f(\underset{\sim}{x}^*) + (1-\theta)f(\underset{\sim}{x}^{**}) \quad , \quad 0 < \theta < 1$$

which gives a contradiction. $\square$

<u>Definition</u>: A set C is a <u>cone</u> with vertex at the origin if $\underset{\sim}{x} \in C \Rightarrow \lambda\underset{\sim}{x} \in C$, $\lambda \geq 0$. C is a cone with vertex at $\underset{\sim}{p}$ if $\{\underset{\sim}{x}-\underset{\sim}{p}\ ;\ x \in C\}$ is a cone with vertex at the origin.

Definition: $\underset{\sim}{x}$ is in the tangent cone $\mathcal{T}(S,\underset{\sim}{x}_0)$ to S at $\underset{\sim}{x}_0$ if $\exists$ sequences $\{\lambda_n\} \geq 0$, $\{\underset{\sim}{x}_n\} \to \underset{\sim}{x}_0$, $\{\underset{\sim}{x}_n\} \subset S$ such that

$$\lim_{n\to\infty} \|\lambda_n(\underset{\sim}{x}_n - \underset{\sim}{x}_0) - \underset{\sim}{x}\| = 0 \quad . \tag{1.5}$$

Example: (i) $S = \{\underset{\sim}{x} ; \|\underset{\sim}{x} - \underset{\sim}{w}\| = r\}$, $\mathcal{T}(S,\underset{\sim}{x}_0) = \{\underset{\sim}{x} ; \underset{\sim}{x}^T(\underset{\sim}{x}_0 - \underset{\sim}{w}) = 0\}$.
(ii) $S = \{\underset{\sim}{x} ; \|\underset{\sim}{x} - \underset{\sim}{w}\| \leq r\}$. $\mathcal{T}(S,\underset{\sim}{x}_0) = E_n$ if $\underset{\sim}{x}_0$ in interior of S, otherwise $\mathcal{T}(S,\underset{\sim}{x}_0) = \{\underset{\sim}{x} ; \underset{\sim}{x}^T(\underset{\sim}{x}_0 - \underset{\sim}{w}) \leq 0\}$.

Lemma 1.2: $\mathcal{T}(S,\underset{\sim}{x}_0)$ is closed.

Proof: Consider a sequence $\{\underset{\sim}{t}_i\} \in \mathcal{T}(S,\underset{\sim}{x}_0)$ such that $\|\underset{\sim}{t}_i - \underset{\sim}{t}\| \to 0$, $i \to \infty$. It is required to show that $\underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0)$. Now $\underset{\sim}{t}_i \in \mathcal{T}(S,\underset{\sim}{x}_0) \Rightarrow \exists$ $\{\lambda_j^i\} \geq 0$, $\{\underset{\sim}{x}_j^i\} \subseteq S$ such that $\lim_{j\to\infty} \|\lambda_j^i(\underset{\sim}{x}_j^i - \underset{\sim}{x}_0) - \underset{\sim}{t}_i\| = 0$. Prescribe $\{\varepsilon_i\} \downarrow 0$. Select $\underset{\sim}{t}_i$ such that $\|\underset{\sim}{t}_i - \underset{\sim}{t}\| < \varepsilon_i/2$, and $j = i(j)$ such that $\|\lambda_j^i(\underset{\sim}{x}_j^i - \underset{\sim}{x}_0) - \underset{\sim}{t}_i\| < \varepsilon_i/2$. Then $\|\lambda_j^i(\underset{\sim}{x}_j^i - \underset{\sim}{x}_0) - \underset{\sim}{t}\| < \varepsilon_i \Rightarrow \underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0)$. $\square$

Lemma 1.3: (Necessary condition for a local minimum.) If $f(\underset{\sim}{x}) \in C^1$ [+/] and if $\underset{\sim}{x}_0$ is a local minimum of f on S then $\nabla f(\underset{\sim}{x}_0)\underset{\sim}{x} \geq 0$, $\forall \underset{\sim}{x} \in \mathcal{T}(S,\underset{\sim}{x}_0)$.

Proof: Let $\underset{\sim}{x}$ be defined by sequences $\{\lambda_n\}$, $\{\underset{\sim}{x}_n\}$. As $\underset{\sim}{x}_0$ is a local minimum $\exists\, \delta > 0$ such that $f(\underset{\sim}{x}^*) \geq f(\underset{\sim}{x}_0)$ $\forall \underset{\sim}{x}^* \in N(\underset{\sim}{x}_0,\delta)$. Consider now the restriction of the sequences $\{\lambda_n\}$, $\{\underset{\sim}{x}_n\}$ such that $\underset{\sim}{x}_n \in N(\underset{\sim}{x}_0,\delta)$.

[+/] $f \in C^1$ at $\underset{\sim}{x}_0$ if $f(\underset{\sim}{x}) = f(\underset{\sim}{x}_0) + \nabla f(\underset{\sim}{x}_0)(\underset{\sim}{x} - \underset{\sim}{x}_0) + o(\|\underset{\sim}{x} - \underset{\sim}{x}_0\|)$. Higher order continuity classes are defined similarly. For example, $f \in C^2$ if the $o(\ )$ term can be estimated in the form $\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{x}_0)^T \nabla^2 f(\underset{\sim}{x}_0)(\underset{\sim}{x} - \underset{\sim}{x}_0) + o(\|\underset{\sim}{x} - \underset{\sim}{x}_0\|^2)$.

We have

$$0 \le f(\underset{\sim}{x}_n) - f(\underset{\sim}{x}_0)$$

$$\le \nabla f(\underset{\sim}{x}_0)(\underset{\sim}{x}_n - \underset{\sim}{x}_0) + o(\|\underset{\sim}{x}_n - \underset{\sim}{x}_0\|)$$

whence (note it is sufficient to consider $\underset{\sim}{x}$ such that $\|\underset{\sim}{x}\| = 1$)

$$0 \le \nabla f(x_0)\lambda_n(\underset{\sim}{x}_n - \underset{\sim}{x}_0) + o(\lambda_n\|\underset{\sim}{x}_n - \underset{\sim}{x}_0\|)$$

$$\le \nabla f(\underset{\sim}{x}_0)\underset{\sim}{x} + o(1) \quad \text{as} \quad n \to \infty \ . \qquad \square$$

<u>Example</u>: (i) If $\underset{\sim}{x}_0 \in S_0$ (the interior of S) then $\mathcal{T}(S,\underset{\sim}{x}_0) = E_n$. Thus $\underset{\sim}{x}$ can be chosen arbitrarily so that $\nabla f(\underset{\sim}{x}_0) = 0$.

(ii) If $S = \{\underset{\sim}{x} ; \|\underset{\sim}{x} - \underset{\sim}{w}\| = r\}$ then $\mathcal{T}(S,\underset{\sim}{x}_0) = \{\underset{\sim}{x} ; \underset{\sim}{x}^T(\underset{\sim}{x}_0 - \underset{\sim}{w}) = 0\}$. In particular if $\underset{\sim}{x} \in \mathcal{T}(S,\underset{\sim}{x}_0)$ then $-\underset{\sim}{x} \in \mathcal{T}(S,\underset{\sim}{x}_0)$. Thus we must have $\nabla f(\underset{\sim}{x}_0)\underset{\sim}{x} = 0 \ \forall \underset{\sim}{x}$ such that $\underset{\sim}{x}^T(\underset{\sim}{x}_0 - \underset{\sim}{w}) = 0$. Thus $\nabla f(\underset{\sim}{x}_0) = \alpha(\underset{\sim}{x}_0 - \underset{\sim}{w})$ for some α .

(iii) If $S = \{x ; \|\underset{\sim}{x} - \underset{\sim}{w}\| \le r\}$ and $\|\underset{\sim}{x}_0 - \underset{\sim}{w}\| = r$ then $\mathcal{T}(S,\underset{\sim}{x}_0) = \{\underset{\sim}{x} ; \underset{\sim}{x}^T(\underset{\sim}{x}_0 - \underset{\sim}{w}) \le 0\}$. In this case we have $\nabla f(\underset{\sim}{x}_0)\underset{\sim}{x} \ge 0$ $\forall \underset{\sim}{x}$ such that $\underset{\sim}{x}^T(\underset{\sim}{x}_0 - \underset{\sim}{w}) \le 0$. Thus $\nabla f(\underset{\sim}{x}_0) = \alpha(\underset{\sim}{w} - \underset{\sim}{x}_0)$ for some nonnegative α .

Let A be a set in E_n .

<u>Definition</u>: The <u>polar cone</u> to A is the set $A^* = \{\underset{\sim}{x} ; \underset{\sim}{x}^T\underset{\sim}{y} \le 0 \ \forall \underset{\sim}{y} \in A\}$. A^* has the following properties.

(i) A^* is a closed convex cone.

(ii) If $A_1 \subseteq A_2$ then $A_2^* \subseteq A_1^*$.

(iii) $A^{**} = A$ if and only if A is a closed convex cone.

(iv) $A^* = (\overline{A^c})^*$ -- the polar cone of the closure of the convex hull of A. The convex hull of a set is the smallest convex set containing it. Thus $A^c = \cap X$, $A \subseteq X$, X convex.

(v) If A is a subspace then $A^\perp = A^*$.

<u>Remark</u>: Lemma 1.3 can be restated: ' if $\underset{\sim}{x}^*$ is a local minimum of f on S then $-\nabla f(\underset{\sim}{x}^*) \in \mathcal{T}(S,\underset{\sim}{x}^*)^*$ '.

<u>Lemma 1.4</u>: If $\underset{\sim}{y} \in \mathcal{T}(S,\underset{\sim}{x}_0)^*$ then $-\underset{\sim}{y}$ is the gradient of a function having a local minimum on S at $\underset{\sim}{x}_0$.

<u>Remark</u>: It is sufficient to consider the case $\|\underset{\sim}{y}\| = 1$, $\underset{\sim}{x}_0 = 0$.

<u>Proof</u>: Let $C_e = \{\underset{\sim}{x} ; \underset{\sim}{x}^T \underset{\sim}{y} \leq \frac{\|\underset{\sim}{x}\|}{e}\}$ $e = 1,2,\ldots$. We first show that for each e, $\exists\ \varepsilon(e) > 0$ such that $N(0,\varepsilon(e)) \subset C_e$. For assume this is not the case. Then $\exists\ \{\underset{\sim}{x}_p\} \subset E_n - C_e$ with $\underset{\sim}{x}_p \in N(0,1/p)$, $p = 1,2,\ldots$ such that

$$\frac{\underset{\sim}{x}_p^T \underset{\sim}{y}}{\|\underset{\sim}{x}_p\|} > \frac{1}{e}, \quad p = 1,2,\ldots \quad . \tag{1.6}$$

The sequence $\left\{\frac{\underset{\sim}{x}_p}{\|\underset{\sim}{x}_p\|}\right\}$ is bounded and therefore contains a convergent subsequence $\left\{\frac{\hat{\underset{\sim}{x}}_i}{\|\hat{\underset{\sim}{x}}_i\|}\right\} \to \underset{\sim}{z}$. By definition $\underset{\sim}{z} \in \mathcal{T}(S,0)$, but, by (1.6),

$$\underset{\sim}{z}^T \underset{\sim}{y} > \frac{1}{e} > 0$$

which contradicts $\underset{\sim}{y} \in \mathcal{T}(S,0)^*$.

Now let $\hat{\varepsilon}_k = \sup\{\varepsilon, N(0,\varepsilon) \subset C_k\}$. We define

$$\varepsilon_1 = \min(1, \hat{\varepsilon}_1) ,$$

$$\varepsilon_k = \min(\tfrac{1}{2}\, \varepsilon_{k-1}, \hat{\varepsilon}_k) \quad , \quad k > 1 ,$$

and

$$P(\underset{\sim}{z}) = 2\|\underset{\sim}{z}\| \quad , \quad \|\underset{\sim}{z}\| \geq \varepsilon_2$$

$$\frac{2\|\underset{\sim}{z}\|}{k-1} \; \frac{\|\underset{\sim}{z}\| - \varepsilon_{k+1}}{\varepsilon_k - \varepsilon_{k+1}} + \frac{2\|\underset{\sim}{z}\|}{k} \; \frac{\varepsilon_k - \|\underset{\sim}{z}\|}{\varepsilon_k - \varepsilon_{k+1}} , \quad \|\underset{\sim}{z}\| \in [\varepsilon_{k+1}, \varepsilon_k] ,$$

$$0 \quad , \quad \|\underset{\sim}{z}\| = 0 .$$

It is clear that $\varepsilon_k > 0$ and ε_k monotonically decreasing. Further $P(\underset{\sim}{z}) \geq 0$, $P(\underset{\sim}{z})$ is an increasing function of $\|\underset{\sim}{z}\|$, and

$$\|\underset{\sim}{z}\| \leq \varepsilon_k \;\Rightarrow\; P(\underset{\sim}{z}) \leq \frac{2\|\underset{\sim}{z}\|}{k-1} \quad .$$

Thus $P(\underset{\sim}{z}) = o(\|\underset{\sim}{z}\|)$ so that $\nabla P(0) = 0$.

Now let $\underset{\sim}{z} = \underset{\sim}{x} - (\underset{\sim}{x}^T\underset{\sim}{y})\underset{\sim}{y}$. We show that, under appropriate conditions, $\underset{\sim}{x}^T\underset{\sim}{y} < P(\underset{\sim}{z})$. It is sufficient to consider $\underset{\sim}{x}^T\underset{\sim}{y} > 0$, and in this case

$$\|\underset{\sim}{x}\| - \underset{\sim}{x}^T\underset{\sim}{y} \;\leq\; \|\underset{\sim}{z}\| \;\leq\; \|\underset{\sim}{x}\| + \underset{\sim}{x}^T\underset{\sim}{y} \quad . \tag{1.7}$$

If $\underset{\sim}{x} \in C_e$ then $\underset{\sim}{x}^T\underset{\sim}{y} \leq \frac{\|\underset{\sim}{x}\|}{e}$. Using (1.7) we have

$$\frac{e-1}{e}\,\|\underset{\sim}{x}\| \;\leq\; \|\underset{\sim}{z}\| \;\leq\; \frac{e+1}{e}\,\|\underset{\sim}{x}\| \quad . \tag{1.8}$$

Now assume $\underset{\sim}{x} \in N(0,\varepsilon)$, $\varepsilon < \varepsilon_3$. Then $\|\underset{\sim}{x}\| \in [\varepsilon_{k+1}, \varepsilon_k]$ for some $k \geq 3$ whence $\underset{\sim}{x} \in C_k$. This gives

$$\underset{\sim}{x}^T \underset{\sim}{y} \leq \frac{\|\underset{\sim}{x}\|}{k} \leq \frac{\|\underset{\sim}{z}\|}{k-1} \quad . \tag{1.9}$$

However, $\|\underset{\sim}{z}\| \geq \frac{k-1}{k} \varepsilon_{k+1} > \varepsilon_{k+2}$ whence

$$P(\underset{\sim}{z}) \geq \frac{2\|\underset{\sim}{z}\|}{k+1} \tag{1.10}$$

so that, combining (1.9) and (1.10)

$$\underset{\sim}{x}^T \underset{\sim}{y} \leq \frac{k+1}{2(k-1)} P(\underset{\sim}{z}) \leq P(\underset{\sim}{z}) \quad . \tag{1.11}$$

Thus the function

$$f(\underset{\sim}{x}) = -\underset{\sim}{x}^T \underset{\sim}{y} + P(\underset{\sim}{x} - (\underset{\sim}{x}^T \underset{\sim}{y})\underset{\sim}{y}) \tag{1.12}$$

has a local minimum on S at $\underset{\sim}{x}_0 = 0$. Further $f \in C^1$ at 0 , and $\nabla f(0) = -\underset{\sim}{y}$. $\square$

2. Some properties of linear inequalities.

Definition: The set $H(\underset{\sim}{u},\nu) = \{\underset{\sim}{x} ; \underset{\sim}{u}^T \underset{\sim}{x} = \nu\}$ is a hyperplane. Note that the hyperplane separates E_n into two disjoint half spaces $R_+ = \{\underset{\sim}{x} ; \underset{\sim}{u}^T \underset{\sim}{x} \geq \nu\}$, $R_- = \{\underset{\sim}{x} ; \underset{\sim}{u}^T \underset{\sim}{x} < \nu\}$.

Lemma 2.1: (lemma of separating hyperplane). Let S be a closed convex set in E_n , and let $\underset{\sim}{x}_0 \notin S$. Then $\exists$ a hyperplane separating $\underset{\sim}{x}_0$ and S .

Proof: Let $\underset{\sim}{x}_1$ be any point in S . Then $\min_{\underset{\sim}{x} \in S} \|\underset{\sim}{x} - \underset{\sim}{x}_0\| \leq \|\underset{\sim}{x}_1 - \underset{\sim}{x}_0\| = r$.

The function $\|\underset{\sim}{x} - \underset{\sim}{x}_0\|$ is continuous on the closed set $S \cap \{\underset{\sim}{x} ; \|\underset{\sim}{x} - \underset{\sim}{x}_0\| \leq r\}$ and hence the minimum is attained. Let this point be $\underset{\sim}{x}^*$. From Figure 2.1 it is suggested that

$$(\underset{\sim}{x} - \underset{\sim}{x}^*)^T (\underset{\sim}{x}^* - \underset{\sim}{x}_0) = 0 \tag{2.1}$$

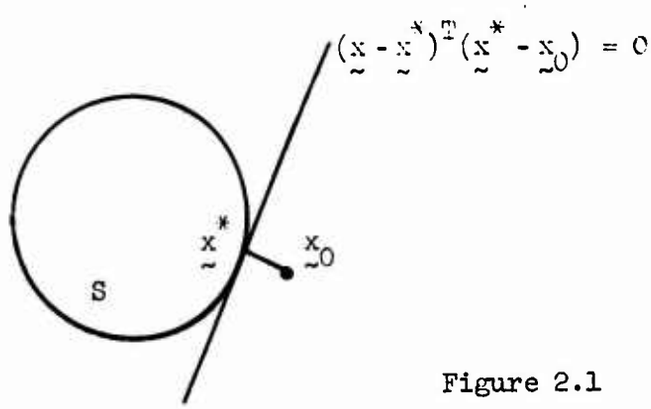

Figure 2.1

is an appropriate hyperplane. To verify this, note that $\underset{\sim}{x}_0 \in R_-$ so that it remains to show that $S \subset R_+$. Let $\underset{\sim}{x} \in S$ then for $0 \leq \theta \leq 1$,

$$\|\theta \underset{\sim}{x} + (1-\theta)\underset{\sim}{x}^* - \underset{\sim}{x}_0\|^2 \geq \|\underset{\sim}{x}^* - \underset{\sim}{x}_0\|^2$$

so that

$$\theta^2\|\underset{\sim}{x} - \underset{\sim}{x}^*\|^2 + 2\theta(\underset{\sim}{x} - \underset{\sim}{x}^*)^T(\underset{\sim}{x}^* - \underset{\sim}{x}_0) \geq 0$$

and, letting $\theta \to 0$,

$$(\underset{\sim}{x} - \underset{\sim}{x}^*)^T(\underset{\sim}{x}^* - \underset{\sim}{x}_0) \geq 0$$

whence $\underset{\sim}{x} \in R_+$. $\square$

<u>Definition</u>: C is finitely generated if

$C = \{\underset{\sim}{x} ; \underset{\sim}{x} = \sum_{i=1}^{p} \lambda_i \underset{\sim}{c}_i , \lambda_i \geq 0 , i = 1,2,\ldots,p\}$. It is clear that C is a cone. It can be shown that C is closed.

<u>Lemma 2.2</u>: (Farkas Lemma). Let A be a $p \times n$ matrix. If for every solution $\underset{\sim}{y}$ of the system of linear inequalities

$$A\underset{\sim}{y} \geq 0 \tag{2.2}$$

it is true that

$$\underset{\sim}{a}^T \underset{\sim}{y} \geq 0 \tag{2.3}$$

then $\exists\ \underset{\sim}{x} \geq 0$ such that $A^T \underset{\sim}{x} = \underset{\sim}{a}$.

<u>Proof</u>: Let C be the cone generated by $\rho_i(A)$, $i = 1,2,\ldots,p$. Then the result of Farkas lemma is that if (2.2) $\Rightarrow$ (2.3) then $\underset{\sim}{a} \epsilon C$. We assume $\underset{\sim}{a} \notin C$ and seek a contradiction. By Lemma 2.1 there exists a separating hyperplane. To construct it let $\underset{\sim}{x}^*$ be the closest point in C to $\underset{\sim}{a}$. Then $\|\lambda \underset{\sim}{x}^* - \underset{\sim}{a}\|^2$ has a minimum at $\lambda = 1$. Differentiating and setting $\lambda = 1$ gives

$$(\underset{\sim}{x}^* - \underset{\sim}{a})^T \underset{\sim}{x}^* = 0 \quad . \tag{2.4}$$

By (2.1) the equation of the separating hyperplane is

$$(\underset{\sim}{x} - \underset{\sim}{x}^*)^T (\underset{\sim}{x}^* - \underset{\sim}{a}) = \underset{\sim}{x}^T (\underset{\sim}{x}^* - \underset{\sim}{a}) = 0 \tag{2.5}$$

which shows that it passes through the origin.

By Lemma 2.1 $C \subset R_+$ whence

$$\underset{\sim}{v}^T A(\underset{\sim}{x}^* - \underset{\sim}{a}) \geq 0$$

for arbitrary $\underset{\sim}{v} \geq 0$ so that

$$A(\underset{\sim}{x}^* - \underset{\sim}{a}) \geq 0 \quad , \tag{2.6}$$

but $\underset{\sim}{a} \in R_-$ whence

$$\underset{\sim}{a}^T (\underset{\sim}{x}^* - \underset{\sim}{a}) < 0 \tag{2.7}$$

which gives the desired contradiction. □

Remark: Another way of looking at this result is that at most one of the following pair of systems can have a solution.

(i) $\quad A\underset{\sim}{x} = \underset{\sim}{b} \ , \quad \underset{\sim}{x} \geq 0 \ .$

(ii) $\quad A^T\underset{\sim}{y} \geq 0 \ , \quad \underset{\sim}{b}^T\underset{\sim}{y} < 0 \ .$

This is an example of a 'theorem of the alternative'.

3. Multiplier relations.

We consider now the mathematical programming problem (MPP)

$$\min f(\underset{\sim}{x}) \quad \text{subject to}$$

$$g_i(\underset{\sim}{x}) \geq 0 \ , \quad i \in I_1 \ ,$$

$$h_i(\underset{\sim}{x}) = 0 \ , \quad i \in I_2 \ .$$

We assume that f , g_i , $i \in I_1$, and h_i , $i \in I_2$, are in C^2 and that the constraints on the problem are not contradictory. This corresponds to the problem discussed in Section 1 with S given by

$$S = \{\underset{\sim}{x} \, ; \, g_i(\underset{\sim}{x}) \geq 0 \ , \ i \in I_1 \ , \ h_i(\underset{\sim}{x}) = 0 \ , \ i \in I_2\} \ . \tag{3.1}$$

At any point $\underset{\sim}{x}_0 \in S$ let B_0 be the index set for the constraints satisfying $g_i(\underset{\sim}{x}_0) = 0$. If $i \in B_0$ we say that g_i is <u>active</u> at $\underset{\sim}{x}_0$.

<u>Definition</u>: S is <u>Lagrange regular</u> at $\underset{\sim}{x}_0$ iff for every f such that (i) f has a minimum on S at $\underset{\sim}{x}_0$, and (ii) $f \in C^1$ at $\underset{\sim}{x}_0$ (i.e., $f \in F_0$) $\exists \ \underset{\sim}{u} \ , \ \underset{\sim}{v}$ such that

$$\text{(i)} \quad \nabla f(\underset{\sim}{x}_0) = \sum_{i \in B_0} u_i \nabla g_i(\underset{\sim}{x}_0) + \sum_{i \in I_2} v_i \nabla h_i(\underset{\sim}{x}_0) \tag{3.2}$$

$$\text{(ii)} \quad u_i \geq 0 \ , \quad i \in B_0 \ .$$

This can also be written

(i) $$\nabla f(\underset{\sim}{x}_0) = \sum_{i \in I_1} u_i \nabla g_i(\underset{\sim}{x}_0) + \sum_{i \in I_2} v_i \nabla h_i(\underset{\sim}{x}_0) \quad ,$$

(ii) $$\underset{\sim}{u}^T \underset{\sim}{g}(\underset{\sim}{x}_0) = 0 \quad , \quad \text{and}$$

(iii) $$\underset{\sim}{u} \geq 0$$

where zero multipliers are introduced corresponding to the inactive constraints.

<u>Remark</u>: If (3.2) holds for $f \in F_0$, then f satisfies the <u>Kuhn-Tucker conditions</u>.

<u>Example</u>: It is important to realize that (3.2) need not hold. Consider the MPP

$$\min f = -x_1 \ ,$$

$$\text{subject to} \quad g_1 = x_1 \geq 0 \ , \ g_2 = x_2 \geq 0 \ , \ g_3 = (1 - x_1)^3 - x_2 \geq 0 \ .$$

From Figure 3.1 it is clear that the minimum is attained at $x_1 = 1$, $x_2 = 0$, and here g_1 and g_3 are active. We nave

$$\nabla g_1 = -\nabla g_3 = \underset{\sim}{e}_2$$

while

$$\nabla f = -\underset{\sim}{e}_1$$

so that a relation of the form (3.2) is impossible.

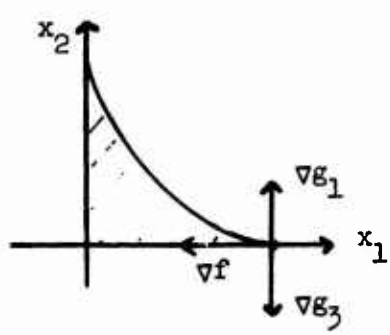

Figure 3.1

Let

$$H_0 = \{\underset{\sim}{x} \,;\, \nabla h_i(\underset{\sim}{x}_0)\underset{\sim}{x} = 0 \,,\; i \in I_2\} \,,$$

$$G_0 = \{\underset{\sim}{x} \,;\, \nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{x} \geq 0 \,,\; i \in B_0\} \,.$$

<u>Lemma 3.1</u>: S is Lagrange regular at $\underset{\sim}{x}_0$ iff $-\nabla f(\underset{\sim}{x}_0) \in (G_0 \cap H_0)^*$ for all $f \in F_0$.

<u>Proof</u>: If $-\nabla f(\underset{\sim}{x}_0) \in (G_0 \cap H_0)^*$ then

$$-\nabla f(\underset{\sim}{x}_0)\underset{\sim}{y} \geq 0$$

$\forall \underset{\sim}{y}$ such that

$$\nabla h_i(\underset{\sim}{x}_0)\underset{\sim}{y} \geq 0 \,,$$

$$-\nabla h_i(\underset{\sim}{x}_0)\underset{\sim}{y} \geq 0 \,, \quad i \in I_2$$

$$\nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{y} \geq 0 \,, \quad i \in B_0 \,.$$

Thus, by Farkas Lemma, $\nabla f(\underset{\sim}{x}_0)$ is a linear combination with nonnegative weights of $\nabla g_i(\underset{\sim}{x}_0)$, $i \in B_0$, and $\nabla h_i(\underset{\sim}{x}_0)$, $-\nabla h_i(\underset{\sim}{x}_0)$, $i \in I_2$. Thus (3.2) holds. On the other hand, if (3.2) holds then $\nabla f(\underset{\sim}{x}_0)\underset{\sim}{y} \geq 0$ for all $\underset{\sim}{y} \in G_0 \cap H_0$. $\square$

<u>Remark</u>: Lemma 3.1 shows the difficulty with the above example. Here $\mathcal{T} = \{\underset{\sim}{x} = -\alpha \underset{\sim}{e}_1 \,, \alpha \geq 0\}$, $G_0 \cap H_0 = \{\underset{\sim}{x} = \alpha \underset{\sim}{e}_1 \,, \alpha \text{ unconstrained}\}$. We have $\mathcal{T}^*$ = right half plane, $(G_0 \cap H_0)^*$ the x_2 axis. By Lemma 1.4 for every $\underset{\sim}{x} \in \mathcal{T}^*$ there is a function with a minimum at $(1,0)$ and such that $-\nabla f = \underset{\sim}{x}$. Thus the conditions of Lemma 3.1 are not met in this case.

Lemma 3.2: $(G_0 \cap H_0)^* \subseteq \mathcal{T}(S,\underset{\sim}{x}_0)^*$.

Proof: This result follows if we show that $\mathcal{T}(S,\underset{\sim}{x}_0) \subseteq H_0 \cap G_0$. If $\underset{\sim}{x} \in \mathcal{T}(S,\underset{\sim}{x}_0)$ $\exists$ $\{\underset{\sim}{x}_n\} \to \underset{\sim}{x}$, $\{\underset{\sim}{x}_n\} \subset S$, $\{\lambda_n\} \geq 0$ such that $\{\lambda_n(\underset{\sim}{x}_n - \underset{\sim}{x}_0)\} \to \underset{\sim}{x}$. We have

$$0 = h_i(\underset{\sim}{x}_n) = h_i(\underset{\sim}{x}_0) + \nabla h_i(\underset{\sim}{x}_0)(\underset{\sim}{x}_n - \underset{\sim}{x}_0) + o(\|\underset{\sim}{x}_n - \underset{\sim}{x}_0\|) \ , \quad i \in I_2 \ ,$$

and

$$0 \leq g_i(\underset{\sim}{x}_n) = g_i(\underset{\sim}{x}_0) + \nabla g_i(\underset{\sim}{x}_0)(\underset{\sim}{x}_n - \underset{\sim}{x}_0) + o(\|\underset{\sim}{x}_n - \underset{\sim}{x}_0\|) \ , \quad i \in B_0 \ .$$

Multiplying by λ_n and repeating the argument used in Lemma 1.3 we have

$$\nabla h_i(\underset{\sim}{x}_0)\underset{\sim}{x} = 0 \ , \quad i \in I_2 \ , \quad \nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{x} \geq 0 \ , \quad i \in B \ ,$$

so that $\underset{\sim}{x} \in G_0 \cap H_0$. □

Theorem 3.1: The set S is Lagrange regular at $\underset{\sim}{x}_0$ iff $\mathcal{T}(S,\underset{\sim}{x}_0)^* = (G_0 \cap H_0)^*$.

Proof: If $\mathcal{T}(S,\underset{\sim}{x}_0)^* = (G_0 \cap H_0)^*$ then $-\nabla f(\underset{\sim}{x}_0) \in (G_0 \cap H_0)^*$ $\forall f \in F_0$ by Lemma 1.3. Thus (3.2) holds by Lemma 3.1. If S is Lagrange regular at $\underset{\sim}{x}_0$ then by Lemma 3.1 $-\nabla f(\underset{\sim}{x}_0) \in (G_0 \cap H_0)^*$ $\forall f \in F_0$. $\therefore$, by Lemma 1.4, $\mathcal{T}(S,\underset{\sim}{x}_0)^* \subseteq (G_0 \cap H_0)^*$. Thus $\mathcal{T}(S,\underset{\sim}{x}_0)^* = (G_0 \cap H_0)^*$ by Lemma 3.2. □

Remark: Conditions which ensure that S is Lagrange regular at $\underset{\sim}{x}_0$ are called restraint conditions. Theorem 3.1 gives a <u>necessary and sufficient</u> restraint condition.

Corollary 3.1: (Kuhn Tucker restraint condition). If $\nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{t} \geq 0$, $i \in B_0$, and $\nabla h_i(\underset{\sim}{x}_0)\underset{\sim}{t} = 0$, $i \in I_2 \Rightarrow \underset{\sim}{t}$ is tangent at $\underset{\sim}{x}_0$ to a once

differentiable arc $\underset{\sim}{x} = \underset{\sim}{x}(\theta)$, $\underset{\sim}{x}(0) = \underset{\sim}{x}_0$ contained in $N(\underset{\sim}{x}_0,\delta)$ for some $\delta > 0$ then S is Lagrange regular at $\underset{\sim}{x}_0$.

<u>Proof</u>: It is clear that $\underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0)$ for consider a sequence $\{\theta_i\} \downarrow 0$ and define $\{\underset{\sim}{x}_n\} = \{\underset{\sim}{x}(\theta_n)\}$, $\{\lambda_n\} = \{\frac{1}{\theta_n}\}$ then

$$\{\lambda_n(\underset{\sim}{x}_n - \underset{\sim}{x}_0)\} \to \frac{d\underset{\sim}{x}(0)}{d\theta} = \underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0) \quad .$$

Thus the Kuhn Tucker restraint condition implies $(G_0 \cap H_0) \subseteq \mathcal{T}(S,\underset{\sim}{x}_0)$. The result now follows from Lemma 3.2 and Theorem 3.1. □

<u>Lemma 3.3</u>: Let $k_i(\underset{\sim}{x}) \in C^2$, $k_i(\underset{\sim}{x}^*) = 0$, and $\nabla k_i(\underset{\sim}{x}^*)t = 0$, $i = 1,2,\ldots,s < n$. We assume $\exists\ \varepsilon > 0$ such that the $\nabla k_i(\underset{\sim}{x})$, $i = 1,2,\ldots,s$ are linearly independent for $\|\underset{\sim}{x} - \underset{\sim}{x}^*\| < \varepsilon$. Then $\exists$ a smooth arc $\underset{\sim}{x} = \underset{\sim}{x}(\theta)$, $\underset{\sim}{x}(0) = \underset{\sim}{x}^*$, such that $k_i(\underset{\sim}{x}(\theta)) = 0$, $i = 1,2,\ldots,s$, for $\|\underset{\sim}{x}(\theta) - \underset{\sim}{x}^*\| < \varepsilon$ and $\frac{d\underset{\sim}{x}(0)}{d\theta} = \underset{\sim}{t}$.

<u>Proof</u>: Let $P(\underset{\sim}{x}) = K^T(K\,K^T)^{-1}\,K$ where $\rho_i(K) = \nabla k_i(\underset{\sim}{x})$, $i = 1,2,\ldots,s$. Then $\underset{\sim}{x}(\theta)$ can be found by integrating the differential equation

$$\frac{d\underset{\sim}{x}}{d\theta} = (I - P(\underset{\sim}{x}))\underset{\sim}{t} \tag{3.3}$$

subject to the initial condition $\underset{\sim}{x}(0) = \underset{\sim}{x}^*$. □

<u>Remark</u>: Let the k_i be as given in the statement of Lemma 3.3. Then the linear independence of the ∇k_i in a region containing $\underset{\sim}{x}^*$ is a consequence of the linear independence at $\underset{\sim}{x}^*$. For consider the matrix KK^T . At $\underset{\sim}{x} = \underset{\sim}{x}^*$ this matrix is positive definite as K has rank s .

Thus the smallest eigenvalue is positive. Clearly it is a continuous function of $\underset{\sim}{x}$ so that it remains positive in a small enough neighborhood of $\underset{\sim}{x}^*$, and in this neighborhood the $\nabla k_i(\underset{\sim}{x})$ are linearly independent.

<u>Lemma 3.4</u>: (Restraint condition A). S is Lagrange regular at $\underset{\sim}{x}_0$ if the set of vectors $\nabla g_i(\underset{\sim}{x}_0)$, $I \epsilon B_0$, $\nabla h_i(\underset{\sim}{x}_0)$, $i \epsilon I_2$ are linearly independent.

<u>Proof</u>: This is a consequence of Corollary 3.1 and Lemma 3.3. For let $\underset{\sim}{t} \in G_0 \cap H_0$, and let $B(t)$ be the index set such that $\nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{t} = 0$, $i \epsilon B(t)$. Then by Lemma 3.3 a smooth arc can be constructed such that $\underset{\sim}{x} = \underset{\sim}{x}(\theta))$, $g_i(\underset{\sim}{x}(\theta)) = 0$, $i \epsilon B(t)$, $h_i(\underset{\sim}{x}(\theta))$, $i \epsilon I_2$, $g_i(\underset{\sim}{x}(\theta)) \geq 0$, $i \epsilon I_1 - B(t)$, $\underset{\sim}{x}(\theta) \in N(\underset{\sim}{x}, \delta)$ for some $\delta > 0$, and $\frac{d\underset{\sim}{x}(0)}{d\theta} = \underset{\sim}{t}$. □

<u>Lemma 3.5</u>: (Restraint condition B). If $\nabla h_i(\underset{\sim}{x}_0)$, $i \epsilon I_2$ are linearly independent, and if $\exists\, \underset{\sim}{t}$ such that $\nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{t} > 0$, $i \epsilon B_0$, $\nabla h_i(\underset{\sim}{x}_0)\underset{\sim}{t} = 0$, $i \epsilon I_2$, then S is Lagrange regular at $\underset{\sim}{x}_0$.

<u>Proof</u>: Assume $\underset{\sim}{w} \in G_0 \cap H_0$ but $\underset{\sim}{w} \notin \mathcal{T}(S, \underset{\sim}{x}_0)$. Prescribe $\{\epsilon_k\} \downarrow 0$ and set $\underset{\sim}{w}_k = \underset{\sim}{w} + \epsilon_k \underset{\sim}{t}$. Then $\nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{w}_k > 0$, $i \in B_0$, $\nabla h_i(\underset{\sim}{x}_0)\underset{\sim}{w}_k = 0$, $i \in I_2$.

Now construct $\underset{\sim}{x}_k = \underset{\sim}{x}_k(\theta)$ such that $\underset{\sim}{x}_k(0) = \underset{\sim}{x}_0$, $\frac{d\underset{\sim}{x}_k(0)}{d\theta} = \underset{\sim}{w}_k$, $h_i(\underset{\sim}{x}_k(\theta)) = 0$, $i \epsilon I_2$, for $\underset{\sim}{x}_k(\theta)$ is some neighborhood of $\underset{\sim}{x}_0$. By continuity there will be a subneighborhood (say $N(\underset{\sim}{x}_0, \delta_k)$ for some $\delta_k > 0$) such that (i) $\nabla g_i(\underset{\sim}{x}_k(\theta)) \frac{d\underset{\sim}{x}_k(\theta)}{d\theta} \geq 0$, $i \in B_0$, and (ii) $g_i(\underset{\sim}{x}_k(\theta)) \geq 0$, $i \in I_1 - B_0$ for $\underset{\sim}{x}_k(\theta) \in N(\underset{\sim}{x}_0, \delta_k)$. The argument of Corollary 3.1 now gives $\underset{\sim}{w}_k \in \mathcal{T}(S, \underset{\sim}{x}_0)$. But, by construction, $\{\underset{\sim}{w}_k\} \to \underset{\sim}{w}$. Thus $\underset{\sim}{w} \in \mathcal{T}(S, \underset{\sim}{x}_0)$ as $\mathcal{T}$ is closed. □

4. Second order conditions.

In certain cases it is possible to further characterize local minima of f on S by looking at second derviative information.

Lemma 4.1: Let $w(\underset{\sim}{x}) \in C^2$, w have a local minimum on S at $\underset{\sim}{x}_0$, and $\nabla w(\underset{\sim}{x}_0) = 0$. Then $\underset{\sim}{t}^T \nabla^2 w(\underset{\sim}{x}_0)\underset{\sim}{t} \geq 0 \quad \forall \underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0)$. If $\underset{\sim}{t}^T \nabla^2 w(\underset{\sim}{x}_0)\underset{\sim}{t} > 0 \quad \forall \underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0)$ then $\exists\, \delta > 0$, $m > 0$ such that $w(x) \geq w(\underset{\sim}{x}_0) + m\|\underset{\sim}{x} - \underset{\sim}{x}_0\|^2$, $\underset{\sim}{x} \in N(\underset{\sim}{x}_0,\delta)$.

Proof: Let $\{\underset{\sim}{x}_n\}$, $\{\lambda_n\}$ be defining sequences for $\underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0)$. Then for n large enough we have, as $\nabla w(\underset{\sim}{x}_0) = 0$,

$$0 \leq w(\underset{\sim}{x}_n) - w(\underset{\sim}{x}_0) = \frac{1}{2}(\underset{\sim}{x}_n - \underset{\sim}{x}_0)^T \nabla^2 w(\underset{\sim}{x}_0)(\underset{\sim}{x}_n - \underset{\sim}{x}_0) + o(\|\underset{\sim}{x}_n - \underset{\sim}{x}_0\|^2)$$

$$\therefore 0 \leq \lambda_n^2(w(\underset{\sim}{x}_n) - w(\underset{\sim}{x}_0)) = \frac{1}{2}\underset{\sim}{t}^T \nabla^2 w(\underset{\sim}{x}_0)\underset{\sim}{t} + o(1) \quad \text{as } \underset{\sim}{x}_n \to \underset{\sim}{x}_0 .$$

Now assume $\underset{\sim}{t}^T \nabla^2 w(\underset{\sim}{x}_0)\underset{\sim}{t} > 0 \quad \forall \underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0)$ and $\exists$ no $m > 0$ such that $w(\underset{\sim}{x}) \geq w(\underset{\sim}{x}_0) + m\|\underset{\sim}{x} - \underset{\sim}{x}_0\|^2$ for $\underset{\sim}{x}$ in any neighborhood of $\underset{\sim}{x}_0$. This implies that for any integer q, $\exists\, \underset{\sim}{x}_q \in S$ such that (i) $\underset{\sim}{x}_q \in N(\underset{\sim}{x}_0, 1/q)$, (ii) $w(\underset{\sim}{x}_q) - w(\underset{\sim}{x}_0) < \frac{1}{q}\|\underset{\sim}{x}_q - \underset{\sim}{x}_0\|^2$. Select a subsequence of the $\underset{\sim}{x}_q$ such that $\left\{\dfrac{\underset{\sim}{x}_q - \underset{\sim}{x}_0}{\|\underset{\sim}{x}_q - \underset{\sim}{x}_0\|}\right\} \to \underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0)$. Then (ii) $\Rightarrow \underset{\sim}{t}^T \nabla^2 w(\underset{\sim}{x}_0)\underset{\sim}{t} \leq 0$ which gives a contradiction. $\square$

Definition: The Lagrangian function associated with the MPP is given by

$$\mathcal{L}(\underset{\sim}{x},\underset{\sim}{u},\underset{\sim}{v}) = f(\underset{\sim}{x}) - \sum_{i \in I_1} u_i g_i(\underset{\sim}{x}) - \sum_{i \in I_2} v_i h_i(\underset{\sim}{x}) . \tag{4.1}$$

It will frequently be convenient to suppress the dependence of $\mathcal{L}$ on $\underset{\sim}{u}$ and $\underset{\sim}{v}$ in the case where these are implied by the Kuhn Tucker conditions. In this case (3.2) becomes

$$\nabla \mathcal{L}(\underset{\sim}{x}_0) = 0 \quad . \tag{4.2}$$

<u>Lemma 4.2</u>: Let S be Lagrange regular at $\underset{\sim}{x}_0$, $f(\underset{\sim}{x})$ have a local minimum on S at $\underset{\sim}{x}_0$, and $S_1 = \{\underset{\sim}{x} ; \underset{\sim}{x} \epsilon S , g_i(\underset{\sim}{x}) = 0 , i \epsilon B_0\}$ then

$$\underset{\sim}{t}^T \nabla^2 \mathcal{L}(\underset{\sim}{x}_0) \underset{\sim}{t} \geq 0 \ , \ \forall \underset{\sim}{t} \in \mathcal{T}(S_1, \underset{\sim}{x}_0) \quad . \tag{4.3}$$

<u>Proof</u>: Note that $\mathcal{L} = f$ on S_1 so that $\mathcal{L}$ has a minimum on S_1 at $\underset{\sim}{x}_0$. Also, as S is Lagrange regular at $\underset{\sim}{x}_0$, $\nabla \mathcal{L}(\underset{\sim}{x}_0) = 0$. Thus the result follows from Lemma 4.1. □

<u>Remark</u>: If S_1 is Lagrange regular at $\underset{\sim}{x}_0$ then $\underset{\sim}{t}^T \nabla^2 \mathcal{L}(\underset{\sim}{x}_0) \underset{\sim}{t} \geq 0 \ \forall \underset{\sim}{t}$ such that $\nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{t} = 0$, $i \epsilon B_0$ and $\nabla h_i(\underset{\sim}{x}_0)\underset{\sim}{t} = 0$, $i \epsilon I_2$.

<u>Example</u>: Consider

$$g_1 = x_1^2 + (x_2 + 1)^2 - 1 \geq 0 \quad , \quad g_2 = 1 - x_1^2 - (x_2 - 1)^2 \geq 0 \quad .$$

S is illustrated diagramatically in Figure 4.1. At $x_1 = x_2 = 0$, $\nabla g_1 = \nabla g_2 = (0,2)$. However S is Lagrange regular at the origin -- for example, $\underset{\sim}{e}_2$ satisfies $\nabla g_1 \underset{\sim}{e}_2 > 0$, $\nabla g_2 \underset{\sim}{e}_2 > 0$ so that restraint condition B applies. In this case S_1 is the single point $\underset{\sim}{x} = \underset{\sim}{\theta}$ so that $\mathcal{T}(S_1, \underset{\sim}{\theta})$ is null.

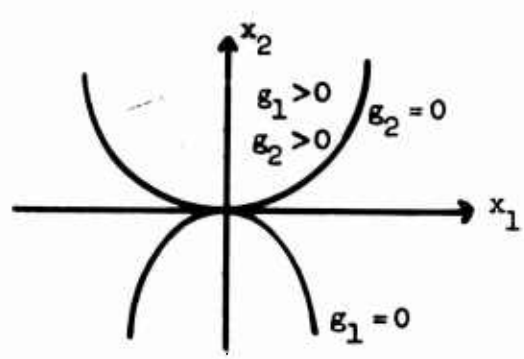

Figure 4.1

Lemma 4.3: If $\underset{\sim}{t}^T \nabla^2 \mathcal{L}(\underset{\sim}{x}_0)\underset{\sim}{t} > 0$ $\forall \underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0)$ such that $\nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{t} = 0$ $\forall i \in B_0$ such that $u_i > 0$ then $\exists$ m , $\delta > 0$ such that

$$f(\underset{\sim}{x}) \geq f(\underset{\sim}{x}_0) + m\|\underset{\sim}{x} - \underset{\sim}{x}_0\|^2 , \quad \underset{\sim}{x} \in N(\underset{\sim}{x}_0,\delta) \quad . \tag{4.4}$$

Proof: Assume $\nexists$ no m , $\delta > 0$ such that (4.4) holds. Then for each integer q $\exists$ $\underset{\sim}{x}_q$ such that (i) $\underset{\sim}{x}_q \in N(\underset{\sim}{x}_0, 1/q)$, (ii) $f(\underset{\sim}{x}_q) - f(\underset{\sim}{x}_0) < \frac{1}{q}\|\underset{\sim}{x}_q - \underset{\sim}{x}_0\|^2$. Select a subsequence of the $\underset{\sim}{x}_q$ such that

$\dfrac{\underset{\sim}{x}_q - \underset{\sim}{x}_0}{\|\underset{\sim}{x}_q - \underset{\sim}{x}_0\|} \to \underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0)$. Set $G = \sum_{i \in B_0} u_i g_i(\underset{\sim}{x})$. Then $G \geq 0$ on S , $G(\underset{\sim}{x}_0) = 0$, and $f = \mathcal{L} + G$. For the subsequence defining $\underset{\sim}{t}$ we have

$$\frac{\mathcal{L}(\underset{\sim}{x}_q) - \mathcal{L}(\underset{\sim}{x}_0)}{\|\underset{\sim}{x}_q - \underset{\sim}{x}_0\|^2} + \frac{G(\underset{\sim}{x}_q)}{\|\underset{\sim}{x}_q - \underset{\sim}{x}_0\|^2} < \frac{1}{q} \quad , . \tag{4.5}$$

Thus

$$\underset{\sim}{t}^T \nabla^2 \mathcal{L}(\underset{\sim}{x}_0)\underset{\sim}{t} + \limsup_{q \to \infty} \frac{G(\underset{\sim}{x}_q)}{\|\underset{\sim}{x}_q - \underset{\sim}{x}_0\|^2} \leq 0 \tag{4.6}$$

As $G(\underset{\sim}{x}_q) \geq 0$, the second term is bounded and nonnegative. Therefore

$$0 = \lim_{q \to \infty} \frac{G(\underset{\sim}{x}_q)}{\|\underset{\sim}{x}_q - \underset{\sim}{x}_0\|} = \sum_{i \in B_0} u_i \nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{t} \quad . \tag{4.7}$$

Thus

$$\nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{t} = 0 , \quad \forall i \in B_0 \text{ such that } u_i > 0 \tag{4.8}$$

so that (4.6) states that $\exists$ $\underset{\sim}{t} \in \mathcal{T}(S,\underset{\sim}{x}_0)$ such that $\underset{\sim}{t}$ satisfies (4.8) and that $\underset{\sim}{t}^T \nabla^2 \mathcal{L}(\underset{\sim}{x}_0)\underset{\sim}{t} \leq 0$. This gives a contradiction. □

Consider now the system

$$\nabla \mathcal{L}(\underset{\sim}{x}_0)^T = 0$$

$$u_i g_i(\underset{\sim}{x}) = 0 \ , \quad i = 1,2,\ldots,m \ ,$$

$$h_i(\underset{\sim}{x}) = 0 \ , \quad i = 1,2,\ldots,p \tag{4.9}$$

where explicit enumerations of I_1 and I_2 are assumed.

<u>Definition</u>: $J(\underset{\sim}{x}_0)$ is the Jacobian of the system (4.9) with respect to $(\underset{\sim}{x},\underset{\sim}{u},\underset{\sim}{v})$.

$$J(\underset{\sim}{x}_0) = \begin{bmatrix} \nabla^2 \mathcal{L}(\underset{\sim}{x}_0) & -\nabla g_1(\underset{\sim}{x}_0)^T & \cdots & -\nabla g_m(\underset{\sim}{x}_0)^T & -\nabla h_1(\underset{\sim}{x}_0)^T & \cdots & -\nabla h_p(\underset{\sim}{x}_0)^T \\ u_1 \nabla g_1(\underset{\sim}{x}_0) & g_1(\underset{\sim}{x}_0) & & & & & \\ \vdots & & \ddots & & & 0 & \\ u_m \nabla g_m(\underset{\sim}{x}_0) & & & g_m(\underset{\sim}{x}_0) & & & \\ \nabla h_1(\underset{\sim}{x}_0) & & & & & & \\ \vdots & & & 0 & & 0 & \\ \nabla h_p(\underset{\sim}{x}_0) & & & & & & \end{bmatrix} \tag{4.10}$$

<u>Lemma 4.4</u>: If $J(\underset{\sim}{x}_0)$ is nonsingular, then $\underset{\sim}{x}_0$ is an isolated local minimum of f on S .

<u>Remark</u>: Note that the condition $J(\underset{\sim}{x}_0)$ nonsingular imposes strong conditions on the problem. For example,

(i) the active constraint gradients must be linearly independent, and

(ii) if $g_i(\underset{\sim}{x}_0) = 0$ then $u_i > 0$ (this condition is called <u>strict complementarity</u>).

In particular S_1 is Lagrange regular at $\underset{\sim}{x}_0$.

<u>Proof</u>: If J is singular there is a vector $\begin{bmatrix} \underset{\sim}{y} \\ \underset{\sim}{a} \\ \underset{\sim}{b} \end{bmatrix}$ satisfying

$$J \begin{bmatrix} \underset{\sim}{y} \\ \underset{\sim}{a} \\ \underset{\sim}{b} \end{bmatrix} = 0 \quad . \tag{4.11}$$

This relation gives

(i) $\nabla h_i(\underset{\sim}{x}_0)\underset{\sim}{y} = 0$, $i = 1,2,\ldots,p$,

(ii) $u_i \nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{y} + a_i g_i(\underset{\sim}{x}_0) = 0$, $i = 1,2,\ldots,m$, and

(iii) $\nabla^2 \mathcal{L}(x_0)\underset{\sim}{y} - \sum_{i=1}^{m} a_i \nabla g_i(x_0)^T - \sum_{i=1}^{p} b_i \nabla h_i(x_0)^T = 0$.

From (ii) we see that $u_i > 0 \Rightarrow \nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{y} = 0$ while $u_i = 0 \Rightarrow a_i = 0$.
Now consider the problem

$$\min \underset{\sim}{y}^T \nabla^2 \mathcal{L}(\underset{\sim}{x}_0)\underset{\sim}{y}$$

subject to $\nabla g_i(\underset{\sim}{x}_0)\underset{\sim}{y} = 0$, $i \epsilon B_0$, $\nabla h_i(\underset{\sim}{x}_0)\underset{\sim}{y} = 0$, $i \epsilon I_2$, and $\|\underset{\sim}{y}\|^2 = 1$.
Clearly the constraint gradients are linearly independent as $2\underset{\sim}{y} = \nabla(\|\underset{\sim}{y}\|^2)$ is in the orthogonal complement of the set spanned by the other constraint gradients. Thus the set of feasible $\underset{\sim}{y}$ is Lagrange regular at every point by restraint condition A. Let $\underset{\sim}{y}_0$ minimize the objective function (the minimum exists as the constraint set is compact), then the Lagrange regularity ensures that $\exists$ multipliers λ , a_i , $i \epsilon B_0$, b_i , $i \epsilon I_2$ such that

$$2\nabla^2 \mathcal{L}(\underset{\sim}{x}_0)\underset{\sim}{y}_0 - 2\lambda \underset{\sim}{y}_0 - \sum_{i\in B_0} a_i \nabla g_i(\underset{\sim}{x}_0)^T - \sum_{i\in I_2} b_i \nabla h_i(\underset{\sim}{x}_0)^T = 0 \tag{4.12}$$

whence

$$\lambda = \underset{\sim}{y}_0^T \nabla^2 \mathcal{L}(\underset{\sim}{x}_0)\underset{\sim}{y}_0 = \min \underset{\sim}{y}^T \nabla^2 \mathcal{L}(\underset{\sim}{x}_0)\underset{\sim}{y} \geq 0$$

Now if $\lambda = 0$, (4.12) shows that conditions (i) - (iii) above are satisfied and hence $J(\underset{\sim}{x}_0)$ singular. Thus if $J(\underset{\sim}{x}_0)$ nonsingular, then $\lambda > 0$. In this case Lemma 4.3 shows that the minimum of the MPP is isolated. □

5. Convex programming problems.

If $g_i(\underset{\sim}{x})$ concave, $i\in I_1$, then the set $S = \{\underset{\sim}{x} ; g_i(\underset{\sim}{x}) \geq 0 , i\in I_2\}$ is convex. The problem of minimizing a convex function on S is called a convex programming problem. In this section certain properties of this problem are studied. We require the following characterization of convex functions.

Lemma 5.1: If $f(\underset{\sim}{x}) \in C^1$ then $f(\underset{\sim}{x})$ is convex on S iff

$$f(\underset{\sim}{x}) + \nabla f(\underset{\sim}{x})(\underset{\sim}{y}-\underset{\sim}{x}) \leq f(\underset{\sim}{y}) \quad , \quad \underset{\sim}{x},\underset{\sim}{y} \in S \tag{5.1}$$

Proof: If f convex then, for $0 \leq \lambda \leq 1$,

$$f(\underset{\sim}{x} + (1-\lambda)(\underset{\sim}{y}-\underset{\sim}{x})) \leq f(\underset{\sim}{x}) + (1-\lambda)(f(\underset{\sim}{y}) - f(\underset{\sim}{x}))$$

whence, if $\lambda < 1$,

$$\frac{f(\underset{\sim}{x} + (1-\lambda)(\underset{\sim}{y}-\underset{\sim}{x})) - f(\underset{\sim}{x})}{1-\lambda} \leq f(\underset{\sim}{y}) - f(\underset{\sim}{x}) \quad .$$

The necessity follows on letting $\lambda \to 1$. Now if (5.1) holds then

$$f(\lambda \underset{\sim}{x} + (1-\lambda)\underset{\sim}{y}) + \lambda \nabla f(\lambda \underset{\sim}{x} + (1-\lambda)\underset{\sim}{y})(\underset{\sim}{y}-\underset{\sim}{x}) \leq f(\underset{\sim}{y}) \quad , \tag{5.2}$$

$$f(\lambda \underset{\sim}{x} + (1-\lambda)\underset{\sim}{y}) - (1-\lambda)\nabla f(\lambda \underset{\sim}{x} + (1-\lambda)\underset{\sim}{y})(\underset{\sim}{y}-\underset{\sim}{x}) \leq f(\underset{\sim}{x}) \quad . \tag{5.3}$$

Multiplying (5.2) by $(1-\lambda)$, (5.3) by λ and adding gives (1.4) which demonstrates sufficiency. □

<u>Lemma 5.2</u>: If $S = \{\underset{\sim}{x} ; g_i(\underset{\sim}{x}) \geq 0 , g_i \text{ concave}, i \epsilon I_1\}$ has an interior point $\underset{\sim}{x}^*$, then every point of S is Lagrange regular.

<u>Proof</u>: Consider $\underset{\sim}{x}_0 \epsilon S$. Let $i \epsilon B_0$ then Lemma 5.1 gives

$$\nabla g_i(\underset{\sim}{x}_0)(\underset{\sim}{x}^* - \underset{\sim}{x}_0) \geq g_i(\underset{\sim}{x}^*) > 0 \tag{5.4}$$

as $g_i(\underset{\sim}{x}_0) = 0$, $i \epsilon B_0$. Thus restraint condition B is satisfied. □

<u>Lemma 5.3</u>: If f convex satisfies the Kuhn Tucker conditions at $\underset{\sim}{x}_0$ then f has a minimum on S at $\underset{\sim}{x}_0$.

<u>Proof</u>: In this case (3.2) gives

$$\nabla f(\underset{\sim}{x}_0) = \sum_{i \epsilon I_1} u_i \nabla g_i(\underset{\sim}{x}_0) \quad , \quad \underset{\sim}{u}^T \underset{\sim}{g}(\underset{\sim}{x}_0) = 0 \quad , \quad u_i \geq 0 \quad .$$

Let $\underset{\sim}{x}$ be any other point of S , then

$$f(\underset{\sim}{x}) \geq f(\underset{\sim}{x}) - \sum_{i \epsilon I_1} u_i g_i(\underset{\sim}{x}) = \mathcal{L}(\underset{\sim}{x}) \tag{5.5}$$

where $\mathcal{L}(\underset{\sim}{x})$ is convex on S as the $g_i(\underset{\sim}{x})$, $i \epsilon I_1$, are concave. Thus

$$\begin{aligned} f(\underset{\sim}{x}) &\geq \mathcal{L}(\underset{\sim}{x}_0) + \nabla \mathcal{L}(\underset{\sim}{x}_0)(\underset{\sim}{x} - \underset{\sim}{x}_0) \\ &= f(\underset{\sim}{x}_0) \quad . \quad \square \end{aligned}$$

Remark: If f has an interior the Kuhn Tucker conditions are both necessary and sufficient for a minimum of the convex programming problem.

Definition: The primal function for the convex programming problem is

$$\omega(\underset{\sim}{z}) = \inf_{\underset{\sim}{x} \in S_z} f(x) \quad , \quad S_z = \{\underset{\sim}{x} \, ; \, \underset{\sim}{g}(\underset{\sim}{x}) \geq \underset{\sim}{z}\} \quad . \tag{5.6}$$

Note that if $\underset{\sim}{z}_1 \geq \underset{\sim}{z}_2$ then $S_{z_1} \subseteq S_{z_2}$ so that $\omega(\underset{\sim}{z}_1) \geq \omega(\underset{\sim}{z}_2)$ and that if S has an interior then S_z nonempty for $\underset{\sim}{z} \geq 0$ and small enough.

Lemma 5.3: $\omega(\underset{\sim}{z})$ is convex.

Proof: If $\underset{\sim}{x}_1 \in S_{z_1}$, $\underset{\sim}{x}_2 \in S_{z_2}$ then, by concavity of g_i , $i \in I_1$,

$$\underset{\sim}{g}(\lambda \underset{\sim}{x}_1 + (1-\lambda)\underset{\sim}{x}_2) \geq \lambda \underset{\sim}{z}_1 + (1-\lambda)\underset{\sim}{z}_2 \quad , \quad 0 \leq \lambda \leq 1 \, .$$

Thus $\lambda \underset{\sim}{x}_1 + (1-\lambda)\underset{\sim}{x}_2 \in S_{\lambda z_1 + (1-\lambda) z_2}$. We have

$$\begin{aligned}
\omega(\lambda \underset{\sim}{z}_1 + (1-\lambda)\underset{\sim}{z}_2) &\leq \inf_{\underset{\sim}{x}_1 \in S_{z_1}, \underset{\sim}{x}_2 \in S_{z_2}} f(\lambda \underset{\sim}{x}_1 + (1-\lambda)\underset{\sim}{x}_2) \\
&\leq \inf_{\underset{\sim}{x}_1 \in S_{z_1}, \underset{\sim}{x}_2 \in S_{z_2}} (\lambda f(\underset{\sim}{x}_1) + (1-\lambda) f(\underset{\sim}{x}_2)) \quad \text{by convexity} \\
&\leq \lambda \inf_{\underset{\sim}{x}_1 \in S_{z_1}} f(\underset{\sim}{x}_1) + (1-\lambda) \inf_{\underset{\sim}{x}_2 \in S_{z_2}} f(\underset{\sim}{x}_2) \\
&\leq \lambda \omega(\underset{\sim}{z}_1) + (1-\lambda)\omega(\underset{\sim}{z}_2) \quad . \qquad \square
\end{aligned}$$

Definition: The dual function is

$$\phi(\underset{\sim}{z}^*) = \inf_{\underset{\sim}{x} \in \Omega} f(\underset{\sim}{x}) - \underset{\sim}{g}^T(\underset{\sim}{x})\underset{\sim}{z}^* \quad , \quad \underset{\sim}{z}^* \geq 0 \tag{5.7}$$

where Ω is the region on which f , $-g_i$, $i \in I_1$, are convex.

Lemma 5.4: $\phi(\underset{\sim}{z}^*)$ is concave.

Proof: Let $0 \leq \lambda \leq 1$, and $\underset{\sim}{z}_1^*, \underset{\sim}{z}_2^* \geq 0$, then

$$\begin{aligned}
\phi(\lambda \underset{\sim}{z}_1^* + (1-\lambda)\underset{\sim}{z}_2^*) &= \inf_{\underset{\sim}{x}} \{f(\underset{\sim}{x}) - \underset{\sim}{g}^T(\underset{\sim}{x})(\lambda \underset{\sim}{z}_1^* + (1-\lambda)\underset{\sim}{z}_2^*)\} \\
&= \inf_{\underset{\sim}{x}} \{\lambda(f - \underset{\sim}{g}^T \underset{\sim}{z}_1^*) + (1-\lambda)(f - \underset{\sim}{g}^T \underset{\sim}{z}_2^*)\} \\
&\geq \lambda \inf_{\underset{\sim}{x}} (f - \underset{\sim}{g}^T \underset{\sim}{z}_1^*) + (1-\lambda) \inf_{\underset{\sim}{x}} (f - \underset{\sim}{g}^T \underset{\sim}{z}_2^*) \\
&\geq \lambda\, \phi(\underset{\sim}{z}_1^*) + (1-\lambda)\phi(\underset{\sim}{z}_2^*) \quad . \qquad \square
\end{aligned}$$

Lemma 5.5: Let $\Gamma = \{\underset{\sim}{z}\,;\, \exists\ \underset{\sim}{x} \in \Omega$ such that $\underset{\sim}{g}(\underset{\sim}{x}) \geq \underset{\sim}{z}\}$. Then

$$\phi(\underset{\sim}{z}^*) = \inf_{\underset{\sim}{z} \in \Gamma} (\omega(\underset{\sim}{z}) - \underset{\sim}{z}^T \underset{\sim}{z}^*) \quad . \tag{5.8}$$

Proof:

$$\begin{aligned}
\phi(\underset{\sim}{z}^*) &= \inf_{\underset{\sim}{x}} (f(\underset{\sim}{x}) - g(\underset{\sim}{x})^T \underset{\sim}{z}^*) \quad , \\
&\leq \inf_{\underset{\sim}{x} \in S_z} (f(\underset{\sim}{x}) - \underset{\sim}{z}^T \underset{\sim}{z}^*) \quad , \\
&= \omega(\underset{\sim}{z}) - \underset{\sim}{z}^T \underset{\sim}{z}^*
\end{aligned}$$

$$\therefore \quad \phi(\underset{\sim}{z}^*) \leq \inf_{\underset{\sim}{z} \in \Gamma} (\omega(\underset{\sim}{z}) - \underset{\sim}{z}^T \underset{\sim}{z}^*) \quad . \tag{5.9}$$

Now let $\underset{\sim}{g}(\underset{\sim}{x}_1) = \underset{\sim}{z}_1$. Then

$$f(\underset{\sim}{x}_1) - \underset{\sim}{g}^T(\underset{\sim}{x}_1)\underset{\sim}{z}^* \geq \inf_{\underset{\sim}{x}\in S_{z_1}} (f(\underset{\sim}{x}) - \underset{\sim}{z}_1^T \underset{\sim}{z}^*)$$

$$\geq \omega(\underset{\sim}{z}_1) - \underset{\sim}{z}_1^T \underset{\sim}{z}^*$$

$$\geq \inf_{\underset{\sim}{z}\in\Gamma} (\omega(\underset{\sim}{z}) - \underset{\sim}{z}^T\underset{\sim}{z}^*)$$

$$\therefore \quad \inf_{\underset{\sim}{x}_1} (f(\underset{\sim}{x}_1) - \underset{\sim}{g}^T(\underset{\sim}{x}_1)\underset{\sim}{z}^*) \geq \inf_{\underset{\sim}{z}\in\Gamma} (\omega(\underset{\sim}{z}) - \underset{\sim}{z}^T\underset{\sim}{z}^*) \quad . \tag{5.10}$$

The result follows from the inequalities (5.9) and (5.10). □

<u>Theorem 5.1</u>: (Duality theorem). (i) $\sup_{\underset{\sim}{z}^*\geq 0} \phi(\underset{\sim}{z}^*) \leq \inf_{\underset{\sim}{x}\in S} f(\underset{\sim}{x})$.

(ii) If S has an interior, and $\exists\, \underset{\sim}{x}_0$ such that the Kuhn Tucker conditions are satisfied, then $\exists\, \underset{\sim}{z}^*$ maximizing $\phi(\underset{\sim}{z}^*)$ and equality holds in (i).

<u>Proof</u>: From Lemma 5.5 we have that

$$\phi(\underset{\sim}{z}^*) \leq \omega(\underset{\sim}{\theta}) = \inf_{\underset{\sim}{x}\in S} f(\underset{\sim}{x})$$

holds for each $\underset{\sim}{z}^* \geq 0$. Thus

$$\sup_{\underset{\sim}{z}^*\geq 0} \phi(\underset{\sim}{z}^*) \leq \inf_{\underset{\sim}{x}\in S} f(\underset{\sim}{x}) \quad . \tag{5.11}$$

If $\exists\, \underset{\sim}{x}_0$ such that the Kuhn Tucker conditions are satisfied then $\underset{\sim}{x}_0$ minimizes f on S . Defining $\underset{\sim}{z}^* = \{u_1,\ldots,u_m\}^T$ where the $u_i \geq 0$ are the multipliers in the Kuhn Tucker conditions we see that $\phi(\underset{\sim}{z}^*) = f(\underset{\sim}{x}_0)$. □

Corollary 5.1: (Wolfe's form of the duality theorem). Consider the primal problem minimize the convex function $f(\underset{\sim}{x})$ subject to the concave constraints $g_i(\underset{\sim}{x}) \geq 0$, $i = 1,2,\ldots,m$, and the dual problem maximize $\mathcal{L}(\underset{\sim}{x},\underset{\sim}{u})$ subject to $\nabla_x \mathcal{L} = 0$, $\underset{\sim}{u} \geq 0$. If a solution to the primal exists then the dual problem has a solution and the objective function values are equal.

Remark: (i) The linear programming problem

$$\min \underset{\sim}{a}^T \underset{\sim}{x} \quad \text{subject to} \quad A\underset{\sim}{x} - \underset{\sim}{b} \geq 0 \tag{5.12}$$

is a special case of a convex programming problem as linear functions have the special property of being both convex and concave -- this is an immediate consequence of Lemma 5.1. This property of linear constraints permits the previous discussion to be extended to permit linear equality constraints. Note that if the linear equality constraints are not to be contradictory, then their gradients must be linearly independent.

(ii) If the restraint condition B is satisfied at $\underset{\sim}{x}_0$, and $f(\underset{\sim}{x})$ has a minimum on S at $\underset{\sim}{x}_0$ then $\underset{\sim}{x}_0$ also solves the linear programming problem

$$\min f(\underset{\sim}{x}_0) + \nabla f(\underset{\sim}{x}_0)(\underset{\sim}{x} - \underset{\sim}{x}_0)$$

subject to

(i) $g_i(\underset{\sim}{x}_0) + \nabla g_i(\underset{\sim}{x}_0)(\underset{\sim}{x} - \underset{\sim}{x}_0) \geq 0$, $i \epsilon I_1$, and

(ii) $h_i(\underset{\sim}{x}_0) + \nabla h_i(\underset{\sim}{x}_0)(\underset{\sim}{x} - \underset{\sim}{x}_0) \geq 0$,

$-h_i(\underset{\sim}{x}_0) - \nabla h_i(\underset{\sim}{x}_0)(\underset{\sim}{x} - \underset{\sim}{x}_0) \geq 0$, $i \epsilon I_2$,

as the Kuhn Tucker conditions are both necessary and sufficient for a

solution to the linear programming problem. That the converse need not be true is readily seen from the example min $-x$ subject to $1-x^2-y^2 \geq 0$ which has a minimum at $x = 1$, $y = 0$. The associated linear programming problem is min $-x$ subject to $1-x \geq 0$ which has the solution $(1,y)$ for any y. Thus additional conditions are required if the converse is to hold (for example, Lemmas 4.3 or 4.4 could be used).

<u>Example</u>: (i) (Duality in linear programming). Consider the primal problem

$$\text{minimize } \underset{\sim}{a}^T \underset{\sim}{x} \quad \text{subject to} \quad A\underset{\sim}{x} - \underset{\sim}{b} \geq 0 \ .$$

The corresponding dual is

$$\text{maximize } \underset{\sim}{b}^T \underset{\sim}{u} \quad \text{subject to} \quad A^T\underset{\sim}{u} - \underset{\sim}{a} = 0 \ , \quad \underset{\sim}{u} \geq 0 \ .$$

If the primal has a solution then so does the dual and the objective function values are equal.

(ii) (The cutting plane algorithm).

(a) Consider the set $S = \{\underset{\sim}{x} \, ; \, g_i(\underset{\sim}{x}) \geq 0 \text{ and } g_i \text{ concave}, \ i \in I_1\}$. If $\underset{\sim}{x}^* \notin S$ then $g_i(\underset{\sim}{x}^*) < 0$ for at least one i. Let α satisfy $g_\alpha(\underset{\sim}{x}^*) \leq g_i(\underset{\sim}{x}^*)$, $i \in I_1$. Consider the half space $U = \{\underset{\sim}{x} \, ; \, g_\alpha(\underset{\sim}{x}^*) + \nabla g_\alpha(\underset{\sim}{x}^*)(\underset{\sim}{x} - \underset{\sim}{x}^*) \geq 0\}$. Then $x^* \notin U$. Now if $g_\alpha(\underset{\sim}{x}) \geq 0$ then, as g_α concave,

$$g_\alpha(\underset{\sim}{x}^*) + \nabla g_\alpha(\underset{\sim}{x}^*)(\underset{\sim}{x} - \underset{\sim}{x}^*) \geq g_\alpha(\underset{\sim}{x}) \geq 0 \ .$$

Thus $g_\alpha(\underset{\sim}{x}) \geq 0 \Rightarrow \underset{\sim}{x} \in U$ so that

$$S_\alpha = \{\underset{\sim}{x} \, ; \, g_\alpha(\underset{\sim}{x}) \geq 0\} \subseteq U \ .$$

We have $S \subseteq S_\alpha \subseteq U$. Thus the hyperplane $g_\alpha(\underset{\sim}{x}^*) + \nabla g_\alpha(\underset{\sim}{x}^*)(\underset{\sim}{x} - \underset{\sim}{x}^*) = 0$ separates $\underset{\sim}{x}^*$ and S .

(b) The convex programming problem minimize $f(\underset{\sim}{x})$ subject to $\underset{\sim}{x} \epsilon S$ is equivalent to the problem minimize x_{n+1} subject to $\underset{\sim}{x} \epsilon S$, $x_{n+1} - f(\underset{\sim}{x}) \geq 0$ where x_{n+1} is a new independent variable (note that the new constraint is concave). This equivalence follows from the Kuhn Tucker conditions by noting that the new constraint must be active. Thus a convex programming problem can be replaced by the problem of minimizing a linear objective function subject to an enlarged constraint set.

(c) Consider the problem of minimizing $\underset{\sim}{c}^T \underset{\sim}{x}$ subject to $\underset{\sim}{x} \epsilon S$ and S bounded. In particular we assume that $S \subseteq R_0 = \{\underset{\sim}{x} ; A\underset{\sim}{x} - \underset{\sim}{b} \geq 0\}$. We can now state the cutting plane algorithm

(0) $i = 0$.

(i) Let $\underset{\sim}{x}_i$ minimize $\underset{\sim}{c}^T \underset{\sim}{x}$ subject to $\underset{\sim}{x} \epsilon R_i$.

(ii) Determine α such that $g_\alpha(\underset{\sim}{x}_i) \leq g_j(\underset{\sim}{x}_i)$, $j \epsilon I_1$.

(iii) If $g_\alpha(\underset{\sim}{x}_i) \geq 0$ go to (v).

(iv) Set $R_{i+1} = R_i \cap \{\underset{\sim}{x} ; g_\alpha(\underset{\sim}{x}_i) + \nabla g_\alpha(\underset{\sim}{x}_i)(\underset{\sim}{x} - \underset{\sim}{x}_i) \geq 0\}$,
$i := i+1$, go to (i)

(v) Stop.

Note that step (i) requires the solution of a linear programming problem.

(d) The cutting plane algorithm generates a sequence of points $\underset{\sim}{x}_i$ with the property that

$$\underset{\sim}{c}^T \underset{\sim}{x}_0 \leq \underset{\sim}{c}^T \underset{\sim}{x}_1 \leq \cdots \leq \underset{\sim}{c}^T \underset{\sim}{x}_i \leq \cdots \leq \min_{\underset{\sim}{x} \epsilon S} \underset{\sim}{c}^T \underset{\sim}{x}$$

as $R_0 \supseteq R_1 \supseteq \cdots \supseteq S$. Thus the sequence $\{\underset{\sim}{c}^T \underset{\sim}{x}_i\}$ is increasing and

bounded above and therefore convergent. Let $\underset{\sim}{x}^*$ be a limit point of the $\{\underset{\sim}{x}_i\}$. Then $\underset{\sim}{x}^* \epsilon S$ and therefore solves the convex programming problem. To prove this, assume $\underset{\sim}{x}^* \notin S$. Then

$$\min_i g_i(\underset{\sim}{x}^*) = g_\alpha(\underset{\sim}{x}^*) = -\lambda < 0 \quad .$$

Let a subsequence $\{\underset{\sim}{x}_j\} \to \underset{\sim}{x}^*$, then $\exists$ k such that

(i) $\quad \|\underset{\sim}{x}_k - \underset{\sim}{x}^*\| < \frac{\lambda}{2C}$, and

(ii) $\quad g_\alpha(\underset{\sim}{x}_k) < -\frac{\lambda}{2}$

where $C \geq \|\nabla g_i(\underset{\sim}{x})\|$, $x \epsilon R_0$, $i \epsilon I_1$.

Let

$$\min_i g_i(\underset{\sim}{x}_k) = g_\beta(\underset{\sim}{x}_k) \quad .$$

Then $g_\beta(\underset{\sim}{x}_k) < -\frac{\lambda}{2}$. Now $\underset{\sim}{x}^*$ a limit point of $\{\underset{\sim}{x}_i\} \Rightarrow \underset{\sim}{x}^* \epsilon \cap R_i$. In particular, $\underset{\sim}{x}^* \epsilon R_{k+1}$ whence

$$g_\beta(\underset{\sim}{x}_k) + \nabla g_\beta(\underset{\sim}{x}_k)(\underset{\sim}{x}^* - \underset{\sim}{x}_k) \geq 0 \quad .$$

But

$$\|\nabla g_\beta(\underset{\sim}{x}_k)(\underset{\sim}{x}^* - \underset{\sim}{x}_k)\| < \frac{\lambda}{2C} C < \frac{\lambda}{2}$$

so that

$$g_\beta(\underset{\sim}{x}_k) + \nabla g_\beta(\underset{\sim}{x}_k)(\underset{\sim}{x}^* - \underset{\sim}{x}_k) < -\frac{\lambda}{2} + \|\nabla g_\beta(\underset{\sim}{x}_k)(\underset{\sim}{x}^* - \underset{\sim}{x}_k)\|$$

$$< 0$$

which gives a contradiction.

Notes

1. For properties of tangent cones, see Hestenes. Luenberger discusses polar cones (which he calls negative conjugate cones) on pp. 157-159. Lemma 1.4 is due to Gould and Tolle. The proof is due to Nashed et al.

2. Hestenes is a good general reference for this section and includes a proof that a finitely generated cone is closed. The proof given here of Farkas Lemma is standard (see for example Vajda's paper). An extensive list of alternative theorems is given in Mangasarian.

3. The main result is due to Gould and Tolle. The treatment of the other restraint conditions follows Fiacco and McCormick.

4. The treatment of second order conditions is based on Hestenes. Similar material is given in Fiacco and McCormick.

5. The treatment of duality is based on Luenberger. A related treatment is given by Whittle who is good value on applications. Vajda is a good reference for the mathematical programming application. Wolfe's papers in both the Abadie books discuss various aspects of the cutting plane method.

References

J. Abadie (Editor) (1967): Nonlinear Programming.

(1970): Integer and Nonlinear Programming. North Holland.

M. S. Bazaraa, J. J. Goode, M. Z. Nashed, and C. M. Shetty (1971): Nonlinear programming without differentiability in Banach Spaces: Necessary and Sufficient Constraint Qualification. To appear in Applicable Analysis.

A. V. Fiacco and G. P. McCormick (1968): Nonlinear Programming and Sequential Unconstrained Minimization Techniques. Wiley.

F. J. Gould and J. W. Tolle (1971): A necessary and sufficient constraint qualification for constrained optimization, SIAM Applied Maths., 20, pp. 164-172.

M. R. Hestenes (1966): Calculus of Variations and Optimal Control Theory. Wiley.

O. L. Mangasarian (1969): Nonlinear Programming. McGraw-Hill.

Peter Whittle (1971): Optimization under Constraints. Wiley.

II. Descent Methods for Unconstrained Minimization

1. General properties of descent methods.

The class of descent methods for minimizing an unconstrained function $F(\underset{\sim}{x})$ solve the problem iteratively by means of a sequence of one dimensional minimizations. The main idea is illustrated in Figure 1.1. At the current point $\underset{\sim}{x}_i$ a direction $\underset{\sim}{t}_i$ is provided, and the closest minimum to $\underset{\sim}{x}_i$ of the function

$$G_i(\lambda) = F(\underset{\sim}{x}_i + \lambda \underset{\sim}{t}_i)$$

sought. At $\underset{\sim}{x}_{i+1}$ we have

$$G_i'(\lambda_i) = \nabla F(\underset{\sim}{x}_{i+1})\underset{\sim}{t}_i = 0 \tag{1.1}$$

where $\underset{\sim}{x}_{i+1} = \underset{\sim}{x}_i + \lambda_i \underset{\sim}{t}_i$.

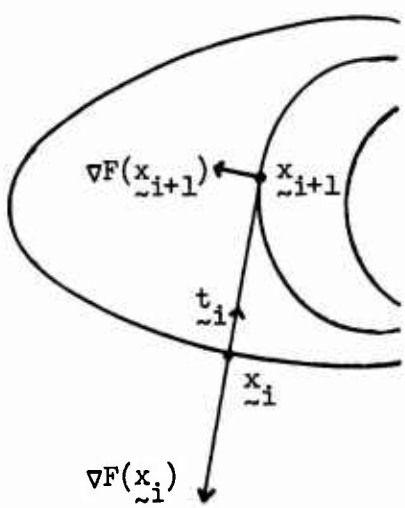

Figure 1.1

Definition: A step in which $\underset{\sim}{x}_{i+1}$ is determined by satisfying the above conditions is said to satisfy the descent condition. We consider $\underset{\sim}{t}_i$ a profitable search direction if $F(\underset{\sim}{x}_i + \lambda \underset{\sim}{t}_i)$ decreases initially as λ increases from zero. This condition is formalized as follows.

Definition: (i) The vector $\underset{\sim}{t}$ is downhill for minimizing F at $\underset{\sim}{x}$ if $\nabla F(\underset{\sim}{x})\underset{\sim}{t} < 0$. (ii) The sequence of unit vectors $\{\underset{\sim}{t}_i\}$ is downhill for minimizing F at the sequence of points $\{\underset{\sim}{x}_i\}$ if $\exists\, \delta > 0$, independent of i , such that $\nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i \leq -\delta\|\nabla F(\underset{\sim}{x}_i)\|$.

Example: The sequence of vectors $\{-\nabla F(\underset{\sim}{x}_i) / \|\nabla F(\underset{\sim}{x}_i)\|\}$ satisfies the downhill condition with $\delta = 1$. In this case we say that $\underset{\sim}{t}_i$ is in the direction of steepest descent.

An estimate of the value of λ minimizing G_i is readily given. We have

$$0 = \nabla F(\underset{\sim}{x}_{i+1})\underset{\sim}{t}_i = \nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i + \lambda_i\, \underset{\sim}{t}_i^T\, \nabla^2 F(\bar{\underset{\sim}{x}}_i)t_i$$

where $\bar{\underset{\sim}{x}}_i = \underset{\sim}{x}_i + \bar{\lambda}_i\underset{\sim}{t}_i$ is an appropriate mean value. Thus

$$\lambda_i = \frac{-\nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i}{\underset{\sim}{t}_i^T\, \nabla^2 F(\underset{\sim}{x}_i)\underset{\sim}{t}_i} \geq \frac{\delta\|\nabla F(\underset{\sim}{x}_i)\|}{\|\nabla^2 F(\bar{\underset{\sim}{x}}_i)\|} \quad . \tag{1.2}$$

Theorem 1.1: (Ostrowski's descent theorem). Let $R = \{\underset{\sim}{x}\,;\, F(\underset{\sim}{x}) \leq F^*\}$, and assume that F bounded below and $\underset{\sim}{t}^T \nabla^2 F(\underset{\sim}{x})\underset{\sim}{t} \leq K\|\underset{\sim}{t}\|^2$, $\underset{\sim}{x} \epsilon R$. Define

$$\underset{\sim}{x}_{i+1} = \underset{\sim}{x}_i + \frac{\delta\|\nabla F(\underset{\sim}{x}_i)\|}{K}\underset{\sim}{t}_i \,, \qquad \text{and}$$

$$\nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i \leq -\delta\|\nabla F(\underset{\sim}{x}_i)\| \,, \quad \|\underset{\sim}{t}_i\| = 1 \quad ,$$

for $i = 1,2,\ldots$ where $\delta > 0$. Then $\{F(\underset{\sim}{x}_i)\}$ converges, and the limit points of $\{\underset{\sim}{x}_i\}$ are stationary values of F .

Proof: As $\{\underset{\sim}{t}_i\}$ downhill then $\{\underset{\sim}{x}_i\} \subset R$. Expanding by the mean value theorem we obtain

$$F(\underset{\sim}{x}_{i+1}) = F(\underset{\sim}{x}_i) + \frac{\delta\|\nabla F(\underset{\sim}{x}_i)\|}{K} \nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i + \frac{1}{2}\left(\frac{\delta\|\nabla F(\underset{\sim}{x}_i)\|}{K}\right)^2 \underset{\sim}{t}_i^T \nabla^2 F(\bar{\underset{\sim}{x}}_i)\underset{\sim}{t}_i$$

where $\bar{\underset{\sim}{x}}_i$ is a mean value. We have

$$F(\underset{\sim}{x}_{i+1}) \leq F(\underset{\sim}{x}_i) - \frac{(\delta\|\nabla F(\underset{\sim}{x}_i)\|)^2}{K} + \frac{1}{2}\left(\frac{\delta\|\nabla F(\underset{\sim}{x}_i)\|}{K}\right)^2 K$$

$$\leq F(\underset{\sim}{x}_i) - \frac{\delta^2\|\nabla F(\underset{\sim}{x}_i)\|^2}{K} \quad . \tag{1.3}$$

Thus the sequence $\{F(\underset{\sim}{x}_i)\}$ is decreasing and bounded below and therefore convergent. Further, from (1.3),

$$\|\nabla F(\underset{\sim}{x}_i)\| \leq \frac{1}{\delta}\sqrt{2K(F(\underset{\sim}{x}_i) - F(\underset{\sim}{x}_{i+1}))} \tag{1.4}$$

$$\to 0 \ , \quad i \to \infty \ .$$

Thus $\nabla F(\underset{\sim}{x}^*) = 0$ if $\underset{\sim}{x}^*$ is a limit point of $\{\underset{\sim}{x}_i\}$. □

Remark: By (1.2) the step taken in the direction $\underset{\sim}{t}_i$ underestimates the step to the minimum of G_i . Thus (1.3) holds if the descent condition is satisfied so that the conclusions of the theorem are valid also in this case.

Theorem 1.2: (Goldstein's descent theorem). Let $R = \{\underset{\sim}{x} ; F(\underset{\sim}{x}) \leq F^*\}$ be bounded, and assume $F \epsilon C^1$ and bounded below on R . Define

$$\Delta(\underset{\sim}{x}_i,\lambda) = F(\underset{\sim}{x}_i) - F(\underset{\sim}{x}_i + \lambda \underset{\sim}{t}_i) \quad ,$$

$$\psi(\underset{\sim}{x}_i,\lambda) = - \frac{\Delta(\underset{\sim}{x}_i,\lambda)}{\lambda \nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i}$$

where $\{\underset{\sim}{t}_i\}$ downhill, and the $\{\underset{\sim}{x}_i\}$ are generated by the algorithm

(i) $\underset{\sim}{x}_{i+1} = \underset{\sim}{x}_i$ if $\Delta(\underset{\sim}{x}_i,\lambda) = 0$.

(ii) If $\psi(\underset{\sim}{x}_i,1) < \sigma$ where $0 < \sigma < 1/2$
then choose λ_i such that $\sigma \leq \psi(\underset{\sim}{x}_i,\lambda_i) \leq 1-\sigma$,
else choose $\lambda_i = 1$.

(iii) $\underset{\sim}{x}_{i+1} = \underset{\sim}{x}_i + \lambda_i \underset{\sim}{t}_i$.

Then the limit points of $\{\underset{\sim}{x}_i\}$ are stationary points of F .

<u>Proof</u>: $\Delta(\underset{\sim}{x}_i,\lambda) = -\lambda \nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i + o(\lambda)$.

Thus $\Delta(\underset{\sim}{x}_i,\lambda) = 0 \Rightarrow \|\nabla F(\underset{\sim}{x}_i)\| = 0$ as $\{\underset{\sim}{t}_i\}$ downhill so that $\underset{\sim}{x}_i$ is a stationary point. Otherwise $\nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i < 0$ so that $\psi(\underset{\sim}{x}_i,\lambda) = 1 + o(1)$ whence $\psi(\underset{\sim}{x}_i,0) = 1$. Also the boundedness of R implies that $\Delta(\underset{\sim}{x}_i,\lambda) < 0$ for some λ large enough so that, as $\psi(\underset{\sim}{x}_i,\lambda)$ is continuous, λ_i can be found to satisfy condition (ii) of the algorithm. Note that $\{\underset{\sim}{x}_i\} \subset R$. We have

$$F(\underset{\sim}{x}_i) - F(\underset{\sim}{x}_{i+1}) = -\lambda_i \psi(\underset{\sim}{x}_i,\lambda_i)\nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i$$

$$\geq \lambda_i \sigma \delta \|\nabla F(\underset{\sim}{x}_i)\| \quad . \tag{1.5}$$

Thus $\{F(\underset{\sim}{x}_i)\}$ decreasing and bounded below and therefore convergent. To show that the limit points of $\{\underset{\sim}{x}_i\}$ are stationary values of F consider

the subsequence $\{\underset{\sim}{x}_{\mu_i}\} \to \underset{\sim}{x}^*$ and assume $\|\nabla F(\underset{\sim}{x}^*)\| > \varepsilon > 0$. Then $\|\nabla F(\underset{\sim}{x}_{\mu_i})\| > \varepsilon$ for $i > i_0$. This implies that $\inf_i \lambda_{\mu_i} = \lambda_0 > 0$ as otherwise $\sup_i \psi(\underset{\sim}{x}_{\mu_i}, \lambda_{\mu_i}) = 1$ contradicting $\psi(\underset{\sim}{x}_i, \lambda_i) \leq 1 - \sigma$. Thus

$$\|\nabla F(\underset{\sim}{x}_{\mu_i})\| \leq \frac{F(\underset{\sim}{x}_{\mu_i}) - F(\underset{\sim}{x}_{\mu_i+1})}{\lambda_0 \sigma \delta} . \tag{1.6}$$

The right hand side $\to 0$ as $i \to \infty$ which establishes a contradiction. □

<u>Remark</u>: There are two aspects of this theorem which are of particular interest. (i) It is necessary to assume only that $f \in C^1$ in R . However, the boundedness of R is used explicitly. (ii) The algorithm for determining the step length λ_i is <u>readily implemented</u>. A value of λ satisfying condition (ii) of the algorithm will be said to satisfy the <u>Goldstein condition</u>.

<u>Theorem 1.3</u>: (i) Let the vector sequence in the Goldstein algorithm be defined by

$$\underset{\sim}{s}_i = -A_i \nabla F(\underset{\sim}{x}_i)^T , \quad \underset{\sim}{t}_i = \underset{\sim}{s}_i / \|\underset{\sim}{s}_i\| \tag{1.7}$$

where A_i is positive definite, bounded, and $X(A_i) = \dfrac{\lambda_{min}(A_i)}{\lambda_{max}(A_i)} \geq \omega > 0$, $i = 1,2,\ldots$. Then $\{\underset{\sim}{t}_i\}$ is downhill with constant $\delta = \omega$.

(ii) Assume that $\{\underset{\sim}{x}_i\} \to \underset{\sim}{x}^*$, and that $\|A_i^{-1} - \nabla^2 F(\underset{\sim}{x}_i)\| = o(1)$, then $\lambda_i = \|\underset{\sim}{s}_i\|$ satisfies the Goldstein condition for i large enough.

(iii) The ultimate rate of convergence of the algorithm is superlinear for this choice of λ_i .

Proof: (i)

$$\nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i = - \frac{\nabla F(\underset{\sim}{x}_i)A_i \nabla F(\underset{\sim}{x}_i)^T}{\|\nabla F(\underset{\sim}{x}_i)A_i\|}$$

$$\leq - \frac{\lambda_{min}(A_i)\|\nabla F(x_i)\|}{\lambda_{max}(A)}$$

$$\leq - \omega\|\nabla F(\underset{\sim}{x}_i)\| \quad . \tag{1.8}$$

(ii) $$\psi(\underset{\sim}{x}_i,\lambda) = - \frac{F(\underset{\sim}{x}_i) - F(\underset{\sim}{x}_i + \lambda \underset{\sim}{t}_i)}{\lambda \nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i}$$

$$= \frac{\lambda \nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i + \frac{\lambda^2}{2} \underset{\sim}{t}_i^T \nabla^2 F(\bar{\underset{\sim}{x}}_i)\underset{\sim}{t}_i}{\lambda \nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i}$$

where $\bar{\underset{\sim}{x}}_i$ is a mean value dependent on λ . Now, writing

$$\nabla^2 F(\bar{\underset{\sim}{x}}_i) = A_i^{-1} + E_i$$

and noting that $\|E_i\| \to 0$ as $i \to \infty$, we have

$$\psi(\underset{\sim}{x}_i,\lambda) = 1 - \frac{\lambda}{2}\left\{\frac{1}{\|\underset{\sim}{s}_i\|} + \frac{\underset{\sim}{t}_i^T E_i \underset{\sim}{t}_i}{\nabla F(\underset{\sim}{x}_i)\underset{\sim}{t}_i}\right\}$$

so that

$$\left|\psi(\underset{\sim}{x}_i,\lambda) - \left(1 - \frac{\lambda}{2\|\underset{\sim}{s}_i\|}\right)\right| \leq \frac{\lambda}{2\|\underset{\sim}{s}_i\|} \frac{\|E_i\| \, \|\underset{\sim}{s}_i\|}{\omega\|\nabla F(\underset{\sim}{x}_i)\|}$$

$$\leq \frac{\lambda}{2\|\underset{\sim}{s}_i\|} \frac{\|E_i\| \, \|A_i\|}{\omega} \quad . \tag{1.9}$$

In particular

$$|\psi(\underset{\sim}{x}_i, \|\underset{\sim}{s}_i\|) - \frac{1}{2}| \leq \frac{1}{2} \frac{\|E_i\| \, \|A_i\|}{\omega} \to 0 , \quad i \to \infty .$$

(iii) Another application of the mean value theorems gives

$$\begin{aligned} \underset{\sim}{s}_i &= - A_i \nabla F(\underset{\sim}{x}_i)^T = - A_i(\nabla F(\underset{\sim}{x}_i) - \nabla F(\underset{\sim}{x}^*))^T \\ &= - A_i(\nabla^2 F(\underset{\sim}{x}_i)(\underset{\sim}{x}_i - \underset{\sim}{x}^*) + o(\|\underset{\sim}{x}_i - \underset{\sim}{x}^*\|)) \\ &= - (\underset{\sim}{x}_i - \underset{\sim}{x}^*) + o(\|\underset{\sim}{x}_i - \underset{\sim}{x}^*\|) \quad . \end{aligned} \tag{1.10}$$

Thus

$$\begin{aligned} \underset{\sim}{x}_{i+1} - \underset{\sim}{x}^* &= \underset{\sim}{x}_i + \gamma_i \underset{\sim}{s}_i - \underset{\sim}{x}^* \\ &= (1 - \gamma_i)(\underset{\sim}{x}_i - \underset{\sim}{x}^*) + o(\|\underset{\sim}{x}_i - \underset{\sim}{x}^*\|) \quad . \end{aligned} \tag{1.11}$$

From (1.11) the choice $\gamma_i = 1$ $(\lambda_i = \|\underset{\sim}{s}_i\|)$ gives superlinear convergence. □

<u>Remark</u>: Theorem 1.3 shows that if $\nabla^2 F$ is positive definite in the neighbourhood of an unconstrained minimum, then it is possible to have algorithms with superlinear convergence without the necessity of satisfying the descent condition. It is not generally considered economic to compute the second partial derivatives of F, and considerable emphasis has been placed on developing approximations to the inverse Hessian using only first derivative information. Although the steepest descent direction is initially in the direction of most rapid decrease of the function it gives in general only linear convergence.

2. Methods based on conjugate directions.

The problem of minimizing a positive definite quadratic form is an important special case of the general unconstrained optimization problem. In particular it is frequently used as a model problem for the development of new algorithms. It is argued that in a neighborhood of the minimum, a general function having a positive definite Hessian at the minimum will be well represented by a quadratic form so that methods which work well in this particular case should work well in general.

Let F be given by

$$F(\underset{\sim}{x}) = a + \underset{\sim}{b}^T \underset{\sim}{x} + \frac{1}{2} \underset{\sim}{x}^T C \underset{\sim}{x} \tag{2.1}$$

where C is a positive definite, necessarily symmetric matrix. We have

$$\nabla F(\underset{\sim}{x}) = \underset{\sim}{b}^T + \underset{\sim}{x}^T C \quad . \tag{2.2}$$

Consider now a descent step from $\underset{\sim}{x}_i$ in the direction $\underset{\sim}{t}_i$. The descent condition gives

$$0 = \nabla F(\underset{\sim}{x}_{i+1}) \underset{\sim}{t}_i = \underset{\sim}{t}_i^T (C(\underset{\sim}{x}_i + \lambda \underset{\sim}{t}_i) + \underset{\sim}{b})$$

whence

$$\lambda = - \frac{\underset{\sim}{t}_i^T \underset{\sim}{g}_i}{\underset{\sim}{t}_i^T C \underset{\sim}{t}_i} \tag{2.3}$$

where $g_i = \nabla F(\underset{\sim}{x}_i)^T$. To calculate the change in the value of F in a descent step we have

$$F(\underset{\sim}{x}_i + \lambda_i \underset{\sim}{t}_i) - F(\underset{\sim}{x}_i) = \lambda_i \underset{\sim}{b}^T \underset{\sim}{t}_i + \lambda_i \underset{\sim}{x}_i^T C \underset{\sim}{t}_i + \frac{1}{2} \lambda_i^2 \underset{\sim}{t}_i^T C \underset{\sim}{t}_i$$

$$= \lambda_i \underset{\sim}{g}_i^T \underset{\sim}{t}_i + \frac{1}{2} \lambda_i^2 \underset{\sim}{t}_i^T C \underset{\sim}{t}_i$$

$$= - \frac{1}{2} \frac{(\underset{\sim}{g}_i^T \underset{\sim}{t}_i)^2}{\underset{\sim}{t}_i^T C \underset{\sim}{t}_i} \quad . \tag{2.4}$$

<u>Example</u>: (Linear convergence of the method of steepest descent).

Let $F = \frac{1}{2} \underset{\sim}{x}^T C \underset{\sim}{x}$. Then (2.4) gives

$$F(\underset{\sim}{x}_{i+1}) - F(\underset{\sim}{x}_i) = - \frac{1}{2} \frac{(\underset{\sim}{x}_i^T C^2 \underset{\sim}{x}_i)^2}{\underset{\sim}{x}_i^T C^3 \underset{\sim}{x}_i}$$

$$= - \frac{1}{2} \frac{(\underset{\sim}{w}_i^T \underset{\sim}{w}_i)^2}{\underset{\sim}{w}_i^T C \underset{\sim}{w}_i}$$

where $\underset{\sim}{w}_i = C\underset{\sim}{x}_i$.

We have $F(\underset{\sim}{x}_i) = \frac{1}{2} \underset{\sim}{w}_i^T C^{-1} \underset{\sim}{w}_i$ so that

$$F(\underset{\sim}{x}_{i+1}) = \left(1 - \frac{1}{2} \frac{(\underset{\sim}{w}_i^T \underset{\sim}{w}_i)^2}{\underset{\sim}{w}_i^T C \underset{\sim}{w}_i \; \underset{\sim}{w}_i^T C^{-1} \underset{\sim}{w}_i} \right) F(\underset{\sim}{x}_i) \quad .$$

The Kantorovich inequality gives

$$\frac{(\underset{\sim}{w}^T \underset{\sim}{w})^2}{\underset{\sim}{w}^T C^{-1} \underset{\sim}{w} \; \underset{\sim}{w}^T C \underset{\sim}{w}} \geq \frac{4 \sigma_1 \sigma_n}{(\sigma_1 + \sigma_n)^2}$$

where σ_1 and σ_n are the smallest and largest eigenvalues of C respectively, whence

$$F(\underset{\sim}{x}_{i+1}) \leq \left(\frac{\sigma_n - \sigma_1}{\sigma_n + \sigma_1}\right)^2 F(\underset{\sim}{x}_i)$$

which shows that the rate of convergence of steepest descent is at least linear.

To show that it is exactly linear consider the particular case in which

$$\underset{\sim}{x}_i = \alpha_1^i \underset{\sim}{v}_1 + \alpha_n^i \underset{\sim}{v}_n$$

where $\underset{\sim}{v}_1$ and $\underset{\sim}{v}_n$ are the normalized eigenvectors associated with σ_1 and σ_n respectively. We have

$$\underset{\sim}{x}_{i+1} = \alpha_1^{i+1} \underset{\sim}{v}_1 + \alpha_n^{i+1} \underset{\sim}{v}_n = (1 - \lambda_i \sigma_1)\alpha_1^i \underset{\sim}{v}_1 + (1 - \lambda_i \sigma_n)\alpha_n^i \underset{\sim}{v}_n$$

with (from (2.3))

$$\lambda_i = \frac{(\alpha_1^i)^2 \sigma_1^2 + (\alpha_n^i)^2 \sigma_n^2}{(\alpha_1^i)^2 \sigma_1^3 + (\alpha_n^i)^2 \sigma_n^3} ,$$

so that

$$\alpha_1^{i+1} = \frac{(\alpha_n^i)^2 (\sigma_n - \sigma_1)\sigma_n^2}{(\alpha_1^i)^2 \sigma_1^3 + (\alpha_n^i)^2 \sigma_n^3} \alpha_1^i$$

and

$$\alpha_n^{i+1} = - \frac{(\alpha_1^i)^2 (\sigma_n - \sigma_1) \sigma_1^2}{(\alpha_1^i)^2 \sigma_1^3 + (\alpha_n^i)^2 \sigma_n^3} \alpha_n^i .$$

In particular

$$\frac{\alpha_1^{i+1}}{\alpha_n^{i+1}} = - \frac{\sigma_n^2}{\sigma_1^2} \frac{\alpha_n^i}{\alpha_1^i} = \frac{\alpha_1^{i-1}}{\alpha_n^{i-1}}$$

so that the ratios α_1^i/α_n^i assume just two values for all i (depending on i even or odd). Now

$$|\alpha_1^{i+1}| \geq \left(1 - \frac{\sigma_1}{\sigma_n}\right) \frac{(\alpha_n^i)^2}{(\alpha_1^i)^2 + (\alpha_n^i)^2} |\alpha_1^i| ,$$

and

$$|\alpha_n^{i+1}| \geq \left(1 - \frac{\sigma_1}{\sigma_n}\right)\left(\frac{\sigma_1}{\sigma_n}\right)^2 \frac{(\alpha_1^i)^2}{(\alpha_1^i)^2 + (\alpha_n^i)^2} |\alpha_n^i|$$

so that

$$|\alpha_1^{i+1}| + |\alpha_n^{i+1}| \geq \left(1 - \frac{\sigma_1}{\sigma_n}\right)\left(\frac{\sigma_1}{\sigma_n}\right)^2 \frac{\alpha_1^i \alpha_n^i}{(\alpha_1^i)^2 + (\alpha_n^i)^2} (|\alpha_1^i| + |\alpha_n^i|)$$

$$\geq \left(1 - \frac{\sigma_1}{\sigma_n}\right)\left(\frac{\sigma_1}{\sigma_n}\right)^2 \gamma(|\alpha_1^i| + |\alpha_n^i|)$$

where $\gamma = \min\left\{ \dfrac{\left|\dfrac{\alpha_1^i}{\alpha_n^i}\right|}{1 + \left|\dfrac{\alpha_1^i}{\alpha_n^i}\right|} , \dfrac{\left|\dfrac{\alpha_1^{i+1}}{\alpha_n^{i+1}}\right|}{1 + \left|\dfrac{\alpha_1^{i+1}}{\alpha_n^{i+1}}\right|} \right\} < 1$, and γ is independent of i. This inequality shows that the rate of convergence of steepest descent is linear.

<u>Definition</u>: Directions $\underset{\sim}{t}_1, \underset{\sim}{t}_2$ are <u>conjugate</u> with respect to C if

$$\underset{\sim}{t}_1^T C \underset{\sim}{t}_2 = 0 \quad . \tag{2.5}$$

In what follows it will frequently be convenient to speak about a

'direction of search' without intending to imply that its norm is unity. However, the null vector is excluded from any set of mutually conjugate directions. It is clear that any set of mutually conjugate directions are linearly independent.

Example: The eigenvectors of C are conjugate. The property of being both conjugate and orthogonal specializes the eigenvectors.

Lemma 2.1: Let $\underset{\sim}{t}_1,\ldots,\underset{\sim}{t}_n$ be a set of mutually conjugate directions (with respect to C). Starting from $\underset{\sim}{x}_1$ let $\underset{\sim}{x}_2,\underset{\sim}{x}_3,\ldots,\underset{\sim}{x}_{n+1}$ be points produced by descent steps applied to (2.1). Then

$$\underset{\sim}{g}_i^T \underset{\sim}{t}_j = 0 \quad , \quad j = 1,2,\ldots,i-1 \quad . \tag{2.6}$$

Proof: The descent condition gives $\underset{\sim}{g}_i^T \underset{\sim}{t}_{i-1} = 0$ so it is necessary only to verify the result for $j < i-1$. We have

$$\begin{aligned}
\underset{\sim}{g}_i^T \underset{\sim}{t}_s &= (C\underset{\sim}{x}_i + \underset{\sim}{b})^T \underset{\sim}{t}_s \\
&= (C\underset{\sim}{x}_{s+1} + \underset{\sim}{b} + \sum_{k=s+1}^{i-1} \lambda_k C \underset{\sim}{t}_k)^T \underset{\sim}{t}_s \\
&= \underset{\sim}{g}_{s+1}^T \underset{\sim}{t}_s + \sum_{k=s+1}^{i-1} \lambda_k \underset{\sim}{t}_k^T C \underset{\sim}{t}_s \quad , \quad s = 1,2,\ldots,i-2 \\
&= 0 \quad . \quad \square
\end{aligned}$$

Corollary 2.1: The minimum of a positive definite quadratic form can be found by making at most one descent step along each of n mutually conjugate directions.

Proof: From Lemma 2.1 we have $\underset{\sim}{g}_{n+1}^T \underset{\sim}{t}_i = 0$, $i = 1,2,\ldots,n$. Thus $\underset{\sim}{g}_{n+1}$ is orthogonal to n linearly independent directions and therefore vanishes identically. □

Remark: A method which minimizes a quadratic form in a finite number of steps is said to have a quadratic termination property.

Example: The sequence of vectors

$$\underset{\sim}{t}_1 = - \underset{\sim}{g}_1 \quad ,$$

$$\underset{\sim}{t}_i = - \underset{\sim}{g}_i + \frac{\|\underset{\sim}{g}_i\|^2}{\|\underset{\sim}{g}_{i-1}\|^2} \underset{\sim}{t}_{i-1} \quad , \quad i = 2,\ldots,n \tag{2.7}$$

are conjugate. The algorithm based on this choice is called the method of conjugate gradients.

We now consider the generation of sequences of conjugate directions to provide a basis for a descent calculation. To do this we note that the minimum of (2.1) is at $\underset{\sim}{x} = -C^{-1}\underset{\sim}{b}$ so that if we minimize in the direction $\underset{\sim}{t} = -C^{-1}(C\underset{\sim}{x}_1 + \underset{\sim}{b}) = -C^{-1}\nabla F(\underset{\sim}{x}_1)$ then the minimum is found in a single step. In general C^{-1} is not known in advance, so that we are lead to consider processes in which each step consists of two parts (i) a descent calculation in the direction

$$\underset{\sim}{t}_i = - H_i \underset{\sim}{g}_i \tag{2.8}$$

where H_i is the current estimate of C^{-1} , and (ii) the calculation of a correction to H_i which serves both the purposes of making the $\underset{\sim}{t}_i$ conjugate and making H_i approach C^{-1} . It is convenient in what follows to assume that the H_i are symmetric. This seems a natural condition given the symmetry of C but is in fact not necessary.

If we assume that $\underset{\sim}{t}_s$, $s < i$, are mutually conjugate then the condition that each be conjugate to $\underset{\sim}{t}_i$ is

$$\underset{\sim}{t}_i^T C \underset{\sim}{t}_s = \underset{\sim}{g}_i^T H_i C \underset{\sim}{t}_s = 0 \quad , \quad s < i ,$$

and, by Lemma 2.1, this is certainly satisfied if

$$H_i C \underset{\sim}{t}_s = \rho_s \underset{\sim}{t}_s \quad , \quad s < i .$$

We write this equation in the equivalent form (multiplying both sides by λ_s)

$$H_i \underset{\sim}{y}_s = \rho_s \underset{\sim}{d}_s \quad , \quad s < i , \tag{2.9}$$

where

$$\underset{\sim}{d}_i = \underset{\sim}{x}_{i+1} - \underset{\sim}{x}_i \quad , \quad \underset{\sim}{y}_i = \underset{\sim}{g}_{i+1} - \underset{\sim}{g}_i \quad . \tag{2.10}$$

Consider the symmetric updating formula

$$H_{i+1} = H_i + \xi_i \underset{\sim}{d}_i \underset{\sim}{d}_i^T + \eta_i H_i \underset{\sim}{y}_i \underset{\sim}{y}_i^T H_i - \zeta_i (\underset{\sim}{d}_i \underset{\sim}{y}_i^T H_i + H_i \underset{\sim}{y}_i \underset{\sim}{d}_i^T) \tag{2.11}$$

where ξ_i , η_i , ζ_i are to be determined (or prescribed). We have $H_{i+1} \underset{\sim}{y}_s = H_i \underset{\sim}{y}_s = \rho_s \underset{\sim}{d}_s$, $s < i$, provided (2.9) holds as $\underset{\sim}{d}_i^T \underset{\sim}{y}_s = \lambda_i \lambda_s \underset{\sim}{t}_i^T C \underset{\sim}{t}_s = 0$, and $\underset{\sim}{y}_i^T H_i \underset{\sim}{y}_s = \lambda_s \underset{\sim}{y}_i^T H_i C \underset{\sim}{t}_s = \rho_s \lambda_s \underset{\sim}{y}_i^T \underset{\sim}{d}_s = \rho_s \lambda_s \lambda_i \underset{\sim}{d}_i^T C \underset{\sim}{d}_s = 0$. Thus (2.9) is satisfied for $i := i+1$ if

$$0 = 1 + \eta_i (\underset{\sim}{y}_i^T H_i \underset{\sim}{y}_i) - \zeta_i (\underset{\sim}{d}_i^T \underset{\sim}{y}_i) \quad , \tag{2.12}$$

and

$$\rho_i = \xi_i (\underset{\sim}{d}_i^T \underset{\sim}{y}_i) - \zeta_i (\underset{\sim}{y}_i^T H_i \underset{\sim}{y}_i) \quad . \tag{2.13}$$

If ξ_i and η_i are expressed in terms of ρ_i and ζ_i from (2.12) and (2.13) we have

$$\eta_i = -\frac{1}{\underset{\sim}{y}_i^T H_i \underset{\sim}{y}_i} + \zeta_i \frac{\underset{\sim}{d}_i^T \underset{\sim}{y}_i}{\underset{\sim}{y}_i^T H_i \underset{\sim}{y}_i} ,$$

and

$$\xi_i = \frac{\rho_i}{\underset{\sim}{d}_i^T \underset{\sim}{y}_i} + \zeta_i \frac{\underset{\sim}{y}_i^T H_i \underset{\sim}{y}_i}{\underset{\sim}{d}_i^T \underset{\sim}{y}_i} ,$$

so that equation (2.11) becomes

$$H_{i+1} = H_i + \rho_i \frac{\underset{\sim}{d}_i \underset{\sim}{d}_i^T}{\underset{\sim}{d}_i^T \underset{\sim}{y}_i} - \frac{H_i \underset{\sim}{y}_i \underset{\sim}{y}_i^T H_i}{\underset{\sim}{y}_i^T H_i \underset{\sim}{y}_i}$$

$$+ \zeta_i \left\{ \frac{\underset{\sim}{y}_i^T H_i \underset{\sim}{y}_i}{\underset{\sim}{d}_i^T \underset{\sim}{y}_i} \underset{\sim}{d}_i \underset{\sim}{d}_i^T + \frac{\underset{\sim}{d}_i^T \underset{\sim}{y}_i}{\underset{\sim}{y}_i^T H_i \underset{\sim}{y}_i} H_i \underset{\sim}{y}_i \underset{\sim}{y}_i^T H_i - \underset{\sim}{d}_i \underset{\sim}{y}_i^T H_i - H_i \underset{\sim}{y}_i \underset{\sim}{d}_i^T \right\}$$

$$= D(\rho_i, H_i) + \zeta_i \tau_i \underset{\sim}{v}_i \underset{\sim}{v}_i^T \tag{2.14}$$

where

$$\underset{\sim}{v}_i = \underset{\sim}{d}_i - \frac{1}{\tau_i} H_i \underset{\sim}{y}_i , \tag{2.15}$$

and

$$\tau_i = \frac{\underset{\sim}{y}_i^T H_i \underset{\sim}{y}_i}{\underset{\sim}{d}_i^T \underset{\sim}{y}_i} . \tag{2.16}$$

<u>Example</u>: The particular case $\rho_i = 1$, $\zeta_i = 0$, $i = 1,2,\ldots$ gives the variable metric or DFP formula which is the most frequently used member of the family.

The class of formulae described by (2.14) generate recursively a set of conjugate directions so that the first of our aims is satisfied. It still remains to show the relationship between the H_i and C^{-1}. To do

this note that (2.9) can be written (introducing the symmetric square root $C^{1/2}$ of the positive definite matrix C).

$$C^{1/2} H_i C^{1/2} C^{1/2} \underset{\sim}{t}_s = \rho_s C^{1/2} \underset{\sim}{t}_s \quad , \quad s = 1,2,\ldots,i-1 \quad ,$$

or, more briefly,

$$\hat{H}_i \hat{\underset{\sim}{t}}_s = \rho_s \hat{\underset{\sim}{t}}_s \quad , \quad s = 1,2,\ldots,i-1 \quad . \tag{2.9a}$$

Defining the matrix $\hat{T}$ by $\kappa_i(\hat{T}) = \dfrac{\hat{\underset{\sim}{t}}_s}{(\underset{\sim}{t}_s^T C \underset{\sim}{t}_s)^{1/2}}$, $i = 1,2,\ldots,n$, and the diagonal matrix $\mathbf{P}$ by $\mathbf{P}_{ii} = \rho_i$, $i = 1,2,\ldots,n$, we can write (2.9a) in the case $i = n+1$ in the form

$$\hat{H}_{n+1} \hat{T} = \hat{T} \mathbf{P} \quad . \tag{2.9b}$$

Now $\hat{T}$ is an orthogonal matrix so that

$$\hat{H}_{n+1} = \hat{T} \mathbf{P} \hat{T}^T \quad ,$$

whence

$$H_{n+1} = C^{-1/2} \hat{T} \mathbf{P} \hat{T}^T C^{-1/2} \quad . \tag{2.17}$$

In particular, if $\mathbf{P} = \rho I$,

$$H_{n+1} = \rho C^{-1} \quad . \tag{2.18}$$

<u>Remark</u>: Remember the motivation for developing the recursion (2.14) is the search for efficient descent directions. Specifically we are looking not only for conjugate directions but also for good estimates of the inverse Hessian. This indicates that $\rho = 1$ is the natural choice (or at least ρ = constant), and almost all published methods use $\rho = 1$. However, from (2.17), the choice of ρ variable may well have scaling advantages in the initial phases of a computation with a general objective function. Presumably the strategy for choosing ρ should make $\rho \to$ constant to ensure a fast rate of ultimate convergence.

Lemma 2.2: Provided the descent condition is satisfied, $H_{i+1}\underset{\sim}{g}_{i+1} \parallel \underset{\sim}{v}_i$ or null.

Remark: In what follows it is convenient to drop the i subscripts. Quantities subscripted $i+1$ will be starred. In what follows we assume ρ is constant.

Proof: We have (using the descent condition, the definition of $\underset{\sim}{t}$, and $\underset{\sim}{d} = \lambda \underset{\sim}{t}$)

$$D(\rho,H)\underset{\sim}{g}^* = (H + \rho \frac{\underset{\sim}{d}\underset{\sim}{d}^T}{\underset{\sim}{d}^T\underset{\sim}{y}} - \frac{H\underset{\sim}{y}\underset{\sim}{y}^T H}{\underset{\sim}{y}^T H \underset{\sim}{y}})\underset{\sim}{g}^*$$

$$= H\underset{\sim}{y} + H\underset{\sim}{g} - \frac{H\underset{\sim}{y}\underset{\sim}{y}^T H(\underset{\sim}{y} + \underset{\sim}{g})}{\underset{\sim}{y}^T H \underset{\sim}{y}}$$

$$= -\frac{1}{\lambda}(d - \frac{\underset{\sim}{y}^T\underset{\sim}{d}}{\underset{\sim}{y}^T H \underset{\sim}{y}} H\underset{\sim}{y}) \quad .$$

Whence

$$H^*\underset{\sim}{g}^* = -(\frac{1}{\lambda} + \zeta\tau\underset{\sim}{v}^T\underset{\sim}{g}^*)\underset{\sim}{v} \quad . \quad \square \tag{2.19}$$

Remark: (i) The condition that $H^*\underset{\sim}{g}^* = 0$ when $\underset{\sim}{v} \neq 0$ gives a condition which determines ζ . We have

$$\underset{\sim}{v}^T\underset{\sim}{g}^* = -\frac{1}{\tau}\underset{\sim}{g}^{*T} H\underset{\sim}{y} = -\frac{1}{\tau}\underset{\sim}{g}^{*T} H \underset{\sim}{g}^*$$

so that (from (2.19))

$$\zeta = \frac{1}{\lambda\underset{\sim}{g}^{*T} H \underset{\sim}{g}^*} \quad . \tag{2.20}$$

Provided this value of ζ is excluded from consideration then $\underset{\sim}{x}^*$ is independent of ζ. Note that this result is true for a general function as no properties specific to a quadratic form have been used in its derivation.

(ii) We can only have $\underset{\sim}{v} = 0$ with $\underset{\sim}{d}$ and $\underset{\sim}{g}^*$ nonnull if H is singular, and in this case H^* is also singular and the null space of H^* is at least as large as that of H. This follows from (2.15) which can vanish only if (a) $H\underset{\sim}{g}^* = 0$ and $1 + \frac{1}{\lambda\tau} = 0$, or (b) $H\underset{\sim}{g}$ and $H\underset{\sim}{g}^*$ are parallel. Now if H is singular $\exists\, \underset{\sim}{w}$, $\underset{\sim}{w}^T H = 0$. Thus $\underset{\sim}{w}^T \underset{\sim}{d} = 0$, and hence $\underset{\sim}{w}^T H^* = 0$.

Clearly it is important that H_i positive definite $\Rightarrow H_{i+1}$ positive definite, $i = 1,2,\ldots$ in order that premature termination should be avoided ($H^* \underset{\sim}{g}^* = 0$ and H^* positive definite $\Rightarrow \underset{\sim}{g}^* = 0$ whence $\underset{\sim}{x}^*$ is a stationary point). Conditions which ensure this are given in the following lemma (due to Powell).

<u>Lemma 2.3</u>: If $0 < \rho, \tau < \infty$, H positive semidefinite, and $HH^+ \underset{\sim}{v} = \underset{\sim}{v}$ (where H^+ is the generalized inverse of H), then H^* is positive semidefinite, and the null space of H^* is equal to that of H provided

$$\zeta > \frac{-\underset{\sim}{y}^T \underset{\sim}{d}}{(\underset{\sim}{d}^T H^+ \underset{\sim}{d})(\underset{\sim}{y}^T H \underset{\sim}{y}) - (\underset{\sim}{d}^T \underset{\sim}{y})^2} \quad . \tag{2.21}$$

<u>Proof</u>: We first note the identity

$$D(\rho,H) = (I + \underset{\sim}{u}\underset{\sim}{y}^T)H(I + \underset{\sim}{y}\underset{\sim}{u}^T) \tag{2.22}$$

where

$$\underset{\sim}{u} = \frac{1}{\sqrt{(\underset{\sim}{d}^T \underset{\sim}{y})(\underset{\sim}{y}^T H \underset{\sim}{y})}} \left\{ \sqrt{\rho}\, \underset{\sim}{d} - \frac{1}{\sqrt{\tau}} H \underset{\sim}{y} \right\} . \tag{2.23}$$

and

$$\det(I + \underset{\sim}{u}\underset{\sim}{y}^T) = 1 + \underset{\sim}{y}^T \underset{\sim}{u} = \sqrt{\frac{\rho}{\tau}} \tag{2.24}$$

so that, by the assumptions, $I + \underset{\sim}{u}\underset{\sim}{y}^T$ is nonsingular. Now $\underset{\sim}{y}^T \underset{\sim}{v} = 0$ so that H^* can be written

$$H^* = (I + \underset{\sim}{u}\underset{\sim}{y}^T)(H + \zeta\tau \underset{\sim}{v}\underset{\sim}{v}^T)(I + \underset{\sim}{y}\underset{\sim}{u}^T) . \tag{2.25}$$

Thus the problem reduces to considering $H + \zeta\tau \underset{\sim}{v}\underset{\sim}{v}^T$. We have

$$H + \zeta\tau \underset{\sim}{v}\underset{\sim}{v}^T = H(I + \zeta\tau H^+ \underset{\sim}{v}\underset{\sim}{v}^T) .$$

The null spaces of H and H^* will agree provided $I + \zeta\tau H^+ \underset{\sim}{v}\underset{\sim}{v}^T$ is nonsingular. The condition for singularity is

$$\begin{aligned} 0 &= \det(I + \zeta\tau H^+ \underset{\sim}{v}\underset{\sim}{v}^T) \\ &= 1 + \zeta\tau \underset{\sim}{v}^T H^+ \underset{\sim}{v} . \end{aligned}$$

Noting that $HH^+H = H$, and $HH^+\underset{\sim}{v} = \underset{\sim}{v} \Rightarrow HH^+\underset{\sim}{d} = \underset{\sim}{d}$ we have

$$\begin{aligned} 1 + \zeta\tau \underset{\sim}{v}^T H^+ \underset{\sim}{v} &= 1 + \zeta\tau \left(\underset{\sim}{d}^T H^+ \underset{\sim}{d} - 2\frac{\underset{\sim}{d}^T \underset{\sim}{y}}{\tau} + \frac{\underset{\sim}{y}^T H \underset{\sim}{y}}{\tau^2}\right) \\ &= 1 + \zeta\tau \left(\underset{\sim}{d}^T H^+ \underset{\sim}{d} - \frac{(\underset{\sim}{d}^T \underset{\sim}{y})^2}{\underset{\sim}{y}^T H \underset{\sim}{y}}\right) \end{aligned}$$

and this vanishes provided

$$\zeta = \frac{-\underset{\sim}{y}^T \underset{\sim}{d}}{(\underset{\sim}{d}^T H^+ \underset{\sim}{d})(\underset{\sim}{y}^T H \underset{\sim}{y}) - (\underset{\sim}{y}^T \underset{\sim}{d})^2} .$$

The stated result is a consequence of this and the observation that decreasing ζ below this value will make H^* indefinite. □

Remark: (i) The condition on τ is automatically satisfied if H is positive definite and the descent condition is satisfied for then $\underset{\sim}{d}^T \underset{\sim}{y} = -\underset{\sim}{g}^T \underset{\sim}{d} = \lambda \underset{\sim}{g}^T H \underset{\sim}{g}$. However the lemma does not require that the descent condition be satisfied and remains valid even though the exact minimum in the direction $\underset{\sim}{t}$ is not found. In this case the condition on τ is necessary.

Corolary 2.2: If H_1 positive definite, and $H_{i+1} = D(\rho, H_i)$, $i = 1,2,\ldots$ then provided the descent condition is satisfied for $i = 1,2,\ldots$ then H_{i+1} is positive definite.

Proof: This is a consequence of (2.22) and the above remark which shows that if H is positive definite, and if the descent condition is satisfied, then $I + \underset{\sim}{u}\underset{\sim}{y}^T$ is nonsingular. □

Theorem 2.1: (Dixon's equivalence theorem). If (i) the formula (2.14) is used to generate descent directions, (ii) ζ_i satisfies (2.21) for $i = 1,2,\ldots$ and H_1 is positive definite, and (iii) the descent condition is satisfied in each descent step, then the sequence of points generated by the algorithm depends only on F , H_1 , ρ , and $\underset{\sim}{x}_1$ and is independent of ζ_i , $i = 1,2,\ldots$.

Remark: It is important to note that F is not restricted to be a quadratic form in this result.

Proof: Let $D_1 = H_1$, $D_i = D(\rho, D_{i-1})$, $i = 2,3,\ldots$. We show that if $H_i = D_i + \alpha \underset{\sim}{d}_i \underset{\sim}{d}_i^T$, then $H_{i+1} = D_{i+1} + \beta \underset{\sim}{d}_{i+1} \underset{\sim}{d}_{i+1}^T$. By Lemma 2.2 we have $H^* = D(\rho,H) + \gamma \underset{\sim}{d}^* \underset{\sim}{d}^{*T}$. Now

$$D(\rho,H) = D + \rho \frac{\underset{\sim}{d}\underset{\sim}{d}^T}{\underset{\sim}{d}^T\underset{\sim}{y}} - \frac{(D+\alpha \underset{\sim}{d}\underset{\sim}{d}^T)\underset{\sim}{y}\underset{\sim}{y}^T(D+\alpha \underset{\sim}{d}\underset{\sim}{d}^T)}{\underset{\sim}{y}^T(D+\alpha \underset{\sim}{d}\underset{\sim}{d}^T)y} + \alpha \underset{\sim}{d}\underset{\sim}{d}^T$$

$$= D + \rho \frac{\underset{\sim}{d}\underset{\sim}{d}^T}{\underset{\sim}{d}^T\underset{\sim}{y}} - \frac{D\underset{\sim}{y}\underset{\sim}{y}^T D + \alpha(\underset{\sim}{y}^T\underset{\sim}{d})(D\underset{\sim}{y}\underset{\sim}{d}^T + \underset{\sim}{d}\underset{\sim}{y}^T D) + \alpha^2(\underset{\sim}{d}^T\underset{\sim}{y})^2 \underset{\sim}{d}\underset{\sim}{d}^T}{\underset{\sim}{y}^T D\underset{\sim}{y} + \alpha(\underset{\sim}{y}^T\underset{\sim}{d})^2} + \alpha \underset{\sim}{d}\underset{\sim}{d}^T$$

$$= D^* - \frac{-D\underset{\sim}{y}\underset{\sim}{y}^T D\alpha \dfrac{(\underset{\sim}{y}^T\underset{\sim}{d})^2}{\underset{\sim}{y}^T D\underset{\sim}{y}} + \alpha(\underset{\sim}{y}^T\underset{\sim}{d})(D\underset{\sim}{y}\underset{\sim}{d}^T + \underset{\sim}{d}\underset{\sim}{y}^T D) - \alpha(\underset{\sim}{y}^T D\underset{\sim}{y})\underset{\sim}{d}\underset{\sim}{d}^T}{\underset{\sim}{y}^T D\underset{\sim}{y} + \alpha(\underset{\sim}{y}^T\underset{\sim}{d})^2}$$

$$= D^* + \frac{\alpha(\underset{\sim}{y}^T D\underset{\sim}{y})}{\underset{\sim}{y}^T D\underset{\sim}{y} + \alpha(\underset{\sim}{y}^T\underset{\sim}{d})^2}\left(\underset{\sim}{d} - \frac{\underset{\sim}{y}^T\underset{\sim}{d}}{\underset{\sim}{y}^T D\underset{\sim}{y}} D\underset{\sim}{y}\right)\left(\underset{\sim}{d} - \frac{\underset{\sim}{y}^T\underset{\sim}{d}}{\underset{\sim}{y}^T D\underset{\sim}{y}} D\underset{\sim}{y}\right)^T . \qquad (2.26)$$

By Lemma 2.2, $\underset{\sim}{d}^* \parallel D(\rho,H)\underset{\sim}{g}^*$. By (2.26) $D(\rho,H)\underset{\sim}{g}^* \parallel D^*\underset{\sim}{g}^*$. Thus $\underset{\sim}{d}_j \parallel D_j\underset{\sim}{g}_j$, $j = 1,2,\ldots,i \Rightarrow \underset{\sim}{d}_{i+1} \parallel D_{i+1}\underset{\sim}{g}_{i+1}$. But the case $j = 1$ is a consequence of Lemma 2.2 so the result follows by induction. □

Example: Equivalence results for a wide class of conjugate direction algorithms applied to a given positive definite quadratic form can be demonstrated by noting that at the i-th stage we find the minimum in the translation to $\underset{\sim}{x}_1$ of the subspace spanned by $\underset{\sim}{t}_1,\ldots,\underset{\sim}{t}_i$, and that this subspace is also spanned by $H_1\underset{\sim}{g}_1,\ldots,H_1\underset{\sim}{g}_i$. Thus $\underset{\sim}{x}_{i+1}$ depends only on

$\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_i$ and not on the particular updating formula for the inverse Hessian estimate. If $H_1 = I$ this equivalence extends to the conjugate gradient algorithm (2.7).

<u>Lemma 2.4</u>: If the descent condition is satisfied at each stage then the sequence $\underset{\sim}{g}_i^T D_i \underset{\sim}{g}_i$, $i = 1,2,\ldots$ is strictly decreasing provided D_1 is positive definite.

<u>Proof</u>: We have (as $\underset{\sim}{g}^{*T} \underset{\sim}{d} = 0$)

$$\underset{\sim}{g}^{*T} D^* \underset{\sim}{g}^* = \underset{\sim}{g}^{*T} D \underset{\sim}{g}^* - \frac{(\underset{\sim}{g}^{*T} D \underset{\sim}{y})^2}{\underset{\sim}{y}^T D \underset{\sim}{y}} ,$$

$$= \underset{\sim}{g}^{*T} D \underset{\sim}{g}^* - \frac{(\underset{\sim}{g}^{*T} D \underset{\sim}{g}^*)^2}{\underset{\sim}{g}^{*T} D \underset{\sim}{g}^* + \underset{\sim}{g}^T D \underset{\sim}{g}} ,$$

$$= \frac{(\underset{\sim}{g}^{*T} D \underset{\sim}{g}^*)(\underset{\sim}{g}^T D \underset{\sim}{g})}{\underset{\sim}{g}^{*T} D \underset{\sim}{g}^* + \underset{\sim}{g}^T D \underset{\sim}{g}} .$$

Thus

$$\frac{1}{\underset{\sim}{g}^{*T} D^* \underset{\sim}{g}^*} = \frac{1}{\underset{\sim}{g}^{*T} D \underset{\sim}{g}^*} + \frac{1}{\underset{\sim}{g}^T D \underset{\sim}{g}} . \tag{2.27}$$

By Corollary 2.2, the D_i are positive definite so that the desired result follows from (2.27). □

<u>Remark</u>: This result indicates a potential defect of the DFP algorithm. For if the choice of D_1 is poor in the sense that it leads to too small a value of $\underset{\sim}{g}_1^T D_1 \underset{\sim}{g}_1$ then the algorithm has no mechanism to correct this, and must initially generate a sequence of directions which are nearly orthogonal to the gradient. This must also happen if, for any

reason, an abnormally small value of $\underset{\sim}{g}^T D \underset{\sim}{g}$ is generated at some stage. A possible cause of such behaviour is poor scaling of the problem.

<u>Lemma 2.5</u>: Corresponding to the formula (2.14) for updating H there is a similar formula for updating H^{-1}. Specifically we have

$$H^{*-1} = D(\rho,H)^{-1} + \gamma\mu \underset{\sim}{w}\underset{\sim}{w}^T \tag{2.28}$$

where

$$D(\rho,H)^{-1} = H^{-1} + \left(\frac{1}{\rho} + \mu\right) \frac{\underset{\sim}{y}\underset{\sim}{y}^T}{\underset{\sim}{d}^T\underset{\sim}{y}} - \frac{1}{\underset{\sim}{d}^T\underset{\sim}{y}} \left(\underset{\sim}{y}\underset{\sim}{d}^T H^{-1} + H^{-1}\underset{\sim}{d}\underset{\sim}{y}^T\right) , \tag{2.29}$$

$$\mu = \frac{\underset{\sim}{d}^T H^{-1} \underset{\sim}{d}}{\underset{\sim}{d}^T\underset{\sim}{y}} , \tag{2.30}$$

$$\underset{\sim}{w} = \underset{\sim}{y} - \frac{1}{\mu} H^{-1}\underset{\sim}{d} , \tag{2.31}$$

and γ is related to ζ by

$$\gamma = -\frac{\zeta\tau\mu}{1+\zeta\tau\underset{\sim}{v}^T H^{-1}\underset{\sim}{v}} . \tag{2.32}$$

<u>Proof</u>: From (2.22) we have

$$D(\rho,H)^{-1} = \left(I - \sqrt{\frac{\tau}{\rho}}\,\underset{\sim}{y}\underset{\sim}{u}^T\right) H^{-1}\left(I - \sqrt{\frac{\tau}{\rho}}\,\underset{\sim}{u}\underset{\sim}{y}^T\right)$$

and (2.29) follows from this by an elementary calculation. From (2.25)

$$H^{*-1} = \left(I - \sqrt{\frac{\tau}{\rho}}\,\underset{\sim}{y}\underset{\sim}{u}^T\right)\left(H + \zeta\tau\underset{\sim}{v}\underset{\sim}{v}^T\right)^{-1}\left(I - \sqrt{\frac{\tau}{\rho}}\,\underset{\sim}{u}\underset{\sim}{y}^T\right) ,$$

$$= \left(I - \sqrt{\frac{\tau}{\rho}}\,\underset{\sim}{y}\underset{\sim}{u}^T\right)\left(H^{-1} - \frac{\zeta\tau}{1+\zeta\tau\underset{\sim}{v}^T H^{-1}\underset{\sim}{v}} H^{-1}\underset{\sim}{v}\underset{\sim}{v}^T H^{-1}\right)\left(I - \sqrt{\frac{\tau}{\rho}}\,\underset{\sim}{u}\underset{\sim}{y}^T\right)\underset{\sim}{y} \tag{2.33}$$

Now

$$(I - \sqrt{\frac{\tau}{\rho}}\, \underset{\sim}{y}\, \underset{\sim}{u}^T)H^{-1}\underset{\sim}{v} = H^{-1}\underset{\sim}{d} - (\frac{1}{\tau} + \sqrt{\frac{\tau}{\rho}}\, \underset{\sim}{u}^T H^{-1}\underset{\sim}{d} - \frac{1}{\sqrt{\rho\tau}}\, \underset{\sim}{u}^T\, \underset{\sim}{y})\underset{\sim}{y}$$

$$= H^{-1}\underset{\sim}{d} - \mu\, \underset{\sim}{y}$$

$$= -\mu \underset{\sim}{w} \quad , \tag{2.34}$$

so that (2.28) is a direct consequence of (2.33) and (2.34). □

<u>Remark:</u> If we take $\gamma = -\dfrac{1}{\underset{\sim}{y}^T\underset{\sim}{d}}$ in (2.28) then we obtain

$$G(\rho,H^{-1}) = H^{-1} + \frac{1}{\rho}\frac{\underset{\sim}{y}\underset{\sim}{y}^T}{\underset{\sim}{y}^T\underset{\sim}{d}} - \frac{H^{-1}\underset{\sim}{d}\,\underset{\sim}{d}^T H^{-1}}{\underset{\sim}{d}^T H^{-1}\underset{\sim}{d}}$$

$$= (I + \underset{\sim}{z}\underset{\sim}{d}^T)H^{-1}(I + \underset{\sim}{d}\,\underset{\sim}{z}^T) \tag{2.35}$$

where

$$\underset{\sim}{z} = \frac{1}{\sqrt{(\underset{\sim}{d}^T\underset{\sim}{y})(\underset{\sim}{d}^T H^{-1}\underset{\sim}{d})}}\left\{\frac{1}{\sqrt{\rho}}\underset{\sim}{y} - \frac{1}{\sqrt{\mu}}H^{-1}\underset{\sim}{d}\right\} . \tag{2.36}$$

We have

$$D(\rho,H^{-1})^{-1} = G(\rho,H^{-1}) + \frac{\mu}{\underset{\sim}{y}^T\underset{\sim}{d}}\underset{\sim}{w}\underset{\sim}{w}^T$$

$$= (I + \underset{\sim}{z}\,\underset{\sim}{d}^T)(H^{-1} + \frac{\mu}{\underset{\sim}{y}^T\underset{\sim}{d}}\underset{\sim}{w}\,\underset{\sim}{w}^T)(I + \underset{\sim}{d}^T\underset{\sim}{z}) \tag{2.37}$$

as $\underset{\sim}{d}^T\underset{\sim}{w} = 0$.

To summarize these results we have the following:

(i) $$D(\rho,H) = (I + \underset{\sim}{u}\underset{\sim}{y}^T)H(I + \underset{\sim}{y}\underset{\sim}{u}^T) \quad ,$$

$$G(\rho,H^{-1}) = (I + \underset{\sim}{z}\underset{\sim}{d}^T)H^{-1}(I + \underset{\sim}{d}\underset{\sim}{z}^T) \quad .$$

(ii) $$D(\rho,H)^{-1} = G(\rho,H^{-1}) + \frac{\mu}{\underset{\sim}{y}^T\underset{\sim}{d}}\underset{\sim}{w}\underset{\sim}{w}^T \quad ,$$

$$G(\rho,H^{-1})^{-1} = D(\rho,H) + \frac{\tau}{\underset{\sim}{y}^T\underset{\sim}{d}}\underset{\sim}{v}\underset{\sim}{v}^T \quad .$$

	update formula	update formula for inverse
$D(\rho,H)$	$H + \rho\,\dfrac{\underset{\sim}{d}\underset{\sim}{d}^T}{\underset{\sim}{d}^T\underset{\sim}{y}} - \dfrac{H\underset{\sim}{y}\underset{\sim}{y}^T H}{\underset{\sim}{y}^T H\underset{\sim}{y}}$	$H^{-1} + (\frac{1}{\rho} + \mu)\dfrac{\underset{\sim}{y}\underset{\sim}{y}^T}{\underset{\sim}{d}^T\underset{\sim}{y}} - \dfrac{1}{\underset{\sim}{d}^T\underset{\sim}{y}}(\underset{\sim}{y}\underset{\sim}{d}^T H^{-1} + H^{-1}\underset{\sim}{d}\underset{\sim}{y}^T)$
$G(\rho,H^{-1})$	$H^{-1} + \dfrac{1}{\rho}\dfrac{\underset{\sim}{y}\underset{\sim}{y}^T}{\underset{\sim}{y}^T\underset{\sim}{d}} - \dfrac{H^{-1}\underset{\sim}{d}\underset{\sim}{d}^T H^{-1}}{\underset{\sim}{d}^T H^{-1}\underset{\sim}{d}}$	$H + (\rho + \tau)\dfrac{\underset{\sim}{d}\underset{\sim}{d}^T}{\underset{\sim}{d}^T\underset{\sim}{y}} - \dfrac{1}{\underset{\sim}{d}^T\underset{\sim}{y}}(\underset{\sim}{d}\underset{\sim}{y}^T H + H\underset{\sim}{y}\underset{\sim}{d}^T)$

$D(\rho,H)$, $G(\rho,H^{-1})$ have been called dual formulae by Fletcher.

<u>Lemma 2.6</u>: Let A be a symmetric matrix, $A = T\Lambda T^T$ where Λ diagonal ($\Lambda_{ii} = \lambda_i$, $i = 1,2,\ldots,n$) , and T orthogonal. Let λ_i^* , $i = 1,2,\ldots,n$ be the eigenvalues of $A + \sigma\underset{\sim}{a}\underset{\sim}{a}^T$, then either $\sigma > 0$ and $\lambda_i \le \lambda_i^* \le \lambda_{i+1}$, $i = 1,2,\ldots,n$, or $\sigma < 0$ and $\lambda_{i-1} \le \lambda_i^* \le \lambda_i$.

Proof: We have

$$\det\{A+\sigma \underset{\sim}{a}\underset{\sim}{a}^T - \lambda I\} = \prod_{i=1}^{n}(\lambda_i-\lambda)\ \det\{I+\sigma(\Lambda-\lambda I)^{-1}(T^T\underset{\sim}{a})(T^T\underset{\sim}{a})^T\}$$

$$= \prod_{i=1}^{n}(\lambda_i-\lambda)(1+\sigma(T^T\underset{\sim}{a})^T(\Lambda-\lambda I)^{-1}(T^T\underset{\sim}{a}))$$

$$= \prod_{i=1}^{n}(\lambda_i-\lambda)\{1+\sigma\sum_{i=1}^{n}\frac{\nu_i^2}{\lambda_i-\lambda}\}\ ,\quad \nu_i = \rho_i(T^T)\underset{\sim}{a}\ ,$$

and the desired result is an easy consequence of this expression. □

In the following theorem we consider specifically the minimization of a positive definite quadratic form. We assume that the initial estimate of the Hessian H_1 is positive definite, and we make use of the following sequences of updates for the current Hessian estimates

(a) $H_{i+1} = D(\rho,H_i)$, $i = 1,2,\ldots,$ and

(b) $\hat{H}_{i+1} = G(\rho,H_i^{-1})^{-1}$, $i = 1,2,\ldots$.

Further we do not assume that the descent condition is satisfied.

Theorem 2.2: (i) Let $K_i = C^{1/2}H_iC^{1/2}$, and let the eigenvalues of K_i ordered in increasing magnitude be $\lambda_j^{(i)}$, $j = 1,2,\ldots,n$. Then if $\lambda_j^{(1)} \geq \rho$ then $\lambda_j^{(1)} \geq \lambda_j^{(2)} \geq \cdots \geq \rho$, while if $\lambda_j^{(1)} \leq \rho$ then $\lambda_j^{(1)} \leq \lambda_j^{(2)} \leq \cdots \leq \rho$ for $j = 1,2,\ldots,n$. (ii) Let $\hat{K}_i = C^{1/2}\hat{H}_iC^{1/2}$, and let the eigenvalues of $\hat{K}_i$ be $\hat{\lambda}_j^{(i)}$, $j = 1,2,\ldots,n$. If $\hat{\lambda}_j^{(1)} \geq \rho$ then $\hat{\lambda}_j^{(1)} \geq \hat{\lambda}_j^{(2)} \geq \cdots \geq \rho$, while if $\hat{\lambda}_j^{(1)} \leq \rho$ then $\hat{\lambda}_j^{(1)} \leq \hat{\lambda}_j^{(2)} \leq \cdots \leq \rho$ for $j = 1,2,\ldots,n$.

Remark: This result is important because it shows that we have a 'weak' convergence result for these Hessian estimates when minimizing a positive definite quadratic form even when the descent conditions is not satisfied at each step.

Proof: Noting that $C^{1/2}\underset{\sim}{d} = C^{-1/2}\underset{\sim}{y} = \underset{\sim}{a}$, we can write the formula for updating K as

$$K^* = K + \rho \frac{\underset{\sim}{a}\underset{\sim}{a}^T}{\underset{\sim}{a}^T\underset{\sim}{a}} - \frac{K\underset{\sim}{a}\underset{\sim}{a}^T K}{\underset{\sim}{a}^T K \underset{\sim}{a}} \quad .$$

We can break this into the two operations

$$J = K - \frac{K\underset{\sim}{a}\underset{\sim}{a}^T K}{\underset{\sim}{a}^T K \underset{\sim}{a}} \quad , \quad \text{and}$$

$$K^* = J + \rho \frac{\underset{\sim}{a}\underset{\sim}{a}^T}{\underset{\sim}{a}^T\underset{\sim}{a}} \quad .$$

Note that J has a zero eigenvalue, and that $\underset{\sim}{a}$ is the corresponding eigenvector. By Lemma 2.6 we have $\lambda_1(J) = 0$, and $\lambda_{j-1} \leq \lambda_j(J) \leq \lambda_j$ for $j = 2,3,\ldots,n$. The rank one modification which takes J into K^* changes the zero eigenvalue to ρ and leaves the other eigenvalues of J unchanged. Assume that $\lambda_j(J) \leq \rho \leq \lambda_{j+1}(J)$ then reordering the eigenvalues in increasing order of magnitude we have $\lambda_k^* = \lambda_{k+1}(J)$, $k = 1,2,\ldots,j-1$, $\lambda_j^* = \rho$, $\lambda_k^* = \lambda_k(J)$, $k = j+1,\ldots,n$. This establishes the first part of the theorem. The second is demonstrated in similar fashion by noting that $\hat{K}^{-1}$ satisfies a formally similar update relation. This establishes the result for the eigenvalues of $\hat{K}^{-1}$ and hence for their reciprocals. □

Remark: Note that both H_i and $\hat{H}_i$ are positive definite $i = 2,3,\ldots$ if H_1 is positive definite. In this case the result does not depend on the descent conditions being satisfied.

Theorem 2.3: Let H be positive definite, and consider a step $\underset{\sim}{d}$ in the direction $-H\underset{\sim}{g}$. Let $H^* = D(\rho,H)$, $\hat{H} = G(\rho,H^{-1})^{-1} =$

$D(\rho,H) + \frac{\tau}{\underset{\sim}{y}^T\underset{\sim}{d}} \underset{\sim}{v}\underset{\sim}{v}^T$, and $H_\theta = \theta\hat{H} + (1-\theta)H^* = D(\rho,H) + \frac{\theta\tau}{\underset{\sim}{y}^T\underset{\sim}{d}} \underset{\sim}{v}\underset{\sim}{v}^T$. Let

$K = C^{1/2}HC^{1/2}$, and define K^*, $\hat{K}$, K_θ similarly. Let the eigenvalues of K, K^*, $\hat{K}$, K_θ be λ_j, λ_j^*, $\hat{\lambda}_j$, and λ_j^θ respectively, $j = 1,2,\ldots,n$. Let $0 \le \theta \le 1$. If $\lambda_j \ge \rho$ then $\lambda_j \ge \hat{\lambda}_j \ge \lambda_j^\theta \ge \lambda_j^* \ge \rho$ while if $\lambda_j \le \rho$ then $\lambda_j \le \lambda_j^* \le \lambda_j^\theta \le \hat{\lambda}_j \le \rho$. If $\theta \notin [0,1]$ then λ_j^θ need not lie in the interval defined by λ_j and ρ.

Proof: It follows from the definition of H^*, $\hat{H}$, and H_θ and Lemma 2.6 that $\lambda_j^* \le \lambda_j^\theta \le \hat{\lambda}_j$, $j = 1,2,\ldots,n$, provided $0 \le \theta \le 1$. The first part of the result is now a consequence of Theorem 2.2. To show that λ_j^θ need not lie in the interval defined by λ_j and ρ, consider the example

$$C = \begin{bmatrix} 1+\varepsilon & \sqrt{\varepsilon} \\ \sqrt{\varepsilon} & \varepsilon \end{bmatrix}, \quad H = I, \quad \rho = 1, \quad \underset{\sim}{a} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

We have $\lambda_1 = \eta$, $\lambda_2 = 1 + 2\varepsilon - \eta$ where $\eta = \frac{1}{2}(1 + 2\varepsilon - \sqrt{1+\mu\varepsilon})$. Thus η is positive and $O(\varepsilon^2)$. In this case we have

$$K = C, \quad \tau = \frac{\underset{\sim}{a}^T K \underset{\sim}{a}}{\underset{\sim}{a}^T \underset{\sim}{a}} = \varepsilon, \quad C^{1/2}\underset{\sim}{v} = \underset{\sim}{a} - \frac{1}{\varepsilon} K\underset{\sim}{a} = -\begin{bmatrix} 1/\sqrt{\varepsilon} \\ 0 \end{bmatrix}.$$

It is readily verified that $K^* = \begin{bmatrix} \varepsilon & 0 \\ 0 & 1 \end{bmatrix}$, so that $K_{-\varepsilon} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$, $K_{1+\varepsilon} = \begin{bmatrix} 1+2\varepsilon & 0 \\ 0 & 1 \end{bmatrix}$. In both cases eigenvalues lie outside the prescribed interval. In the first case we have $0 < \eta$, and in the second $1+2\varepsilon > 1+2\varepsilon-\eta$.

<u>Remark</u>: This result shows that $\hat{H}$ gives the best improvement in the eigenvalues $< \rho$, while H^* has a similar property for those $> \rho$. This suggests an algorithm in which a choice is made between updating H to $\hat{H}$ or H^* depending on some appropriate criterion. Fletcher suggests that if $\tau > 1$ (that is, $\underset{\sim}{y}^T H \underset{\sim}{y} > \underset{\sim}{y}^T C^{-1} \underset{\sim}{y}$) then H^* should be used, while if $\tau \leq 1$ then $\hat{H}$ is chosen. He has used this criterion in an implementation of Goldstein's algorithm, and has reported satisfactory results.

Notes

1. For Ostrowski's theorem see his book 'Solution of Equations' (2nd edition) or Kowalik and Osborne. Goldstein's theorem is from his paper 'On steepest descent' in SIAM Control, 1965. Theorem 1.3 is abstracted from Goldstein and Price, 'An Effective Algorithm for Minimization', Num. Math. 1967.

2. For background material see Kowalik and Osborne. The form of the update for the inverse Hessian is due to Powell 'Recent Advances in Unconstrained Optimization' to appear in Math. Prog. It is a specialization of a form derived in Huang, 'Unified approach to quadratically terminating algorithms for function minimization', JOTA, 1970. The form (2.14) and the result of Lemma 2.2 are probably due (in the case $\rho = 1$) to Fletcher 'A new approach to variable metric algorithms', Comp. J., 1970, and Broyden, 'Convergence of a class of double rank minimization algorithms', JIMA, 1970. Lemma 2.3 is due to Powell (to be published). The product update form (2.22) is due to Greenstadt (to be published). Dixon's paper containing Theorem 2.1 is to appear in Math. Prog. The significance of (2.27) for the successful performance of the DFP algorithm was noted in Powell's survey paper already cited. Attention was drawn to the dual updating formulae by Fletcher. This material together with Theorems 2.2 and 2.3 are included in his paper already cited.

References

C. G. Broyden (1970): The convergence of a class of double rank minimization algorithms, J. Inst. Math. Applic., 6, pp. 76-90.

R. Fletcher (1970): A new approach to variable metric algorithms, Computer J., 13, pp. 317-322.

A. A. Goldstein (1965): On steepest descent, SIAM J. Control, 3, pp. 147-151.

A. A. Goldstein and J. Price (1967): An effective algorithm for minimization, Num. Math., 10, pp. 184-189.

N. Y. Huang (1970): Unified approach to quadratically convergent algorithms for function minimization, J.O.T.A., 5, pp. 405-423.

J. Kowalik and M. R. Osborne (1968): Methods for Unconstrained Optimization Problems, Elsevier.

A. M. Ostrowski (1966): Solutions of Equations and Systems of Equations (second edition), Academic Press.

M. J. D. Powell (1970): Recent advances in unconstrained optimization, Report TP.430, AERE Harwell.

APPENDIX Numerical Questions Relating to Fletcher's Algorithm

1. Implementation

In this section we consider two questions relating to the implementation of Fletcher's algorithm. These are

(i) an appropriate strategy for determining λ to satisfy the Goldstein condition, and

(ii) the use of the product updating formulae for the inverse Hessian estimate.

In his program Fletcher uses a cubic line search to determine λ. Here we use a somewhat simpler procedure which has the advantage of requiring only additional function values. Also we work with the Choleski decomposition of the inverse Hessian estimate. This has certain numerical advantages which have been outlined by Gill and Murray*/. In particular, it is possible to ensure the positive definiteness of H_i, and this can be lost through the effect of accumulated rounding error when direct evaluation of the updating formulae is used. Another possible advantage of the Choleski decomposition is that we can work with an estimate of the Hessian (that is H^{-1}) rather than with H as division by a triangular matrix does not differ greatly in cost to multiplication. We felt this could well be an advantage in problems with singular or near singular Hessians, in which case H would be likely to contain large numbers.

To implement the line search we note that by Theorem 1.2 we should test first if $\psi(\underset{\sim}{x}_i, \underset{\sim}{t}_i, \|\underset{\sim}{s}_i\|) = \psi(\underset{\sim}{x}_i, \underset{\sim}{s}_i, 1)$ satisfies the Goldstein condition. This requires the evaluation of $F(\underset{\sim}{x}_i + \underset{\sim}{s}_i)$, and this, together with the

*/ NPL Mathematics Division, Report 97, 1970.

known values $F(\underset{\sim}{x}_i)$ and $F'(\underset{\sim}{x}_i) = \nabla F(\underset{\sim}{x}_i)\underset{\sim}{s}_i$, gives sufficient information to determine a quadratic interpolating polynomial to F . We write this as

$$P(\lambda) = F(\underset{\sim}{x}_i) + F'(\underset{\sim}{x}_i)\lambda + A\lambda^2 \tag{A.1}$$

where A is to be determined by setting $P(1) = F(\underset{\sim}{x}_i + \underset{\sim}{s}_i)$. This gives

$$\begin{aligned} A &= F(\underset{\sim}{x}_i + \underset{\sim}{s}_i) - F(\underset{\sim}{x}_i) - F'(\underset{\sim}{x}_i) \\ &= F'(\underset{\sim}{x}_i)(\psi(\underset{\sim}{x}_i, \underset{\sim}{s}_i, 1) - 1) \quad . \end{aligned} \tag{A.2}$$

The minimum of $P(\lambda)$ is given by

$$\lambda = -\frac{F'(\underset{\sim}{x}_i)}{2A} = \frac{1}{2(1 - \psi(\underset{\sim}{x}_i, \underset{\sim}{s}_i, 1))} \quad . \tag{A.3}$$

To test if this is an appropriate value we compute $\psi(\underset{\sim}{x}_i, \underset{\sim}{s}_i, \lambda)$. This gives

$$\psi(\underset{\sim}{x}_i, \underset{\sim}{s}_i, \lambda) = \frac{1}{2} + \frac{1}{2}\left\{1 - \frac{\frac{1}{2}F''(\underset{\sim}{x}_i + \bar{\lambda}\underset{\sim}{s}_i)}{A}\right\} \tag{A.4}$$

where $\bar{\lambda}$ is a mean value. Thus, if F is quadratic and $\psi(\underset{\sim}{x}_i, \underset{\sim}{s}_i, 1) < \sigma$ then λ given by equation (A.3) satisfies the Goldstein condition for any allowable σ (normally σ is chosen small -- say 10^{-4}): For nonquadratic F the test is satisfied if the relative error in estimating $\frac{1}{2}F''(\underset{\sim}{x}_i + \bar{\lambda}\underset{\sim}{s}_i)$ by A is not too large.

This analysis provides the basis for our method which is given below.

<u>Algorithm</u>

(i) Calculate $\|\underset{\sim}{s}_i\|$, set $w = \min(1, \|\underset{\sim}{s}_i\|)$, $\lambda = 1$.

(ii) Evaluate $\psi = \psi(\underset{\sim}{x}_i, \underset{\sim}{s}_i, \lambda)$.

(iii) If $\lambda < \sigma$ then begin $p\lambda = \lambda$,

$$\lambda = \frac{p\lambda}{2(1-\psi)} ,$$

go to (ii), end.

(iv) If $\lambda \leq 1-\sigma$ go to EXIT .

If $\lambda \geq 1$ then begin if $\lambda \geq 1/w$ go to EXIT,

$\lambda = 2\lambda$, end.

else $\lambda = .5(\lambda + p\lambda)$.

go to (ii).

<u>Remark</u>

(i) Numerical experience has shown that the value of λ predicted in (iii) can be too small, and that an additional instruction

If $\lambda < s^*p\lambda$ then $\lambda = s^*p\lambda$

should be included. A value for s of about .1 has proved satisfactory (1/8 was used in the numerical experiments reported in the next section).

(ii) It is readily verified that $\lim_{\lambda \to 0} \psi(\underset{\sim}{x}_i, \underset{\sim}{s}_i, \lambda) = 1$. Thus the algorithm can be expected to return a value of λ satisfying the Goldstein condition unless ψ exhibits rather pathological behavior.

We write the Choleski decomposition of H as

$$H = R^T R \tag{A.5}$$

where R is an upper triangular matrix. Thus we require to find R^* such that

$$R^{*T} R^* = H^* \tag{A.6}$$

where H^* is given by either

(i) $H^* = (I + \underset{\sim}{u}\underset{\sim}{y}^T)R^TR(I + \underset{\sim}{y}\underset{\sim}{u}^T)$, or

(ii) $H^* = (I + \underset{\sim}{u}\underset{\sim}{y}^T)\{R^TR + \zeta\tau\,\underset{\sim}{v}\underset{\sim}{v}^T\}(I + \underset{\sim}{y}\underset{\sim}{u}^T)$.

The second case can be reduced to the first if we write

$$\hat{R}^T\hat{R} = R^TR + \zeta\tau\,\underset{\sim}{v}\underset{\sim}{v}^T \quad . \tag{A.7}$$

To calculate $\hat{R}$ note that

$$R^TR + \zeta\tau\,\underset{\sim}{v}\underset{\sim}{v}^T = [R^T \mid \sqrt{\zeta\tau}\,\underset{\sim}{v}]\begin{bmatrix} R \\ \sqrt{\zeta\tau}\,\underset{\sim}{v}^T \end{bmatrix}$$

$$= [R^T \mid \sqrt{\zeta\tau}\,\underset{\sim}{v}]Q^TQ\begin{bmatrix} R \\ \sqrt{\zeta\tau}\,\underset{\sim}{v}^T \end{bmatrix} \tag{A.8}$$

where Q is orthogonal. Thus we seek an orthogonal matrix Q such that

$$Q\begin{bmatrix} R \\ \sqrt{\zeta\tau}\,\underset{\sim}{v}^T \end{bmatrix} = \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} \quad . \tag{A.9}$$

Let $W(i,j,\{p,q\})$ be the plane rotation such that $W(i,j,\{p,q\})A$ combines the i-th and j-th rows of A , and reduces A_{pq} to zero. It is necessary that p be either i or j . Then Q is given explicitly by

$$Q = \prod_{i=n}^{1} W(i,n+1,\{n+1,i\}) \quad . \tag{A.10}$$

It is readily verified that the zero introduced by each transformation is preserved by the subsequent transformations provided they are carried out in the order indicated.

Consider now the problem of constructing the Choleski decomposition of S^TS where $S = T + \underset{\sim}{a}\underset{\sim}{b}^T$, and T is upper triangular. This corresponds to our problem with the identifications $T = R$ or $\hat{R}$, $\underset{\sim}{a} = R\underset{\sim}{y}$ or $\hat{R}\underset{\sim}{y}$, and $\underset{\sim}{b} = \underset{\sim}{u}$. In this case the decomposition is done in two stages. Our method uses ideas due independently to Stoer, Golub, and Gill and Murray.

(i) We determine an orthogonal matrix Q_1 such that

$$Q_1\underset{\sim}{a} = \|\underset{\sim}{a}\|\underset{\sim}{e}_n \quad . \tag{A.11}$$

If we set

$$Q_1 = \prod_{i=1}^{n-1} W(i,n,\{i,*\}) \tag{A.12}$$

where the $*$ indicates that the rotation is defined by being applied to zero an element of a vector, then $Q_1S = Q_1T + \|\underset{\sim}{a}\|\underset{\sim}{e}_n\underset{\sim}{b}^T$ differs from an upper triangular matrix only in having possible nonzero elements in the last row.

(ii) To complete the determination of R^* we sweep out the elements in the first $(n-1)$ places in the last row of Q_1S by plane rotations. Thus R^* is given by

$$R^* = Q_2(Q_1T + \|\underset{\sim}{a}\|\underset{\sim}{e}_n\underset{\sim}{b}^T) \tag{A.13}$$

where

$$Q_2 = \prod_{i=n-1}^{1} W(i,n,\{n,i\}) \quad . \tag{A.14}$$

It will be seen that the updating of the Choleski factorization can be carried out very cheaply. Depending on the update formula used, the major cost is either $2n$ or $3n$ plane rotations. It should be noted that

$$\underset{\sim}{y}^T H \underset{\sim}{y} = \|R\underset{\sim}{y}\|^2 = \|\underset{\sim}{a}\|^2 \tag{A.15}$$

is required in the update formula. Thus $\underset{\sim}{a}$ can be already available for S.

2. Numerical Results

In this section we report the results of numerical experiments carried out to test some of our techniques. We consider four line search strategies:

(i) a standard cubic interpolation procedure with $\lambda = 1$ as initial search interval,

(ii) a standard cubic interpolation procedure with λ given by the step to the minimum in the previous line search,

(iii) a strategy for satisfying the Goldstein condition in which λ is reduced by the factor $1/8$ if $\psi < \sigma$, and

(iv) the method for satisfying the Goldstein condition given in the previous section.

Product form updating for the Choleski factorization of both H and $H^{-1} = G$ has been implemented, and the results obtained for each are given.

The problems considered include:

(i) Hilbert: Minimization of a quadratic form with matrix given by the Hilbert matrix of order 5. Here

$$F = \frac{1}{2} \sum_{i=1}^{5} \sum_{j=1}^{5} \frac{(x_i-1)(x_j-1)}{i+j-1} ,$$

and the starting point is given by

$$x_i = -4/i \quad , \quad i = 1,2,\ldots,5 .$$

(ii) Banana(n) : The Banana function in the cases $n = 2$ (the Rosenbrock function) and $n = 8$. Here

$$F = \sum_{i=1}^{n-1} \{100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2\} \quad ,$$

and the starting point is given by

$$x_i = -1{\cdot}2 \quad \text{if } i \text{ odd, otherwise } x_i = 1 \ .$$

(iii) Woods: Here

$$F = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 90(x_4 - x_3^2)^2$$
$$+ (1 - x_3)^2 + 10.1((1 - x_2)^2 + (1 - x_4)^2)$$
$$+ 19{\cdot}8(1 - x_2)(1 - x_4) \quad ,$$

and the starting point is

$$\underset{\sim}{x}^T = \{-3, -1, -3, -1\} \ .$$

(iv) Singular: Powell's singular function is designed to test the performance of algorithms on a function with a singular Hessian at the solution. Here

$$F = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4 \quad ,$$

and the starting point is

$$\underset{\sim}{x}^T = \{3, -1, 0, 1\} \ .$$

(v) Helix: Here we define

$$R = \{x_1^2 + x_2^2\}^{1/2} \quad ,$$

$$T = \text{if } x_1 \geq 0 \text{ then } \frac{1}{2\pi} \arctan\left(\frac{x_2}{x_1}\right)$$

$$\text{if } x_1 < 0 \text{ then } \frac{1}{2\pi} \arctan\left(\frac{x_2}{x_1}\right) + .5 \quad ,$$

and set

$$F = 100((x_3 - 10T)^2 + (R-1)^2) + x_3^2 \quad .$$

The starting vector is

$$\underset{\sim}{x}^T = \{-1, 0, 0\} \quad .$$

Numerical results are given in Table A.1. For historical reasons, the test for terminating the calculations was based on the size of $\|\underset{\sim}{s}_i\|$ ($\|\underset{\sim}{s}_i\| \leq EPS/n$ with $EPS = 10^{-8}$). This proved reasonably satisfactory for all cases except the singular function -- in fact in all other cases the ultimate convergence was clearly superlinear, and the results were accordingly only marginally affected by the size of EPS. In the case of singular the convergence test proved difficult to satisfy in most cases (indicated by * in Table A.1), and these computations were terminated by the number of iterations exceeding the specified limit. However, in all cases the answers were correct to at least six decimal places. There is some variation in the H and G columns. This shows the effect of rounding error, as these would be identical in exact arithmetic. The most interesting case is the H column in both cases of the Banana (8) when satisfying the Goldstein condition. In these cases both H and G formulae produce very similar results until the 10-th iteration at which point the H formulae produce much larger reductions in F than do the G . However, this progress is not maintained and at the 20-th iteration (in the case of the line search algorithm of Section 3) the H matrix becomes singular and the iteration is terminated. A restart procedure could have been used at this stage.

The numerical results indicate that the new algorithm is promising. In general, although more iterations are required, we make significantly

fewer function evaluations in comparison with the routine using a standard line search. As only one derivative evaluation is required in each iteration, the real saving can be considerable. We note that on the basis of the evidence presented it is not possible to draw conclusions as to the relative values of the H and G algorithms. However, that both manage to produce very comparable results provides some evidence of their stability.

The program which gave the results presented here is coded in ALGOL W for the IBM 360/67 at Stanford University. The calculations were carried out using long precision (14 hexadecimal digits). A FORTRAN version of the program has been developed at the Australian National University.

	cubic line search				Goldstein algorithm			
	λ^* from prev. tt		$\lambda^I = 1$		$(1/8)^P$		parabolic interpolation	
	H	G	H	G	H	G	H	G
Hilbert	6 / 19	6 / 19	6 / 22	6 / 24	40 / 40	40 / 40	39 / 41	39 / 41
Banana (2)	23 / 70	27 / 74	24 / 75	24 / 73	61 / 78	61 / 78	46 / 56	44 / 58
Banana (8)	74 / 178	75 / 174	74 / 235	72 / 219	NC	100 / 143	NC	80 / 114
Woods	27 / 80	27 / 78	30 / 102	29 / 100	94 / 114	80 / 112	62 / 84	61 / 86
Singular	50 / 157 *	50 / 153 *	32 / 129	60 / 205 *	100 / 111 *	100 / 111 *	69 / 83	76 / 89
Helix	NC	45 / 102	24 / 70	23 / 65	NC	NC	35 / 47	35 / 48

Table A.1 Number of iterations and number of function values in numerical experiments

III. Barrier and Penalty Function Methods

1. Basic properties of barrier functions.

Consider the inequality constrained problem (ICP)

$$\min f(\underset{\sim}{x})$$

$$\text{subject to } g_i(\underset{\sim}{x}) \geq 0 , \quad i = 1,2,\ldots,m \quad (i \in I_1) ,$$

where we assume (as before) that f, g_i, $i \in I_1$, are in C^2. We also assume that $S = \{\underset{\sim}{x} ; g_i(\underset{\sim}{x}) \geq 0 , i \in I_1\}$ is compact, has a nonvoid interior S_0, and satisfies the regularity condition that every neighborhood of points of S contains points of S_0 (this precludes S having 'whiskers'). If $\underset{\sim}{x} \in S$ and $g_i(\underset{\sim}{x}) = 0$ for some i then it is assumed that $\underset{\sim}{x} \notin S_0$.

Definition: $\phi(\underset{\sim}{g}(\underset{\sim}{x}))$ is a barrier function for S if the following conditions are satisfied.

(i) $\phi > 0$, $\underset{\sim}{x} \in S$. If X closed set, $X \subset S_0$, then $\phi \in C^2$ on X.

(ii) $\phi \to \infty$, $g_i \to 0$, $i \in I_1$.

(iii) $\frac{\partial \phi}{\partial g_i} < 0$ if $g_i < \rho_i$ where the ρ_i, $i \in I_1$, are fixed positive constants.

(iv) $\left|\frac{\partial \phi}{\partial g_i}\right|$ bounded on $N(\underset{\sim}{x}, \delta)$ if $g_i > 0$ on $N(\underset{\sim}{x}, \delta)$.

Example: (i) $\phi = \sum_{i=1}^{m} 1/g_i(\underset{\sim}{x})$ (inverse barrier function),

(ii) $\phi = \sum_{i=1}^{m} (\log(1 + g_i(\underset{\sim}{x})) - \log g_i(\underset{\sim}{x}))$.

Remark: In the second example the term with argument $1 + g_i(\underset{\sim}{x})$ merely ensures that the positivity condition is satisfied. It could be replaced by a bound k_i for $\log(1 + g_i(\underset{\sim}{x}))$ on S if this is known. In

practice it is of no consequence. The barrier function $\phi = \sum_{i=1}^{m} (k_i - \log g_i(\underset{\sim}{x}))$ is called the log barrier function.

<u>Definition</u>: $T(\underset{\sim}{x},r)$ is a <u>barrier objective function</u> if

$$T(\underset{\sim}{x},r) = f(\underset{\sim}{x}) + r\phi(\underset{\sim}{g}(\underset{\sim}{x})) \tag{1.1}$$

where $r > 0$.

<u>Lemma 1.1</u>: $\exists\ \underset{\sim}{x} = \underset{\sim}{x}(r) \in S_0$ such that $T(\underset{\sim}{x}(r),r) = \min_{\underset{\sim}{x} \in S} T(\underset{\sim}{x},r)$.

<u>Proof</u>: $T(\underset{\sim}{x},r)$ is bounded below on S , and $T(\underset{\sim}{x},r) \to +\infty$ as $\underset{\sim}{x} \to \partial S$. □

<u>Lemma 1.2</u>: Let $\{r_j\} \downarrow 0$, and let $\underset{\sim}{x}(r_j) = \underset{\sim}{x}_j$. Then

(i) $\{T(\underset{\sim}{x}_j,r_j)\}$ is strictly decreasing,

(ii) $\{f(\underset{\sim}{x}_j)\}$ is nonincreasing, and

(iii) $\{\phi(x_j)\}$ is nondecreasing.

<u>Proof</u>: Let $r_i < r_j$ then

$$\begin{aligned} f(\underset{\sim}{x}_i) + r_i\phi(\underset{\sim}{g}(\underset{\sim}{x}_i)) &\le f(\underset{\sim}{x}_j) + r_i\phi(\underset{\sim}{g}(\underset{\sim}{x}_j)) , \\ &< f(\underset{\sim}{x}_j) + r_j\phi(\underset{\sim}{g}(\underset{\sim}{x}_j)) , \\ &\le f(\underset{\sim}{x}_i) + r_j\phi(\underset{\sim}{g}(\underset{\sim}{x}_i)) . \end{aligned}$$

This demonstrates (i). Subtracting the inside and outside inequalities gives

$$(r_j - r_i)\phi(\underset{\sim}{g}(\underset{\sim}{x}_i)) \ge (r_j - r_i)\phi(\underset{\sim}{g}(\underset{\sim}{x}_j))$$

which gives (ii). From the first inequality we have

$$f(\underset{\sim}{x}_i) \leq f(\underset{\sim}{x}_j) + r_i(\phi(\underset{\sim}{g}(\underset{\sim}{x}_j)) - \phi(\underset{\sim}{g}(\underset{\sim}{x}_i)))$$

$$\leq f(\underset{\sim}{x}_j) \quad . \quad \square$$

<u>Remark</u>: If $T(\underset{\sim}{x},r)$ is strictly convex, then all inequalities are strict.

<u>Theorem 1.1</u>: The sequence $\{T_1(\underset{\sim}{x}_i,r_i)\}$ converges, and

$$\lim_{i\to\infty} T(\underset{\sim}{x}_i,r_i) = \min_{\underset{\sim}{x}\in S} f(\underset{\sim}{x}) \ .$$

<u>Proof</u>: By Lemma 1.2, $\{T(\underset{\sim}{x}_i,r_i)\}$ is decreasing and bounded below and hence convergent. Let $f^* = \min_{\underset{\sim}{x}\in S} f(\underset{\sim}{x})$, then

$$T(\underset{\sim}{x},r_i) \geq f(\underset{\sim}{x}) \geq f^*$$

whence

$$\lim_{i\to\infty} T(\underset{\sim}{x}_i,r_i) \geq f^* \quad . \tag{1.2}$$

Now let $\varepsilon > 0$ be given. Choose $\bar{\underset{\sim}{x}}\in S_0$ such that $f(\bar{\underset{\sim}{x}}) - f^* < \varepsilon/2$ (this is possible because of the regularity condition on S), and choose r_i such that $r_i\phi(\underset{\sim}{g}(\bar{\underset{\sim}{x}})) < \varepsilon/2$. Then

$$\min_{\underset{\sim}{x}} T(\underset{\sim}{x},r_i) \leq T(\bar{\underset{\sim}{x}},r_i) < f^* + \varepsilon$$

whence

$$\lim_{i\to\infty} T(\underset{\sim}{x}_i,r_i) \leq f^* \quad . \ \square \tag{1.3}$$

<u>Corollary 1.1</u>: The limit points of $\{\underset{\sim}{x}_i\}$ are local minima of the ICP.

Remark: The generality of these results should be noted. For example, we have not required S to be Lagrange regular at the limits points of $\{\underset{\sim}{x}_i\}$.

Definition: $Q(\underset{\sim}{x},\underset{\sim}{r})$ is a separable barrier objective function if

$$Q(\underset{\sim}{x},\underset{\sim}{r}) = f(\underset{\sim}{x}) + \sum_{i=1}^{m} r_i \phi_i(g_i(\underset{\sim}{x})) = f(\underset{\sim}{x}) + \underset{\sim}{r}^T \underset{\sim}{\phi}(\underset{\sim}{x}) \tag{1.4}$$

where $\underset{\sim}{r} > 0$, and ϕ_i is a barrier function for $S_i = \{\underset{\sim}{x} ; \underset{\sim}{g}_i(\underset{\sim}{x}) \geq 0\}$, $i = 1,2,\ldots,m$.

The previous results are readily extended to this case and are summarized in the following theorem.

Theorem 1.2: Let $\underset{\sim}{r}_i > \underset{\sim}{r}_{i+1}$, $i = 1,2,\ldots$, and $\lim_{i\to\infty} \underset{\sim}{r}_i = \underset{\sim}{\theta}$. Then

(i) $\min_{\underset{\sim}{x}\in S} Q(\underset{\sim}{x},\underset{\sim}{r}_k)$ is attained for some $\underset{\sim}{x}_k \in S_0$,

(ii) $\{Q(\underset{\sim}{x}_k,\underset{\sim}{r}_k)\}$ is strictly decreasing, $\{f(\underset{\sim}{x}_k)\}$ is nonincreasing, and

(iii) $\lim_{k\to\infty} Q(\underset{\sim}{x}_k,\underset{\sim}{r}_k) = f^*$, and the limit points of $\{\underset{\sim}{x}_k\}$ are local minima of the ICP.

Remark: Given a sequence of positive vectors tending to zero then it is possible to select a subsequence which is strictly decreasing. Conclusions (i) and (iii) remain valid in this more general case.

2. Multiplier relations (first order analysis)

In this section we assume sequences $\{r_k\} \downarrow 0$, $\{\underset{\sim}{x}_k\} \to \underset{\sim}{x}^*$. The condition that $T(\underset{\sim}{x}, r_k)$ is stationary at r_k gives

$$\nabla T(\underset{\sim}{x}_k, r_k) = \nabla f(\underset{\sim}{x}_k) + \sum_{i=1}^{m} r_k \frac{\partial \phi}{\partial g_i} \nabla g_i(\underset{\sim}{x}_k)$$

$$= \nabla f(\underset{\sim}{x}_k) - \sum_{i=1}^{m} u_i^k \nabla g_i(\underset{\sim}{x}_k) = 0 \qquad (2.1)$$

where $u_i^k = - r_k \frac{\partial \phi}{\partial g_i} (\underset{\sim}{g}(\underset{\sim}{x}_k))$. Note that $u_i^k = O(r_k) \to 0$, $k \to \infty$, if $i \notin B_0$, and $u_i^k \geq 0$ for $i \in B_0$ and $k \geq k_0$ by the conditions defining barrier functions. Equation (2.1) is formally similar to the multiplier relations given earlier (MP(3.2)), and it is comparatively straightforward to deduce these relations from (2.1) in certain special cases. We assume that $B_0 = \{1,2,\ldots,t\}$, that the rank of the system of vectors $\nabla g_i(\underset{\sim}{x}^*)$, $i \in B_0$ is $s \leq t$, and that $\nabla g_1(\underset{\sim}{x}^*), \ldots, \nabla g_s(\underset{\sim}{x}^*)$ are linearly independent. We define matrices $C_1(\underset{\sim}{x})$, $C_2(\underset{\sim}{x})$ by $\kappa_i(C_1) = \nabla g_i(\underset{\sim}{x})^T$, $i = 1, ,\ldots,s$, and $\kappa_i(C_2) = \nabla g_{s+i}(\underset{\sim}{x})^T$, $i = 1,2,\ldots,t-s$, and vectors $\underset{\sim}{u}_1^{(k)T} = \{u_1^k, \ldots, u_s^k\}$, $\underset{\sim}{u}_2^{(k)T} = \{u_{s+1}^k \ldots u_s^k\}$.

<u>Lemma 2.1</u>: If $\{\underset{\sim}{u}_2^k\}$ is bounded then the Kuhn-Tucker conditions hold at $\underset{\sim}{x}^*$.

<u>Proof</u>: From (2.1) we have

$$\nabla f(\underset{\sim}{x}_k)^T = C_1(\underset{\sim}{x}_k)\underset{\sim}{u}_1^{(k)} + C_2(\underset{\sim}{x}_k)\underset{\sim}{u}_2^{(k)} + O(r_k) \quad . \qquad (2.2)$$

The linear dependence of the set of vectors $\nabla g_i(\underset{\sim}{x}^*)$, $i \in B_0$, gives

$$C_2(\underset{\sim}{x}^*) = C_1(\underset{\sim}{x}^*)R \tag{2.3}$$

so that (2.2) can be written

$$\nabla f(\underset{\sim}{x}_k)^T = C_1(\underset{\sim}{x}_k)\{\underset{\sim}{u}_1^{(k)} + R\underset{\sim}{u}_2^{(k)}\} + \{C_2(\underset{\sim}{x}_k) - C_1(\underset{\sim}{x}_k)R\}\underset{\sim}{u}_2^{(k)} + O(r_k) \ . \tag{2.4}$$

Provided k is large enough the rank of $C_1(\underset{\sim}{x}_k)$ will be s (see MP Remark following Lemma 3.3). Thus

$$\underset{\sim}{u}_1^{(k)} + R\underset{\sim}{u}_2^{(k)} = \{C_1(\underset{\sim}{x}_k)^T C_1(\underset{\sim}{x}_k)\}^{-1} C_1(\underset{\sim}{x}_k)^T\{\nabla f(\underset{\sim}{x}_k)^T + (C_2(\underset{\sim}{x}_k) - C_1(\underset{\sim}{x}_k)R)\underset{\sim}{u}_2^{(k)} + O(r_k)\} \ . \tag{2.5}$$

As $\underset{\sim}{u}_2^{(k)}$ bounded we conclude from (2.5)

(i) $\underset{\sim}{u}_1^{(k)}$ bounded, and

(ii) $\lim_{k\to\infty} \underset{\sim}{u}_1^{(k)} + R\underset{\sim}{u}_2^{(k)} = \{C_1(\underset{\sim}{x}^*)^T C_1(\underset{\sim}{x}^*)\}^{-1} C_1(\underset{\sim}{x}^*)^T \nabla f(\underset{\sim}{x}^*)$.

As $\{\underset{\sim}{u}_1^{(k)}\}$, $\{\underset{\sim}{u}_2^{(k)}\}$ are bounded and nonnegative (at least for k large enough) this property is shared by the limit points of the sequences. Consider subsequences tending to $\underset{\sim}{u}_1^*$, $\underset{\sim}{u}_2^*$ respectively. From (2.2) we have

$$\nabla f(\underset{\sim}{x}^*)^T = C_1(\underset{\sim}{x}^*)\underset{\sim}{u}_1^* + C_2(\underset{\sim}{x}^*)\underset{\sim}{u}_2^*$$

or

$$\nabla f(\underset{\sim}{x}^*) = \sum_{i\in B_0} u_i^* \nabla g_i(\underset{\sim}{x}^*) \ . \tag{2.6}$$

Thus the Kuhn-Tucker conditions are satisfied. □

<u>Corollary 2.1</u>: If the Kuhn-Tucker conditions do not hold at $\underset{\sim}{x}^*$, then $\{\underset{\sim}{u}_1^k\}$, $\{\underset{\sim}{u}_2^k\}$ are unbounded.

Corollary 2.2: If restraint condition A holds then the $\nabla g_i(\underset{\sim}{x}^*)$ are linearly independent for $i \epsilon B_0$. In this case $\{\underset{\sim}{u}_2^k\}$ is null, and $\{\underset{\sim}{u}_1^k\}$ converges. If restraint condition A holds then the multipliers in the Kuhn-Tucker conditions are uniquely determined.

Lemma 2.2: If restraint condition B holds then $\{\underset{\sim}{u}_2^{(k)}\}$ is bounded.

Remark: By MP Lemma 5.2, this implies that $\{\underset{\sim}{u}_2^{(k)}\}$ is bounded for the convex programming problem provided S has an interior.

Proof: If restraint condition B holds then $\exists\ \underset{\sim}{d}$ such that $\nabla g_i(\underset{\sim}{x}^*)\underset{\sim}{d} > 0$, $i = 1,2,\ldots,t$. From (2.2) we have

$$\sum_{i=1}^{t} u_i^k \nabla g_i(\underset{\sim}{x}_k)\underset{\sim}{d} = \nabla f(\underset{\sim}{x}_k)\underset{\sim}{d} + O(r_k) \quad . \tag{2.7}$$

As $\nabla g_i(\underset{\sim}{x})\underset{\sim}{d}$ is a continuous function, we must have $u_i^k > 0$ and $\nabla g_i(\underset{\sim}{x}_k)\underset{\sim}{d} > 0$, $i = 1,2,\ldots,t$, provided k large enough. Thus (2.7) gives

$$\sum_{i=1}^{t} u_i^k \leq \frac{\nabla f(\underset{\sim}{x}_k)\underset{\sim}{d} + O(r_k)}{\min_i \nabla g_i(\underset{\sim}{x}_k)\underset{\sim}{d}} \quad . \tag{2.8}$$

This relation shows that the u_i^k are bounded as $k \to \infty$. □

Remark: The results of the first section showed that convergence of barrier function algorithms can be proved under very few assumptions. The results of this section show that valuable structural information on the problem is available as a by-product of the computation. Note that the condition that the u_i^k be bounded is a weaker restraint condition than either A or B.

3. Second order conditions.

Consider now a barrier function ϕ and a sequence $\{r_k\} \downarrow 0$ such that $\{\underset{\sim}{x}_k\} \to \underset{\sim}{x}^*$, $\{\underset{\sim}{u}_k\} \to \underset{\sim}{u}^*$. It is convenient to assume the following properties which are satisfied by all barrier functions of practical interest.

(i) $u_i^k = - r_k \dfrac{\partial \phi}{\partial g_i}(\underset{\sim}{x}_k) > 0$, $r_k \dfrac{\partial^2 \phi}{\partial g_i^2}(\underset{\sim}{x}_k) > 0$, $i = 1,2,\ldots,m$, $\forall k$.

(ii) $r_k \dfrac{\partial^2 \phi}{\partial g_i^2}(\underset{\sim}{x}_k) \to +\infty$, $k \to \infty$, $i \in B_0$. (But see Example 4(ii) p. 100 for qualification.)

(iii) $\dfrac{\partial^2 \phi}{\partial g_i \partial g_j} = 0$ if $i \neq j$.

<u>Lemma 3.1</u>: If the matrices $\nabla^2 T(\underset{\sim}{x}_k, r_k)$ are positive definite for $k \geq k_0$ and the $\nabla g_i(\underset{\sim}{x}^*)$, $i \in B_0$, are linearly independent then $\underset{\sim}{v}^T \nabla^2 f(\underset{\sim}{x}^*, \underset{\sim}{u}^*) \underset{\sim}{v} > 0$, $\forall \underset{\sim}{v} \neq 0$ such that $\nabla g_i(\underset{\sim}{x}^*) \underset{\sim}{v} = 0$, $\forall i \in B_0$.

<u>Proof</u>: Differentiating $T(\underset{\sim}{x}, r_k)$ gives

$$\nabla^2 T(\underset{\sim}{x}_k, r_k) = \nabla^2 f(\underset{\sim}{x}_k, \underset{\sim}{u}_k) + r_k \sum_{i=1}^{m} \frac{\partial^2 \phi}{\partial g_i^2} \nabla g_i(\underset{\sim}{x}_k)^T \nabla g_i(\underset{\sim}{x}_k) \tag{3.1}$$

which can be written

$$\nabla^2 T(\underset{\sim}{x}_k, r_k) = \nabla^2 f(\underset{\sim}{x}_k, \underset{\sim}{u}_k) + C_1(\underset{\sim}{x}_k) D_k C_1(\underset{\sim}{x}_k)^T + \sum_{i \in I_1 - B_0} r_k \frac{\partial^2 \phi}{\partial g_i^2} \nabla g_i(\underset{\sim}{x}_k)^T \nabla g_i(\underset{\sim}{x}_k)$$

where

$$(D_k)_{ii} = r_k \frac{\partial^2 \phi}{\partial g_i^2}(\underset{\sim}{x}_k) \quad , \quad i = 1,2,\ldots,t \ .$$

Let

$$P_k = C_1(\underset{\sim}{x}_k)C_1(\underset{\sim}{x}_k)^+ \tag{3.2}$$

then, for arbitrary nonzero $\underset{\sim}{v}$ such that $(I - P_k)\underset{\sim}{v} \neq \underset{\sim}{\theta}$

$$0 < \underset{\sim}{v}^T(I - P_k)\nabla^2 T(\underset{\sim}{x}_k, \underset{\sim}{r}_k)(I - P_k)\underset{\sim}{v}$$

$$= \underset{\sim}{v}^T(I - P_k)\nabla^2 \mathcal{L}(\underset{\sim}{x}_k, \underset{\sim}{u}_k)(I - P_k)\underset{\sim}{v} + o(1)$$

as

$$(I - P_k)C_1(\underset{\sim}{x}_k) = 0 \quad .$$

The desired result follows from this on letting $k \to \infty$. □

<u>Corollary 3.1</u>: If in addition to the conditions of Lemma 3.1 we have also strict complementarity then the second order sufficiency conditions (the conditions of MP Lemma 4.3) hold at $\underset{\sim}{x}^*$.

<u>Remark</u>: The problem of generalizing this result to the case where the active constraint gradients are not linearly independent is the following. In general, when $k < \infty$, $\operatorname{rank}[C_1(\underset{\sim}{x}_k) | C_2(\underset{\sim}{x}_k)] > s$. Thus

$$V_k = \{\underset{\sim}{v} ; \nabla g_i(\underset{\sim}{x}_k)\underset{\sim}{v} = 0 , \forall i \in B_0\} \subset U_k = \{\underset{\sim}{v} ; \underset{\sim}{v} = (I - P_k)\underset{\sim}{u} , \underset{\sim}{u} \in E_n\} .$$

We have

$$\lim_{k \to \infty} U_k = V^* = \{\underset{\sim}{v} ; \nabla g_i(\underset{\sim}{x}^*)\underset{\sim}{v} = 0 , \forall i \in B_0\} \quad .$$

It is not difficult to construct examples in which $\lim_{k \to \infty} V_k \subset V^*$.

Consider $C_1(\underset{\sim}{x}_k) = \underset{\sim}{e}_1$, $C_2(\underset{\sim}{x}_k) = \underset{\sim}{e}_1 + \underset{\sim}{e}_2^T(\underset{\sim}{x} - \underset{\sim}{x}^*)\underset{\sim}{e}_2$. Then

$V_k = \{\underset{\sim}{v} ; \underset{\sim}{e}_1^T \underset{\sim}{v} = \underset{\sim}{e}_2^T \underset{\sim}{v} = 0\} = \lim_{k\to\infty} V_k \subset V^* = \{\underset{\sim}{v} ; \underset{\sim}{e}_1^T \underset{\sim}{v} = 0\}$. The argument of Lemma 3.1 shows only that $\underset{\sim}{v}^T \nabla^2 \mathcal{L}(\underset{\sim}{x}^*, \underset{\sim}{u}^*) \underset{\sim}{v} > 0$ for $\underset{\sim}{v} \in \lim_{k\to\infty} V_k$.

<u>Lemma 3.2</u>: Let

$$W = U + \gamma V V^T ,$$

$$N = \{\underset{\sim}{t} ; \|\underset{\sim}{t}\| = 1 , V^T \underset{\sim}{t} = 0\} ,$$

$$M = \{\underset{\sim}{u} ; \|\underset{\sim}{u}\| = 1 , \underset{\sim}{u} \in N^\perp\} ,$$

$$\nu = \min_{\underset{\sim}{t}\in N} \underset{\sim}{t}^T U \underset{\sim}{t} > 0 , \quad \sigma = \min_{\underset{\sim}{u}\in M} \underset{\sim}{u}^T U \underset{\sim}{u} , \quad \mu = \min_{\underset{\sim}{u}\in M} \|V\underset{\sim}{u}\| > 0 ,$$

$$\eta = \min_{\underset{\sim}{t}\in N, \underset{\sim}{u}\in M} \underset{\sim}{t}^T U \underset{\sim}{u} , \quad \rho = \min(0, \eta)$$

then W is positive definite provided

$$\gamma > \frac{\rho^2 + \nu^2 - \sigma\nu}{\nu\mu^2} . \tag{3.3}$$

<u>Proof</u>: Any unit vector $\underset{\sim}{w}$ can be written

$$w = \alpha \underset{\sim}{u} + \beta \underset{\sim}{t} \quad \text{where } \underset{\sim}{u}\in M , \ \underset{\sim}{t}\in N , \text{ and } \alpha^2 + \beta^2 = 1 .$$

Thus

$$\begin{aligned} \underset{\sim}{w}^T W \underset{\sim}{w} &= \alpha^2 \underset{\sim}{u}^T U \underset{\sim}{u} + 2\alpha\beta \underset{\sim}{u}^T U \underset{\sim}{t} + \beta^2 \underset{\sim}{t}^T U \underset{\sim}{t} + \gamma\alpha^2 \|V^T \underset{\sim}{u}\|^2 \\ &\geq \alpha^2(\sigma + \gamma\mu^2 - \nu) + 2|\alpha|(1-\alpha^2)^{1/2}\rho + \nu \\ &\geq \alpha^2(\sigma + \gamma\mu^2 - \nu) + 2|\alpha|\rho + \nu \end{aligned} \tag{3.4}$$

as $\rho \leq 0$. Provided $\gamma > \frac{\nu-\sigma}{\mu^2}$ (a weaker condition than (3.3)), the right hand side has a minimum at $\alpha = \frac{-\rho}{\sigma + \gamma\mu^2 - \nu} > 0$. The value at this minimum is $\frac{-\rho^2}{\sigma + \gamma\mu^2 - \nu} + \nu$ which is positive if (3.3) holds. □

<u>Corollary 3.2</u>: If $W = U + VDV^T$ where D diagonal, and if the conditions of Lemma 3.2 hold, then W is positive definite provided $\min_i D_{ii} > \frac{\rho^2 + \nu^2 - \sigma\nu}{\nu\mu^2}$. If D is positive definite the result holds provided the smallest eigenvalue of D satisfies this inequality.

<u>Proof</u>: We have

$$\underset{\sim}{w}^T VDV^T \underset{\sim}{w} = \alpha^2 \underset{\sim}{u}^T VDV^T \underset{\sim}{u} = \alpha^2 \left\{ \sum_{i=1}^{t} D_{ii} (\rho_i (V^T) \underset{\sim}{u})^2 \right\}$$

$$\geq \min_i D_{ii} \; \alpha^2 \|V^T \underset{\sim}{u}\|^2 \quad .$$

The result now follows as from equation (3.4) above. □

<u>Lemma 3.3</u>: If the second order sufficiency conditions hold at $\underset{\sim}{x}^*$ then $\nabla^2 T(\underset{\sim}{x}_k, r_k)$ is positive definite for k large enough.

<u>Proof</u>: We have

$$V^* \subseteq U^* = \{\underset{\sim}{v} \,;\, \nabla g_i(\underset{\sim}{x}^*)\underset{\sim}{v} = 0 \;,\; \forall i \in B_0 \text{ such that } u_i^* > 0\} \quad .$$

Thus the second order sufficiency conditions imply that $\underset{\sim}{v}^T \nabla^2 f(\underset{\sim}{x}^*, \underset{\sim}{u}^*) \underset{\sim}{v} > 0$, $\forall \underset{\sim}{v} \in V^*$ such that $\underset{\sim}{v} \neq 0$. From Corollary 3.2 it follows that $\exists\, D_0$ such that $\nabla^2 f(\underset{\sim}{x}^*, \underset{\sim}{u}^*) + [C_1(\underset{\sim}{x}^*) | C_2(\underset{\sim}{x}^*)] D [C_1(\underset{\sim}{x}^*) | C_2(\underset{\sim}{x}^*)]^T$

is positive definite for $D \geq D_0$. By continuity this implies that the corresponding matrix evaluated at $\underset{\sim}{x}_k$ is positive definite for k large enough. The desired result follows from this as $(D_k)_{ii} \to \infty$, $i = 1,2,\ldots,t$ as $k \to \infty$. □

<u>Lemma 3.4</u>: If U is nonsingular, D diagonal, and V of full rank, then the system of linear equations

$$[U + VDV^T]\underset{\sim}{x} = V\underset{\sim}{y} \tag{3.5}$$

has the solution

$$\underset{\sim}{x} = U^{-1}V(I+M)^{-1}M\underset{\sim}{y} \tag{3.6}$$

where $M = (V^T U^{-1} V)^{-1} D^{-1}$ provided $I+M$ is nonsingular. A sufficient condition for $I+M$ nonsingular is $\|M\| < 1$ which is satisfied if $\min_i |D_{ii}|$ is large enough.

<u>Proof</u>: The result follows on substituting (3.6) into (3.5). □

<u>Remark</u>: From (3.6) it follows that

$$\underset{\sim}{x} \sim U^{-1}V(V^T U^{-1} V)^{-1} D^{-1}\underset{\sim}{y} \tag{3.7}$$

as $\min_i |D_{ii}| \to \infty$.

<u>Corollary 3.3</u>: If the right hand side of equation (3.5) is $\underset{\sim}{z}$, a general vector, then the solution is given by

$$\underset{\sim}{x} = U^{-1}V(I+M)^{-1}M(V^T U^{-1} V)^{-1} V^T U^{-1}\underset{\sim}{z}$$
$$+ U^{-1}(I - V(V^T U^{-1} V)^{-1} V^T U^{-1})\underset{\sim}{z} \quad . \tag{3.8}$$

4. Rate of convergence results.

In this section, rate of convergence estimates for barrier function algorithms are considered. Unless stated otherwise the conditions imposed in Section 3 are assumed, together with the condition that $\|\nabla g_i(\underset{\sim}{x}^*)\| \neq 0$, $i \in B_0$.

Lemma 4.1: Provided $\{\underset{\sim}{u}_k\}$ is bounded then

$$f(\underset{\sim}{x}_k) - f(\underset{\sim}{x}^*) = \sum_{i=1}^{t} u_i^k \, g_i(\underset{\sim}{x}_k)$$

$$+ \; O(\max\{\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|^2 , r_k\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|\}) \quad . \tag{4.1}$$

Proof: The result follows by taking the scalar product of (2.2) with $\underset{\sim}{x}_k - \underset{\sim}{x}^*$ and identifying with terms in the Taylor series expansion. □

Definition: We say that u_k is $SO(v_k)$ (strict order v_k) provided

(i) $u_k = O(v_k)$, and

(ii) $\exists\, k_0 < \infty$ and $\mu > 0$ such that $|u_k| \geq \mu |v_k|$ for $k \geq k_0$.

Remark: (4.1) gives an error estimate provided the remainder term is small. A sufficient condition for this is

$$f(\underset{\sim}{x}_k) - f(\underset{\sim}{x}^*) = SO(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|) \quad . \tag{4.2}$$

This implies that for at least one i , $u_i^k g_i(\underset{\sim}{x}_k) = SO(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|)$.

If (4.2) does not hold then for $i = 1,2,\ldots,t$ either

(i) $u_i^k \to 0$, $k \to \infty$, or

(ii) $g_i(\underset{\sim}{x}_k) = o(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|)$.

If (ii) holds then the approach of $\underset{\sim}{x}_k$ to $\underset{\sim}{x}^*$ is tangential to the surface $g_i(\underset{\sim}{x}) = 0$ at $\underset{\sim}{x} = \underset{\sim}{x}^*$.

<u>Lemma 4.2</u>: If the ICP is convex, then

(i) $\underset{\sim}{x}_k, \underset{\sim}{u}_k$ are dual feasible, and

(ii) $f(\underset{\sim}{x}_k) - f(\underset{\sim}{x}^*) \leq \sum_{i=1}^{m} u_i^k g_i(\underset{\sim}{x}_k)$.

<u>Proof</u>: The dual feasibility is a consequence of (2.1) and assumption (i) of Section 3. This follows directly from Wolfe's form of the duality theorem (MP Corollary 5.1). We have

$$\mathcal{L}(\underset{\sim}{x}_k, \underset{\sim}{u}_k) = f(\underset{\sim}{x}_k) - \sum_{i=1}^{m} u_i^k g_i(\underset{\sim}{x}_k) \leq \mathcal{L}(\underset{\sim}{x}^*, \underset{\sim}{u}^*) = f(\underset{\sim}{x}^*) \leq f(\underset{\sim}{x}_k) \qquad (4.3)$$

which demonstrates the second part of the desired result. □

<u>Example</u>: For $i \epsilon B_0$ let $g_i(\underset{\sim}{x}_k) = SO(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|)$, $u_i^* > 0$.

(a) inverse barrier function. We have

$$u_i^k = r_k / g_i(\underset{\sim}{x}_k)^2$$

whence

$$g_i(\underset{\sim}{x}_k) = \sqrt{r_k / u_i^k} \quad .$$

This gives

$$\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\| = O(r_k^{1/2}) \quad .$$

(b) log barrier functions. In this case

$$u_i^k = r_k / g_i(\underset{\sim}{x}_k)$$

which gives

$$g_i(\underset{\sim}{x}_k) = r_k / u_i^k$$

and

$$\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\| = O(r_k) \quad .$$

Thus the strict order condition permits us to deduce a rate of convergence result. We now show that the SO condition is equivalent to the condition of strict complementarity for the inverse and log barrier functions. In these cases the remark following Lemma 4.1 gives us a geometric interpretation of strict complementarity.

To discuss this equivalence, consider the following system of equations which define $\underset{\sim}{x}_k$ and $\underset{\sim}{u}_k$ as functions of r_k .

$$\nabla f(\underset{\sim}{x}_k) - \sum_{i=1}^{m} u_i^k \nabla g_i(\underset{\sim}{x}_k) = 0 \quad ,$$

and

$$u_i^k \Big/ \frac{\partial \phi}{\partial g_i}(\underset{\sim}{x}_k) = - r_k \quad , \quad i = 1,2,\ldots,m \ . \tag{4.4}$$

If the Jacobian $H(\underset{\sim}{x},\underset{\sim}{u})$ of this system with respect to $\underset{\sim}{x}$, $\underset{\sim}{u}$ or an appropriate transform of it is nonsingular then we can study the behavior of $\underset{\sim}{x}(r)$, $\underset{\sim}{u}(r)$ as a function of r by integrating the system of differential equations

$$H(x,u)\begin{bmatrix} \dfrac{d\underset{\sim}{x}}{dr} \\ \\ \dfrac{d\underset{\sim}{u}}{dr} \end{bmatrix} = -\begin{bmatrix} 0 \\ \\ \underset{\sim}{e} \end{bmatrix} \tag{4.5}$$

where $\underset{\sim}{e}^T = \{1,1,\ldots,1\}$. We have

$$H(\underset{\sim}{x},\underset{\sim}{u}) = \begin{bmatrix} \nabla^2 \mathcal{L}(\underset{\sim}{x},\underset{\sim}{u}) & & -\nabla g_i(\underset{\sim}{x})^T \cdots & \nabla g_m(\underset{\sim}{x})^T \\ \dfrac{-u_1}{\left(\dfrac{\partial\phi}{\partial g_1}\right)^2} & \dfrac{\partial^2\phi}{\partial g_1^2}\nabla g_1(\underset{\sim}{x}) & 1\Big/\dfrac{\partial\phi}{\partial g_1} & \\ & \vdots & & \ddots \\ \dfrac{-u_m}{\left(\dfrac{\partial\phi}{\partial g_m}\right)^2} & \dfrac{\partial^2\phi}{\partial g_m^2}\nabla g_m(\underset{\sim}{x}) & & 1\Big/\dfrac{\partial\phi}{\partial g_m} \end{bmatrix} \tag{4.6}$$

Let D be the diagonal matrix

$$D = \begin{bmatrix} I & & & \\ & w_1 & & \\ & & \ddots & \\ & & & w_m \end{bmatrix} \tag{4.7}$$

where

$$w_i = -\left(\frac{\partial\phi}{\partial g_i}\right)^2 \Big/ \frac{\partial^2\phi}{\partial g_i^2} \tag{4.8}$$

then

$$DH = \begin{bmatrix} \nabla^2 \mathcal{L}(\underset{\sim}{x},\underset{\sim}{u}) & -\nabla g_1(\underset{\sim}{x})^T & \cdots & -\nabla g_m(\underset{\sim}{x})^T \\ u_1\nabla g_1(\underset{\sim}{x}) & -\dfrac{\partial\phi}{\partial g_1}\Big/\dfrac{\partial^2\phi}{\partial g_1^2} & & \\ \vdots & & \ddots & \\ u_m\nabla g_m(\underset{\sim}{x}) & & & -\dfrac{\partial\phi}{\partial g_m}\Big/\dfrac{\partial^2\phi}{\partial g_m^2} \end{bmatrix} \tag{4.9}$$

Provided $\frac{\partial\phi}{\partial g_i}\Big/\frac{\partial^2\phi}{\partial g_i^2} \to 0$, $k \to \infty$, $i \in B_0$ then DH will be non-singular for k large enough provided $J(\underset{\sim}{x}^*)$ is nonsingular (see MP Lemma 4.4). This implies

(i) the second order sufficiency conditions hold,

(ii) the active constraint gradients are linearly independent, and

(iii) strict complementarity holds.

In this case

$$DH\begin{bmatrix} \frac{d\underset{\sim}{x}}{dr} \\ \\ \frac{d\underset{\sim}{u}}{dr} \end{bmatrix} = -\begin{bmatrix} 0 \\ \\ \underset{\sim}{w} \end{bmatrix} \tag{4.10}$$

whence

$$\begin{bmatrix} \underset{\sim}{x}_k - \underset{\sim}{x}^* \\ \\ \underset{\sim}{u}_k - \underset{\sim}{u}^* \end{bmatrix} = -\int_0^{r_k} (DH)^{-1}\begin{bmatrix} 0 \\ \\ \underset{\sim}{w} \end{bmatrix} dr \tag{4.11}$$

provided the components of $\underset{\sim}{w}$ are integrable.

<u>Example</u>: (i) inverse barrier function.

We have

$$\frac{\partial\phi}{\partial g_i}\Big/\frac{\partial^2\phi}{\partial g_i^2} = -\frac{g_i}{2} \ , \quad \left(\frac{\partial\phi}{\partial g_i}\right)^2\Big/\frac{\partial^2\phi}{\partial g_i^2} = \frac{1}{2g_i} \ .$$

In particular, it follows that $DH(\underset{\sim}{x}_k)$ is nonsingular for k large enough, while $-w_i = \frac{1}{2}\sqrt{u_i / r}$. We have

$$\begin{bmatrix} \frac{d\underset{\sim}{x}}{dr} \\ \\ \frac{d\underset{\sim}{u}}{dr} \end{bmatrix} = \frac{1}{2r^{1/2}} \; (DH)^{-1} \begin{bmatrix} 0 \\ \frac{}{\sqrt{u_1}} \\ \vdots \\ \sqrt{u_m} \end{bmatrix} \tag{4.12}$$

whence, changing to $r^{1/2}$ as independent variable,

$$\begin{bmatrix} \frac{d\underset{\sim}{x}}{d(r^{1/2})} \\ \\ \frac{d\underset{\sim}{u}}{d(r^{1/2})} \end{bmatrix} = (DH)^{-1} \begin{bmatrix} 0 \\ \frac{}{\sqrt{u_1}} \\ \vdots \\ \sqrt{u_m} \end{bmatrix} . \tag{4.13}$$

Thus we have

$$\begin{bmatrix} \underset{\sim}{x}(r) - \underset{\sim}{x}^* \\ \\ \underset{\sim}{u}(r) - \underset{\sim}{u}^* \end{bmatrix} = r^{1/2}[DH(\underset{\sim}{x}^*)]^{-1} \begin{bmatrix} 0 \\ \frac{}{\sqrt{u_1^*}} \\ \vdots \\ \sqrt{u_m^*} \end{bmatrix} + o(r^{1/2}) . \tag{4.14}$$

(ii) log barrier function. In this case

$$\frac{\partial\phi}{\partial g_i} \Big/ \frac{\partial^2\phi}{\partial g_i^2} = -g_i \; , \; \left(\frac{\partial\phi}{\partial g_i}\right)^2 \Big/ \frac{\partial^2\phi}{\partial g_i^2} = 1$$

so that (4.10) becomes

$$\begin{bmatrix} \frac{d\underset{\sim}{x}}{dr} \\ \\ \frac{d\underset{\sim}{u}}{dr} \end{bmatrix} = J^{-1} \begin{bmatrix} 0 \\ \\ \underset{\sim}{e} \end{bmatrix} . \tag{4.15}$$

In particular, $\underset{\sim}{x}(r)$, $\underset{\sim}{u}(r)$ inherit the differentiability properties of f and g_i, $i = 1,\ldots,m$ for r small enough (if $f,g \in C^p$ then $\underset{\sim}{x},\underset{\sim}{u} \in C^{p-1}$). Thus

$$\begin{bmatrix} \underset{\sim}{x}(r) - \underset{\sim}{x}^* \\ \\ \underset{\sim}{u}(r) - \underset{\sim}{u}^* \end{bmatrix} = r\, J(\underset{\sim}{x}^*)^{-1} \begin{bmatrix} 0 \\ \\ \underset{\sim}{e} \end{bmatrix} + O(r^2) \quad . \tag{4.16}$$

<u>Remark</u>: These results can also be derived by differentiating (2.1) with respect to r. We obtain

$$\nabla^2 T(\underset{\sim}{x}(r),r)\, \frac{d\underset{\sim}{x}}{dr} = -\sum_{i=1}^{m} \frac{\partial \phi}{\partial g_i}\, \nabla g_i(\underset{\sim}{x}(r))^T \quad . \tag{4.17}$$

In this case Lemma 3.3 guarantees that $\nabla^2 T(\underset{\sim}{x}(r),r)$ is positive definite for r small enough, and Corollary 3.3 can be used to give the solution to (4.17). Note that (4.17) results if $\frac{d\underset{\sim}{u}}{dr}$ is eliminated from (4.10).

We can now proceed to the main result.

<u>Theorem 4.1</u>: Provided $J(\underset{\sim}{x}^*)$ is nonsingular, then the strict complementarity and strict order conditions are equivalent for the inverse and log barrier functions.

<u>Proof</u>: The argument is essentially the same for both barrier functions, but is simplest for the log function. Thus only this case is considered here.

For the log penalty function $u_i^k = r_k/g_i(\underset{\sim}{x}_k)$ so that

$$f(\underset{\sim}{x}_k) - f(\underset{\sim}{x}^*) = \sum_{i \in B_0} \frac{r_k}{g_i(\underset{\sim}{x}_k)} g_i(\underset{\sim}{x}_k) + O(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|)$$

$$= t\, r_k + o(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|) \quad . \tag{4.18}$$

If strict complementarity does not hold then for at least one $i \in B_0$

$$\lim_{k \to \infty} \frac{r_k}{g_i(\underset{\sim}{x}_k)} = 0 \quad .$$

As $g_i(\underset{\sim}{x}_k) = \nabla g_i(\underset{\sim}{x}^*)(\underset{\sim}{x}_k - \underset{\sim}{x}^*) + o(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|)$ is $O(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|)$ at most, it follows that $r_k = o(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|)$ and hence, from (4.18), that $f(\underset{\sim}{x}_k) - f(\underset{\sim}{x}^*) = o(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|)$. Thus the SO condition does not hold.

If strict complementarity does hold then asymptotically for small r

$$\frac{r}{\|\nabla g(\underset{\sim}{x}^*)\| \; \|\underset{\sim}{x}(r) - \underset{\sim}{x}^*\|} \sim \leq \frac{r}{g_i(\underset{\sim}{x}(r))} \to u_i^* > 0 \quad , \quad i \in B_0 \; .$$

Thus $r = O(\|\underset{\sim}{x}(r) - \underset{\sim}{x}^*\|)$. As $J(\underset{\sim}{x}^*)$ is nonsingular (4.16) holds and this implies (as $r = O(\|\underset{\sim}{x}(r) - \underset{\sim}{x}^*\|)$) that

$$\|\underset{\sim}{x}(r) - \underset{\sim}{x}^*\| \leq K r + O(r^2)$$

for some $K > 0$. This shows that $r = SO(\|\underset{\sim}{x}(r) - \underset{\sim}{x}^*\|)$ so that, by (4.18), the strict order condition is satisfied. □

<u>Remark</u>: The above argument shows that if strict complementarity does not hold then the strict order condition cannot. The condition that $J(\underset{\sim}{x}^*)$ be nonsingular is required only for the second part of the theorem.

Example: (i) minimize $x_1 + x_2$

subject to $-x_1^2 + x_2 \geq 0$, $x_1 \geq 0$

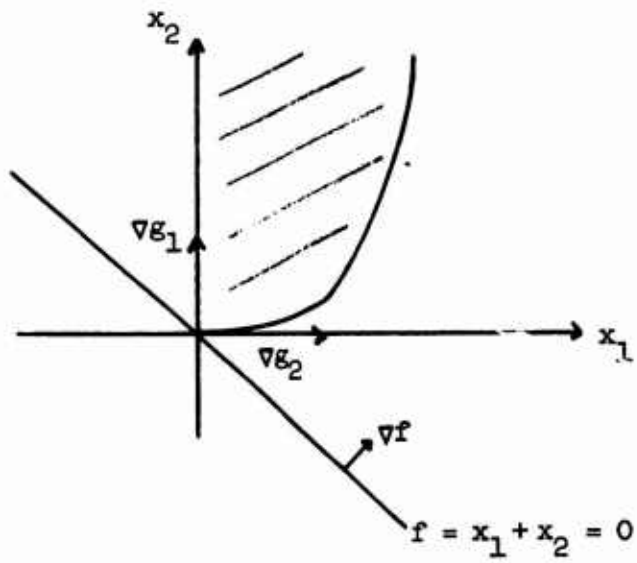

Figure 4.1

From Figure 4.1 it is clear that the minimum is $f = 0$ at $x_1 = x_2 = 0$, and that strict complementarity holds.

(a) inverse barrier function

$$T = x_1 + x_2 + r\left\{\frac{1}{x_2 - x_1^2} + \frac{1}{x_1}\right\}$$

$$\frac{\partial T}{\partial x_1} = 0 = 1 + r\left\{\frac{2x_1}{(x_2 - x_1^2)^2} - \frac{1}{x_1^2}\right\}$$

$$\frac{\partial T}{\partial x_2} = 0 = 1 + r\left\{\frac{-1}{(x_2 - x_1^2)^2}\right\} \quad .$$

This gives a pair of equations for x_1 and x_2 as functions of r. We have

$$x_1 = r^{1/2} - r + O(r^{3/2}) \quad ,$$

$$x_2 = r^{1/2} + r + O(r^{3/2}) \quad .$$

(b) log barrier function

$$T = x_1 + x_2 - r\{\log\,(x_2 - x_1^2) + \log x_1\}$$

$$\frac{\partial T}{\partial x_1} = 0 = 1 - r\left\{\frac{-2x_1}{x_2 - x_1^2} + \frac{1}{x_1}\right\}$$

$$\frac{\partial T}{\partial x_2} = 0 = 1 - r\left\{\frac{1}{x_2 - x_1^2}\right\} \quad .$$

Solving for x_1 and x_2 as functions of r gives

$$x_1 = r - 2r^2 + O(r^3) \quad ,$$

$$x_2 = r + r^2 + O(r^3) \quad .$$

(ii) minimize x_2

subject to $x_2 - x_1^2 \geq 0$, $x_1 \geq 0$.

In this case the minimum is again $f = 0$ at $x_1 = x_2 = 0$. However, $\nabla f(0) = \underset{\sim}{e}_2^T$ is orthogonal to $\nabla g_2(0) = \underset{\sim}{e}_1^T$. Thus, as both constraints are active at zero, strict complementarity does not hold. Note that the constraint $g_2 = x_1 \geq 0$ is redundant, and that the barrier function trajectory is tangential to the constraint surface $g_1 = 0$. Note also that the rate of convergence is reduced, and that $r_k \dfrac{\partial^2 \phi}{\partial g^2}$ does not tend to ∞ for the constraint with the zero multiplier.

(a) inverse barrier function.

$$T = x_2 + r\left\{\frac{1}{x_2 - x_1^2} + \frac{1}{x_1}\right\}$$

$$\frac{\partial T}{\partial x_1} = 0 = r\left\{\frac{2x_1}{(x_2 - x_1^2)^2} - \frac{1}{x_1^2}\right\}$$

$$\frac{\partial T}{\partial x_2} = 0 = 1 + r\left\{\frac{-1}{(x_2 - x_1^2)^2}\right\}$$

whence

$$x_1 = (r/2)^{1/3} \quad , \quad u_1 = 1 \quad , \quad r\,\frac{\partial^2 \phi}{\partial g_1^2} = \frac{2}{r^{1/2}} \quad ,$$

$$x_2 = 3r/2 \quad , \quad u_2 = (4r)^{1/3} \quad , \quad r\,\frac{\partial^2 \phi}{\partial g_2^2} = 4 \quad .$$

(b) log barrier function.

$$T = x_2 - r\{\log(x_2 - x_1^2) + \log x_1\}$$

$$\frac{\partial T}{\partial x_1} = 0 = -r\left\{\frac{-2x_1}{x_2 - x_1^2} + \frac{1}{x_1}\right\}$$

$$\frac{\partial T}{\partial x_2} = 0 = 1 - r\left\{\frac{1}{x_2 - x_1^2}\right\}$$

whence

$$x_1 = (r/2)^{1/2} \quad , \quad u_1 = 1 \quad , \quad r\,\frac{\partial^2 \phi}{\partial g_1^2} = \frac{1}{r}$$

$$x_2 = 3r/2 \quad , \quad u_2 = (2r)^{1/2} \quad , \quad r\,\frac{\partial^2 \phi}{\partial g_2^2} = 2 \quad .$$

(iii) minimize $-x_1$

subject to $(1-x_1)^3 - x_2 \geq 0$, $x_2 \geq 0$.

This is the example used in MP Section 3 (see Figure 3.1). The optimum is $f = -1$ at $x_1 = 1$, $x_2 = 0$. The Kuhn Tucker conditions do not hold at this point.

(a) inverse barrier function.

$$T = -x_1 + r\left\{\frac{1}{(1-x_1)^3 - x_2} + \frac{1}{x_2}\right\}$$

$$\frac{\partial T}{\partial x_1} = 0 = -1 + r\left\{\frac{3(1-x_1)^2}{((1-x_1)^3 - x_2)^2}\right\}$$

$$\frac{\partial T}{\partial x_2} = 0 = r\left\{\frac{1}{((1-x_1)^3 - x_2)^2} - \frac{1}{x_2^2}\right\}$$

whence

$$x_1 = 1 - 2^{1/2}\, 3^{1/4}\, r^{1/4} \quad ,$$

$$x_2 = 2^{1/2}\, 3^{3/4}\, r^{3/4} \quad .$$

In this case $u_1(r) = u_2(r) = \dfrac{1}{2 \cdot 3^{3/2} \cdot r^{1/2}}$

(b) log barrier function.

$$T = -x_1 - r\{\log((1-x_1)^3 - x_2) + \log x_2\}$$

$$\frac{\partial T}{\partial x_1} = 0 = -1 - r\left\{\frac{-3(1-x_1)^2}{(1-x_1)^3 - x_2}\right\}$$

$$\frac{\partial T}{\partial x_2} = 0 = -r\left\{\frac{-1}{(1-x_1)^3 - x_2} + \frac{1}{x_2}\right\}$$

whence

$$x_1 = 1 - 6r \quad ,$$

$$x_2 = 108r^3 \quad .$$

In this case $u_1(r) = u_2(r) = \dfrac{1}{108r^2}$.

The above examples confirm the predictions of our analysis, and for a given fixed sequence of r_k values effective convergence is attained more rapidly (i.e., for earlier members of the sequence) with the log barrier function.

Now let ϕ be a barrier function. Then

$$\phi_1 = \log(\sigma + \phi) \quad , \quad \sigma \geq 1 \tag{4.19}$$

is a barrier function. Let $\underset{\sim}{x}_k$ minimize $f + r_k\phi$, $\hat{\underset{\sim}{x}}_k$ minimize $f + r_k\phi_1$. Then comparing corresponding Lagrange multiplier estimates gives

$$\frac{\dfrac{\partial\phi}{\partial g_i}(\hat{\underset{\sim}{x}}_k)}{(\sigma + \phi(\hat{\underset{\sim}{x}}_k))\dfrac{\partial\phi}{\partial g_i}(\underset{\sim}{x}_k)} \to 1 \quad , \quad r_k \to 0$$

whence

$$\frac{\partial\phi}{\partial g_i}(\underset{\sim}{x}_k) = o\left(\frac{\partial\phi}{\partial g_i}(\hat{\underset{\sim}{x}}_k)\right) \qquad \text{as} \quad r_k \to 0 \quad .$$

Essentially this says that $g_i(\hat{\underset{\sim}{x}}_k) \to 0$ more rapidly than $g_i(\underset{\sim}{x}_k)$, so that a faster rate of convergence is anticipated for the ϕ_1 barrier function.

Consider now the sequence of barrier functions defined recursively by

$$\phi_j^{(1)} = \log(k_j - \log(g_j(\underset{\sim}{x})) \quad ,$$

$$\phi_j^{(i)} = \log(\sigma + \phi_j^{(i-1)}) \quad , \quad i = 2,3,\ldots, \quad \sigma > 1 \ ,$$

$$\phi^{(i)} = \sum_{j=1}^{m} \phi_j^{(i)} \quad . \tag{4.20}$$

In this case the error estimate is

$$f(\underset{\sim}{x}_k^{(i)}) - f(\underset{\sim}{x}^*) = -r_k \sum_{j=1}^{t} \frac{\partial \phi_j^{(i)}}{\partial g_j} g_j(\underset{\sim}{x}_k) = \sum_{j=1}^{t} \frac{r_k}{k_j - \log(g_j)} \prod_{s=2}^{i-1} \frac{1}{\sigma + \phi_j^{(s)}} \tag{4.21}$$

The right hand side of (4.21) tends to zero as $i \to \infty$, and this suggests that increasingly rapid rates of convergence can be obtained by using barrier functions associated with large values of i .

However, an even more interesting result is possible. This shows that in certain circumstances it is possible to choose a barrier function having the property that the solution to the ICP is approximated arbitrarily closely by the result of a single unconstrained minimization, without requiring r to be taken arbitrarily small. Let

$$T^{(i)}(\underset{\sim}{x},\underset{\sim}{r}) = f(\underset{\sim}{x}) + r \sum_{j=1}^{m} \phi_j^{(i)}(g_j(\underset{\sim}{x})) \quad , \quad \text{and}$$

$$Q(\underset{\sim}{x},\underset{\sim}{\lambda}) = f(\underset{\sim}{x}) + \sum_{j=1}^{m} \lambda_j(k_j - \log(g_j(\underset{\sim}{x}))) \quad .$$

Theorem 4.2: Let $Q(\underset{\sim}{x},\underset{\sim}{\lambda})$ have a unique stationary value (necessarily a minimum) in S_0 for each $\underset{\sim}{\lambda} > 0$, and let $\underset{\sim}{x}^{(i)}$ minimize $T^{(i)}(\underset{\sim}{x},r)$ for $i = 1,2,\ldots$ and fixed r. Then the limit points of $\{\underset{\sim}{x}^{(i)}\}$ are local minima of the ICP.

Remark: Note that r does not have to be small in this result.

Proof: If $\underset{\sim}{x}^{(i)}$ minimizes $T^{(i)}(\underset{\sim}{x},r)$ then

$$\nabla f(\underset{\sim}{x}^{(i)}) - r \sum_{j=1}^{m} \frac{1}{k_j - \log(g_j(\underset{\sim}{x}^{(i)}))} \prod_{s=2}^{i-1} \frac{1}{\sigma + \phi_j^{(s)}(g_j(\underset{\sim}{x}^{(i)}))} \frac{1}{g_j(\underset{\sim}{x}^{(i)})} \nabla g_j(\underset{\sim}{x}^{(i)})$$

$$= 0 \quad , \tag{4.22}$$

and this expression has the form

$$\nabla Q(\underset{\sim}{x}^{(i)},\underset{\sim}{\lambda}^{(i)}) = 0 \tag{4.23}$$

where $\lambda_j^{(i)}$ has the numerical value

$$\lambda_j^{(i)} = \frac{r}{k_j - \log(g_j(\underset{\sim}{x}^{(i)}))} \prod_{s=2}^{i-1} \frac{1}{\sigma + \phi_j^{(s)}(g_j(\underset{\sim}{x}^{(i)}))} \quad , \quad j = 1,2,\ldots,m \; . \tag{4.24}$$

Thus the $\{\underset{\sim}{x}^{(i)}\}$ also correspond to a sequence minimizing $Q(\underset{\sim}{x},\underset{\sim}{\lambda}^{(i)})$ by the assumed uniqueness of these stationary values. Now, as $\sigma > 1$, $\phi_j^{(s)} > 0$, $s = 1,2,\ldots,i-1$, $\lambda_j^{(i)}$ can be made arbitrarily small for each j by choosing i large enough. The desired result is thus a consequence of the remark following Theorem 1.2. □

Remark: The conditions of the theorem are satisfied if $f(\underset{\sim}{x})$ convex, $g_i(\underset{\sim}{x})$, $i = 1,2,\ldots,m$ concave, and strict convexity / concavity holds for at least one of these functions.

In what follows it is convenient to use the superscript i to indicate the appropriate member of the log barrier function sequence (4.20).

Lemma 4.3:

$$\frac{\partial^2 \phi^{(i)}}{\partial g_j^2} \Big/ \frac{\partial \phi^{(i)}}{\partial g_j} = -\frac{\rho_j^{(i)}}{g_j}, \quad i = 1,2,\ldots, \tag{4.25}$$

where

$$\rho_j^{(i)} = \rho_j^{(i-1)} + g_j \frac{\partial \phi^{(i)}}{\partial g_j} < \rho_j^{(i-1)}, \tag{4.26}$$

and

$$\rho_j^{(i)} \to 1, \quad g_j \to 0 . \tag{4.27}$$

Proof: Let $\phi^{(0)} = -\log g_j$ then

$$\frac{\partial^2 \phi^{(0)}}{\partial g_j^2} \Big/ \frac{\partial \phi^{(0)}}{\partial g_j} = -\frac{1}{g_j} \tag{4.28}$$

so that $\rho_j^{(0)} = 1$. Now, differentiating the relation

$$\frac{\partial \phi^{(i+1)}}{\partial g_j} = \frac{1}{\sigma + \phi^{(i)}} \frac{\partial \phi^{(i)}}{\partial g_j} \tag{4.29}$$

gives

$$\frac{\partial^2 \phi^{(i+1)}}{\partial g_j^2} = -\frac{1}{(\sigma + \phi^{(i)})^2} \left(\frac{\partial \phi^{(i)}}{\partial g_j} \right)^2 + \frac{1}{\sigma + \phi^{(i)}} \frac{\partial^2 \phi^{(i)}}{\partial g_j^2}$$

so that

$$\frac{\partial^2\phi^{(i+1)}}{\partial g_j^2} \Big/ \frac{\partial\phi^{(i+1)}}{\partial g_j} = -\frac{\partial\phi^{(i+1)}}{\partial g_j} + \frac{\partial^2\phi^{(i)}}{\partial g_j^2} \Big/ \frac{\partial\phi^{(i)}}{\partial g_j}$$

$$= -\frac{1}{g_j}\left(\rho^{(i)} + g_j \frac{\partial\phi^{(i+1)}}{\partial g_j}\right) .$$

This demonstrates (4.25) and (4.26), (4.27) follows on noting that $\rho_j^{(0)} = 1$, and that, from (4.29),

$$g_j \frac{\partial\phi^{(i)}}{\partial g_j} \to 0 \quad , \quad g_j \to 0 \quad , \quad i = 1,2,\ldots \quad . \ \square$$

A consequence of this lemma is that, provided $J(\underset{\sim}{x}^*)$ nonsingular, then $D^{(i)}H^{(i)}$ is nonsingular for $\underset{\sim}{x}(r)$ sufficiently close to $\underset{\sim}{x}^*$.

<u>Lemma 4.4</u>: Let $J(\underset{\sim}{x}^*)$ be nonsingular, and $I_1 = B_0$, then the SO condition is satisfied.

<u>Proof</u>: We have from (4.29) that

$$w_j^{(i)} = -\left(\frac{\partial\phi^{(i)}}{\partial g_j}\right)^2 \Big/ \frac{\partial^2\phi^{(i)}}{\partial g_j^2} = \frac{g_j}{\rho_j^{(i)}} \frac{\partial\phi^{(i)}}{\partial g_j} , \tag{4.31}$$

so that $w_j^{(i)} \to 0$ as $r \to 0$ for $j \in B_0$. Now

$$\left\| \begin{bmatrix} \underset{\sim}{x}_k - \underset{\sim}{x}^* \\ \\ \underset{\sim}{u}_k - \underset{\sim}{u}^* \end{bmatrix} \right\| \le \left\| [D^{(i)}H^{(i)}(\underset{\sim}{x}_k)]^{-1} \begin{bmatrix} I & \\ & P^{(i)} \end{bmatrix} \begin{bmatrix} 0 \\ \\ u_1^k g_1 \\ \vdots \\ u_m^k g_m \end{bmatrix} \right\|$$

$$+ \text{ smaller terms}$$

where $P^{(i)}$ is diagonal, and $P_{jj}^{(i)} = 1/\rho_j^{(i)} \to 1$, $r_k \to 0$, $j = 1,2,\ldots,m$. This result implies that for k large enough

$$\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\| \leq K \sum_{j \in B_0} u_j^k \, g_j(\underset{\sim}{x}_k) \quad .$$

The SO condition is an immediate consequence of this inequality. □

<u>Remark</u>: If $B_0 \subset I_1$ then $w_j^{(i)}$ need not tend to zero for $j \notin B_0$. Thus eventually the largest components of $\underset{\sim}{w}^{(i)}$ will be those associated with the <u>inactive constraints</u>. This implies that $\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\| = O(r_k)$. But g_j, $j \in B_0$, is $o(r_k)$ which suggests that, in general, the SO condition does not apply. This case should be contrasted with the log and inverse cases where the contributions of the inactive constraints do not dominate in $\underset{\sim}{w}$ (in the inverse case the active constraints dominate). We note that the SO condition is only sufficient for (4.1) to provide an error estimate and numerical experience indicates that it is applicable in the calculations with the log sequence. However, the above discussion suggests that to attain the maximum rate of convergence with the members of the log sequence, the inactive constraints should be identified and discarded. A possible way to do this automatically is by the use of a separable barrier objective function

$$Q(\underset{\sim}{x}, \underset{\sim}{p}_k) = f(\underset{\sim}{x}) + r_k \sum_{i=1}^{m} u_i^{k-1} \, \phi(g_i(\underset{\sim}{x})) \tag{4.32}$$

where $\underset{\sim}{p}_k = r_k \underset{\sim}{u}^{k-1}$, r_k is the usual barrier parameter, and u_i^{k-1} is the multiplier estimate obtained from the previous minimization. This objective function has the property of forcing the multiplier estimates

for the inactive constraints to zero at a very fast rate. We have

$$u_i^{k+1} = - r_{k+1} u_i^k \frac{\partial \phi}{\partial g_i} (\underset{\sim}{x}_{k+1})$$

$$\leq r_{k+1} \rho_i u_i^k \leq \cdots \leq \rho_i^{k+1} \prod_{j=1}^{k+1} r_j \, u_i^0 \tag{4.33}$$

where ρ_i is a bound for $\frac{\partial \phi}{\partial g_i} (\underset{\sim}{x}_k)$, $k = 1,2,\ldots$.

This choice can also be favorable in the case of nonstrict complementarity. Consider the previous example

$$\min x_2 \quad \text{subject to} \quad x_2 - x_1^2 \geq 0 \ , \quad x_1 \geq 0 \ .$$

Set $Q = x_2 - r_2\{\log(x_2 - x_1^2) + u_{k-1} \log x_1\}$. Then $\nabla Q = 0$ gives

$$x_2 - x_1^2 = r_k \ , \quad 2x_1^2 = r_k u_{k-1} \ ,$$

so that

$$u_k = \frac{r_k u_{k-1}}{x_1} = 2^{1/2} r_k^{1/2} u_{k-1}^{1/2} \ .$$

Setting $r_k = \alpha^{-k}$, $u_k/2 = \alpha^{\beta_k}$ reduces this to

$$\beta_k = \frac{1}{2} \beta_{k-1} - \frac{k}{2} \ .$$

The solution to this difference equation satisfying the initial condition $\beta_0 = 0$ is

$$\beta_k = - k + (1 - (\tfrac{1}{2})^k) \ .$$

From this it follows that $u_k = O(r_k)$, and hence that $x_1 = O(r_k)$. Thus, for this example, we are able to obtain results as favorable as those in which strict complementarity holds.

<u>Example</u>. Show that the error estimate (4.1) is valid in this case.

There is a penalty to pay for the generality of the barrier function algorithms, and this is a significant burden of calculation associated with each of the successive unconstrained minimizations. This can be explained (at least in part) by looking at the Hessian of the barrier objective function. Experience (in part supported by theoretical results) indicates that the condition number of the Hessian is a good indicator of the degree of difficulty of an unconstrained optimization problem when it is solved by descent methods.

On the assumptions that the second order sufficiency conditions hold at $\underset{\sim}{x}^*$, and that the active constraint gradients are linearly independent, then it is possible to deduce fairly complete information on the eigenvalues and eigenvectors of $\nabla^2 T(\underset{\sim}{x}_k, r_k)$ from (3.8).

(i) There are $n-t$ eigenvectors associated with eigenvalues of $\nabla^2 T(\underset{\sim}{x}_k, r_k)$ that are $O(1)$ as $r_k \to 0$. The smallest eigenvalue tends to

$$m = \min_{\underset{\sim}{v}} \frac{\underset{\sim}{v}^T \nabla^2 \mathcal{L}(\underset{\sim}{x}^*, \underset{\sim}{u}^*) \underset{\sim}{v}}{\underset{\sim}{v}^T \underset{\sim}{v}} \quad , \quad \forall \underset{\sim}{v} \text{ such that } \nabla g_i(\underset{\sim}{x}^*) \underset{\sim}{v} = 0 \quad , \quad \forall i \in B_0 \quad .$$

(ii) There are t eigenvectors associated with eigenvalues of $\nabla^2 T(\underset{\sim}{x}_k, r_k)$ which tend to ∞ as $r_k \to 0$. These eigenvectors are asymptotic to vectors of the form $C_1(\underset{\sim}{x}^*) \underset{\sim}{y}_i$ where $\underset{\sim}{y}_i$ are eigenvectors of the problem

$$[C_1(\underset{\sim}{x}^*)^T C_1(\underset{\sim}{x}^*) - \mu_i A] \underset{\sim}{y}_i = 0$$

where A is a diagonal matrix, $A_{ii} = \dfrac{\partial^2 \phi}{\partial g_i^2} \Big/ \max\limits_{1 \le j \le t} \dfrac{\partial^2 \phi}{\partial g_j^2}$, $i = 1, \ldots, t$.

The corresponding eigenvalues tend to ∞ like $\mu_i r_k \max_{1 \le j \le t} \frac{\partial^2 \phi}{\partial g_j^2}$

-- that is, like $\mu_i \frac{u_\alpha^*}{g_\alpha(\underset{\sim}{x}_k)}$ where α is the maximizing index.

This shows that the condition number of $\nabla^2 T(\underset{\sim}{x}_k, r_k)$ tends to ∞ like $1/g_\alpha(\underset{\sim}{x}_k)$ or like $1/\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|$ if the SO condition is satisfied. In this latter case we have shown that our measure of the cost of a barrier function calculation <u>depends in the main on the accuracy desired rather than on the choice of barrier functions</u>. However, our estimates for the log family indicate that these will be somewhat more expensive than the above estimate except when all constraints are active.

Note that the device introduced to force more effective elimination of the inactive constraints does not force the Hessian to be worse conditioned in the case that strict complementarity does not obtain, at least in the examples that have been worked out. The use of this device would appear to be an important improvement in barrier function algorithms.

5. Analysis of penalty function methods.

Consider now the equality constrained problem (ECP)

$$\min_{\underset{\sim}{x} \in S} f(\underset{\sim}{x}) \quad , \quad S = \{\underset{\sim}{x} \,;\, h_i(\underset{\sim}{x}) = 0 \,,\, i = 1,2,\ldots,q \ (i \in I_2)\} \,. \qquad (3.1)$$

It is assumed that S is nonempty, and that (3.1) has a bounded minimum (say $\bar{f}$).

Remark: The inequality constraint $g(\underset{\sim}{x}) \geq 0$ can be written as the equality constraint

$$h(\underset{\sim}{x}) = \min\,(0, g(\underset{\sim}{x})) = 0 \tag{5.2}$$

so that formally the ICP is a special case of an ECP. However $h(\underset{\sim}{x})$ given by (5.2) can have discontinuous first derivatives.

Definition: $F(\underset{\sim}{x},\lambda)$ is a penalty objective function if

$$F(\underset{\sim}{x},\lambda) = f(\underset{\sim}{x}) + \lambda \sum_{i=1}^{q} \psi(h_i(\underset{\sim}{x})) \tag{5.3}$$

where $\psi(h)$ is a monotonic increasing function of $|h|$, and $\psi(0) = 0$.

Example: Let $\psi(h) = |h|^{1+\alpha}$ then ψ is a penalty function if $\alpha > -1$. If $g(\underset{\sim}{x})$ is concave then, from (5.2), so is $h(\underset{\sim}{x})$, and $\psi(h)$ is convex provided $\alpha \geq 0$. If $\alpha < 0$ then $\frac{d\psi}{dh}$ is unbounded as $h \to 0$.

Theorem 5.1: Let $\{\lambda_i\} \uparrow \infty$, and $\underset{\sim}{x}_r$ minimize $F(\underset{\sim}{x},\lambda_r)$. Then $\{f(\underset{\sim}{x}_r)\}$ nondecreasing, $\{F(\underset{\sim}{x}_r,\lambda_r)\}$ strictly increasing unless $\underset{\sim}{x}_r \epsilon S$, and $\left\{\sum_{i=1}^{q} \psi(h_i(\underset{\sim}{x}_r))\right\}$ nonincreasing. If $\{\underset{\sim}{x}_r\} \to \underset{\sim}{x}^*$ then $\underset{\sim}{x}^*$ solves the ECP.

Proof: Let $\lambda_r < \lambda_s$. Then, provided $\underset{\sim}{x}_r, \underset{\sim}{x}_s \notin S$,

$$F(\underset{\sim}{x}_r,\lambda_r) \leq F(\underset{\sim}{x}_s,\lambda_r) < F(\underset{\sim}{x}_s,\lambda_s) \leq F(\underset{\sim}{x}_r,\lambda_s) \quad .$$

Thus (compare Lemma 1.2) the results for the sequences follow as before. We have

$$\min_{\underset{\sim}{x}} F(\underset{\sim}{x},\lambda) \leq \min_{\underset{\sim}{x}\epsilon S} F(\underset{\sim}{x},\lambda) = \min_{\underset{\sim}{x}\epsilon S} f(\underset{\sim}{x}) = \bar{f} \quad . \tag{5.4}$$

Thus the $F(\underset{\sim}{x}_r,\lambda_r)$ are bounded, and hence $\underset{\sim}{x}^* \epsilon S$. Now

$$\underset{\sim}{x}^* \epsilon S \Rightarrow f(\underset{\sim}{x}^*) \geq \bar{f} \quad ,$$

but, by (5.4),

$$f(\underset{\sim}{x}_r) \leq F(\underset{\sim}{x}_r,\lambda_r) \leq \bar{f}$$

so that

$$\lim_{r\to\infty} f(\underset{\sim}{x}_r) \leq \bar{f} \quad .$$

Thus $f(\underset{\sim}{x}^*) = \bar{f}$, and $\underset{\sim}{x}^*$ solves the ECP. □

<u>Remark</u>: In the more general case in which $\{\underset{\sim}{x}_r\}$ is bounded it follows, by restricting attention to convergent subsequences, that all limit points of $\{\underset{\sim}{x}_r\}$ solve the ECP.

<u>Theorem 5.2</u>: Let $\{\underset{\sim}{x}_r\} \to \underset{\sim}{x}^*$ and assume $\psi(h_i)$ continuously differentiable, and $\nabla h_i(\underset{\sim}{x}^*)$, $i\epsilon I_2$, linearly independent. Define $\underset{\sim}{u}_r$ by

$$u_i^r = -\lambda_r \frac{\partial\psi}{\partial|h_i|} \operatorname{sgn}(h_i) \quad , \quad i = 1,2,\ldots,q \tag{5.5}$$

then $\{\underset{\sim}{u}_r\} \to \underset{\sim}{u}^*$, the vector of Lagrange multipliers for the ECP.

<u>Proof</u>: Define the matrix $B(\underset{\sim}{x}_r)$ by $\kappa_i(B(\underset{\sim}{x}_r)) = \nabla h_i(\underset{\sim}{x}_r)^T$, $i = 1,2,\ldots,q$. The condition that x_r minimize $F(\underset{\sim}{x},\lambda_r)$ gives

$$0 = \nabla F(\underset{\sim}{x}_r,\lambda_r) = \nabla f(\underset{\sim}{x}_r) + \lambda_r \sum_{i=1}^{q} \frac{\partial\psi}{\partial|h_i|} \operatorname{sgn}(h_i)\nabla h_i(\underset{\sim}{x}_r) \tag{5.6}$$

so that

$$\nabla f(\underset{\sim}{x}_r)^T = B(\underset{\sim}{x}_r)\underset{\sim}{u}_r \quad . \tag{5.7}$$

Now $B(\underset{\sim}{x}_r)$ has full rank for $\|\underset{\sim}{x}_r - \underset{\sim}{x}^*\|$ small enough. Thus

$$\underset{\sim}{u}_r = (B(\underset{\sim}{x}_r)^T B(\underset{\sim}{x}_r))^{-1} B(\underset{\sim}{x}_r)^T \nabla f(\underset{\sim}{x}_r)^T$$

$$\to (B(\underset{\sim}{x}^*)^T B(\underset{\sim}{x}^*))^{-1} B(\underset{\sim}{x}^*)^T \nabla f(\underset{\sim}{x}^*)^T = \underset{\sim}{u}^* \quad . \ \square \tag{5.8}$$

<u>Remark</u>: (i) If strict complementarity holds so that $|u_i^*| > 0$, $i = 1,2,\ldots,q$, then the convergence of the Lagrange multiplier estimates implies (from (5.5)) that $\mathrm{sgn}(h_i)$ is constant for i large enough as $\frac{\partial\psi}{\partial|h_i|} > 0$, $\underset{\sim}{x} \epsilon S$. Thus the minimizing sequence approaches S 'from one side'. In this sense S acts like a barrier.

(ii) Note that $h_i(\underset{\sim}{x}) = \min(0, g_i(\underset{\sim}{x})) = 0$ identically in a neighborhood of $\underset{\sim}{x}^*$ if $g_i(\underset{\sim}{x}^*) > 0$. Thus $\nabla h_i(\underset{\sim}{x}^*) = \underset{\sim}{\theta}$ so that, in this trivial sense, the constraint gradients are not linearly independent. However, if strict complementarity holds, then a multiplier result can be proved for the active constraints (do this!). In fact, the strict complementarity restriction can be relaxed somewhat.

<u>Theorem 5.3</u>: If the conditions of Theorem 5.2 hold, and, in addition, $\{\underset{\sim}{u}_r\} \to \underset{\sim}{u}^*$, and $\lambda_r \frac{d^2\psi}{dh_i^2} \to \infty$, $i = 1,2,\ldots,q$, as $r \to \infty$, then the second order sufficiency conditions hold at $\underset{\sim}{x}^*$ if and only if $\nabla^2 F(\underset{\sim}{x}_r, \lambda_r)$ is positive definite for r sufficiently large.

<u>Proof</u>: This is essentially the same as that of Lemmas 3.1 and 3.2. $\square$

<u>Example</u>: Derive the analogues of Lemmas 3.1 and 3.3 which apply when the ECP is obtained by transforming an ICP by means of (5.2).

Remark: The condition that $\lambda_r \frac{d^2\psi}{dh_i^2} \to \infty$ is related to strict complementarity. Consider $\psi = |h_i|^{1+\alpha}$, $\alpha > 0$. Then

$$\frac{d\psi}{dh_i} = (1+\alpha)|h_i|^{\alpha} \operatorname{sgn}(h_i) , \quad -\lambda_r \frac{d\psi}{dh_i} = u_i^r \tag{5.9}$$

so that

$$\lambda_r \frac{d^2\psi}{dh_i^2} = (1+\alpha)\alpha|h_i|^{\alpha-1} = \frac{\alpha|u_i^r|}{|h_i|} . \tag{5.10}$$

Thus $\lambda_r \frac{d^2\psi}{dh^2} \to \infty$, $h_i \to 0$ if $|u_i| > 0$. Strict complementarity is of particular importance for equality constraints derived from inequality constraints by (5.2). In this case, the one sided convergence implied by the multiplier relations is needed if we are to be able to talk about second derivatives at all.

The parallel development of the treatment of the ECP by penalty function methods and the treatment of the ICP by barrier functions can be completed by discussing convergence rates of penalty function algorithms in much the same way as we treated the barrier function case. For example, multiplying (5.6) by $\underset{\sim}{x}_r - \underset{\sim}{x}^*$ gives

$$f(\underset{\sim}{x}^*) - f(\underset{\sim}{x}_r) = \sum_{i=1}^{q} u_i^r h_i(\underset{\sim}{x}_r) + O(\|\underset{\sim}{x}_r - \underset{\sim}{x}^*\|^2) . \tag{5.11}$$

The assumption that the SO condition is satisfied can now be used to provide estimates. From (5.9)

$$(1+\alpha)\lambda_r|h_i|^{\alpha} \operatorname{sgn}(h_i) = -u_i^r$$

so that

$$|h_i| = \left\{ \frac{|u_i^r|}{(1+\alpha)\lambda_r} \right\}^{1/\alpha} . \tag{5.12}$$

This suggests a rate of convergence of $O((\frac{1}{\lambda_r})^{1/\alpha})$, which contrasts favorably with the estimates obtained for the barrier function algorithms. In particular as $\alpha \to 0$, (5.12) suggests that the convergence rate becomes arbitrarily great. However, the results of the previous section also indicate that the condition number of the Hessian will become arbitrarily large as $\alpha \to 0$. The next result provides information on the limiting case $\alpha = 0$.

<u>Theorem 5.4</u>: In the ICP let $f(\underset{\sim}{x})$ be convex, and $g_i(\underset{\sim}{x})$, $i \in I_1$, concave. Let $\underset{\sim}{w}$ be an infeasible point, $\underset{\sim}{x}_0$ an interior point of S , $a = \min_{i \in I_1} g_i(\underset{\sim}{x}_0)$, $b = f(\underset{\sim}{x}_0) - f(\underset{\sim}{x}^*)$, and $\lambda_0 = (b+1)/a$. Then $\underset{\sim}{x}^*$ minimizes

$$F(\underset{\sim}{x},\lambda) = f(\underset{\sim}{x}) - \lambda \sum_{i=1}^{p} \min(0, g_i(\underset{\sim}{x})) \tag{5.13}$$

provided $\lambda \geq \lambda_0$.

<u>Remark</u>: It is necessary to demonstrate the result only for $\lambda = \lambda_0$. For all larger λ it is then a consequence of Theorem 5.1.

<u>Proof</u>: Let $\underset{\sim}{v}$ be the boundary point of S on the join of $\underset{\sim}{w}$ and $\underset{\sim}{x}_0$, and B_v be the index set of constraints active at $\underset{\sim}{v}$. Define

$$s(\underset{\sim}{x}) = f(\underset{\sim}{x}) - \lambda_0 \sum_{i \in B_v} g_i(\underset{\sim}{x}) , \tag{5.14}$$

then

$$s(\underset{\sim}{x}_0) = f(\underset{\sim}{x}_0) - \lambda_0 \sum_{i \in B_v} g_i(\underset{\sim}{x}_0)$$

$$\leq f(\underset{\sim}{x}_0) - (b+1)$$

$$= f(\underset{\sim}{x}^*) - 1$$

$$\leq f(\underset{\sim}{v}) = s(\underset{\sim}{v}) = F(\underset{\sim}{v}, \lambda_0) \quad . \tag{5.15}$$

As $s(\underset{\sim}{x})$ is convex and $\underset{\sim}{v}$ is on the join of $\underset{\sim}{x}_0$ and $\underset{\sim}{w}$, $\exists\, \theta$, $0 < \theta < 1$, such that

$$s(\underset{\sim}{v}) = \theta s(\underset{\sim}{x}_0) + (1-\theta) s(\underset{\sim}{w})$$

$$\leq \theta s(\underset{\sim}{v}) + (1-\theta) s(\underset{\sim}{w})$$

whence

$$s(\underset{\sim}{v}) \leq s(\underset{\sim}{w}) \quad . \tag{5.16}$$

Now $s(\underset{\sim}{w}) \leq F(\underset{\sim}{w}, \lambda_0)$ so that, from (5.15) and (5.16)

$$F(\underset{\sim}{v}, \lambda_0) \leq F(\underset{\sim}{w}, \lambda_0) \quad .$$

Thus $\min_{\underset{\sim}{x}} F(\underset{\sim}{x}, \lambda_0)$ must be attained at a feasible point. □

6. Accelerated penalty and barrier methods.

The problems of poor conditioning of the computational problem and (comparatively) slow convergence make it worthwhile to search for methods for accelerating the convergence of the penalty / barrier function algorithms. Consider the (generalized) penalty objective function

$$P(\underset{\sim}{x}, W, \underset{\sim}{\eta}) = f(\underset{\sim}{x}) + \sum_{i=1}^{q} W_{ii} \psi(h_i(\underset{\sim}{x}) + \eta_i) \tag{6.1}$$

where W is the diagonal matrix of penalty parameters, and the η_i are further parameters to be used in the acceleration process.

At a minimum of P, $\underset{\sim}{x}(W,\underset{\sim}{\eta})$ satisfies

$$\nabla f(\underset{\sim}{x}) - \sum_{i=1}^{q} u_i(W,\underset{\sim}{\eta})\nabla h_i(\underset{\sim}{x}) = 0 \tag{6.2}$$

where $u_i(W,\underset{\sim}{\eta}) = - W_{ii} \dfrac{\partial\psi}{\partial h_i}$. Provided $\|\underset{\sim}{x}(W,\underset{\sim}{\eta}) - \underset{\sim}{x}^*\|$ and $\|\underset{\sim}{u}(W,\underset{\sim}{\eta}) - \underset{\sim}{u}^*\|$ are sufficiently small, the second order sufficiency conditions hold at $\underset{\sim}{x}^*$, and $\nabla h_i(\underset{\sim}{x}^*)$, $i\epsilon I_2$, are linearly independent, then (by Theorem 5.3) $\underset{\sim}{x}^*$ also solves the ECP

$$\min_{\underset{\sim}{x}\epsilon S_{W,\underset{\sim}{\eta}}} f(\underset{\sim}{x}) \quad , \quad S_{W,\underset{\sim}{\eta}} = \{\underset{\sim}{x}\,;\, h_i(\underset{\sim}{x}) = h_i(\underset{\sim}{x}(W,\underset{\sim}{\eta}))\ ,\ i\epsilon I_2\} \ .$$

One sequential strategy for making $\underset{\sim}{x}(W,\underset{\sim}{\eta}) \to \underset{\sim}{x}^*$ is to force $\lambda = \min_i W_{ii} \uparrow \infty$. However, the parameter vector $\underset{\sim}{\eta}$ is also available, and we ask is it possible to adjust it to make

$$h_i(\underset{\sim}{x}(W,\underset{\sim}{\eta})) = 0 \quad , \quad i\epsilon I_2 \ . \tag{6.3}$$

Let $\dfrac{\partial x}{\partial \eta}$ be the matrix with components $\dfrac{\partial x_i}{\partial \eta_j}$, $i = 1,\ldots,n$, $j = 1,2,\ldots,q$. If $\nabla^2 P(\underset{\sim}{x}(W,\underset{\sim}{\eta}),W,\underset{\sim}{\eta})$ is nonsingular, then, by the implicit function theorem, we can solve (6.2) for $\underset{\sim}{x} = \underset{\sim}{x}(\underset{\sim}{\eta})$ holding W fixed. We have

$$\nabla^2 P \frac{\partial x}{\partial \eta} = - \sum_{i=1}^{q} W_{ii} \frac{\partial^2\psi}{\partial h_i \partial \eta_i} \nabla h_i^T \tag{6.4}$$

where all quantities are evaluated at $\underset{\sim}{x}(W,\underset{\sim}{\eta})$. Defining the diagonal

matrix V by $V_{ii} = W_{ii} \frac{\partial^2 \psi}{\partial h_i^2} = W_{ii} \frac{\partial^2 \psi}{\partial h_i \partial \eta_i}$, $i = 1,2,\ldots,q$, and the matrix B by $\kappa_i(B) = \nabla h_i^T$, $i = 1,2,\ldots,q$, then (6.4) can be written

$$(\nabla^2 \mathcal{L} + B V B^T) \frac{\partial x}{\partial \eta} = - BV \quad . \tag{6.5}$$

Choose V_0 to make $U = \nabla^2 \mathcal{L} + B V_0 B^T$ positive definite, and set $V_1 = V - V_0$. Then, by (3.6), if $\min_i V_{ii}$ is sufficiently large

$$\frac{\partial x}{\partial \eta} \sim - U^{-1} B (B^T U^{-1} B)^{-1} + O(V^{-1}) \quad . \tag{6.6}$$

This relation can be justified if

(i) the second order sufficiency conditions hold at $\underset{\sim}{x}^*$ and $\nabla h_i(\underset{\sim}{x})$, $i \in I_2$, are linearly independent,

(ii) $\|\underset{\sim}{x} - \underset{\sim}{x}^*\|$ and $\|\underset{\sim}{u} - \underset{\sim}{u}^*\|$ are sufficiently small,

(iii) $\min_i V_{ii} \to \infty$ as $\min_i W_{ii} \to \infty$ for $\underset{\sim}{\eta} = \underset{\sim}{\theta}$, and

(iv) $\min_i W_{ii}$ sufficiently large.

Consider now the use of Newton's method for solving (6.3). This suggests that a correction $\delta \underset{\sim}{\eta}$ to $\underset{\sim}{\eta}$ be found by solving

$$\nabla \underset{\sim}{h} \frac{\partial x}{\partial \eta} \delta \underset{\sim}{\eta} = B^T \frac{\partial x}{\partial \eta} \delta \underset{\sim}{\eta} = - \underset{\sim}{h} \quad . \tag{6.7}$$

But, by (6.6),

$$B^T \frac{\partial x}{\partial \eta} \sim - I + O(V^{-1}) \tag{6.8}$$

so that

$$\delta \underset{\sim}{\eta} = \underset{\sim}{h} + O(V^{-1}) \quad . \tag{6.9}$$

Thus we expect the simple correction $\delta\underset{\sim}{\eta} = \underset{\sim}{h}$ to approximate arbitrarily closely to a second order process provided V is sufficiently large.

<u>Algorithm</u> (ECP)

(i) Initialize $\underset{\sim}{\eta}^{(1)}$, $W^{(1)}$.

(ii) Minimize $P(\underset{\sim}{x}, W^{(k)}, \underset{\sim}{\eta}^{(k)})$ to determine $\underset{\sim}{x}_k$, $\underset{\sim}{u}_k$.

(iii) IF $\sum_{i=1}^{q} u_i^k \, h_i(\underset{\sim}{x}_k) < \text{TOL}$ THEN STOP.

(iv) FOR I = 1 STEP 1 UNTIL Q DO

IF $\text{ABS}(h_i(\underset{\sim}{x}_k)) < \text{DECR} * \text{ABS}(h_i(\underset{\sim}{x}_{k-1}))$

THEN $\eta_i^{k+1} = \eta_i^k + h_i(\underset{\sim}{x}_k)$

ELSE $W_{ii}^{(k+1)} = \text{EXP} * W_{ii}^{(k)}$

$$\eta_i^{k+1} = \left(\frac{\partial\psi}{\partial h_i}\right)^{-1}\left(-\frac{u_i^k}{W_{ii}^{(k+1)}}\right) .$$

(v) K := K+1.

(vi) GO TO (ii).

<u>Remark</u>: The idea behind the algorithm is that the correction (6.9) is used whenever the convergence of h_i to zero is satisfactory. Otherwise it is assumed that W_{ii} is too small and it is increased accordingly. η_i is modified at the same time to ensure that $u_i^k \to u_i^*$ as $h_i \to 0$. $(\frac{\partial\psi}{\partial h})^{-1}$ indicates the inverse function to $\frac{\partial\psi}{\partial h}$.

For the ICP we consider the modified barrier function

$$R(\underset{\sim}{x}, W^{(k)}, \underset{\sim}{\eta}^{(k)}) = f(\underset{\sim}{x}) + \sum_{i=1}^{m} W_{ii}^{(k)} \, u_i^{k-1} \, \phi(g_i(\underset{\sim}{x}) + \eta_i^k) \qquad (6.10)$$

where $\underset{\sim}{u}_{k-1}$ is the vector of multiplier estimates from the previous minimization, and $W^{(k)}$ is now the diagonal matrix of barrier parameters. We note that, in the particular case in which all constraints are active, the previous analysis is applicable, at least formally, and suggests a correction

$$\delta\underset{\sim}{\eta}^{(k)} = \underset{\sim}{\eta}^{(k)} + \underset{\sim}{g}(\underset{\sim}{x}_k) \tag{6.11}$$

with order of magnitude departure from a second order iteration of $O\left(\left(\min_i W_{ii}^{(k)} u_i^{k-1} \frac{\partial^2 \phi}{\partial g_i^2}\right)^{-1}\right)$. However, we require automatic selection of the active constraints if we are to make use of this result, and it is important to note that this is provided naturally in the algorithm by the options

(i) if $g_i \to 0$ at a satisfactory rate then $\eta_i^{k+1} = \eta_i^k + g_i$, and

(ii) if the convergence rate is too slow, then decrease the barrier parameter.

This second option can be expected to apply to the inactive constraints, and will drive the contribution to (6.4) from this source rapidly to zero by (4.33). Note that the boundedness of the barrier terms requires that $g_i + \eta_i$ be positive. If $\underset{\sim}{\eta}^{(1)}$ is set to zero then (6.11) ensures that this condition will be met initially. Provided strict complementarity holds, the convergence of the multiplier estimates will ensure that it must hold ultimately. Of course, the calculation must be started from a feasible point.

Algorithm (ICP)

(i) Initialize $\underset{\sim}{\eta}^{(1)}$, $W^{(1)}$, $\underset{\sim}{u}_0$.

(ii) Minimize $R(\underset{\sim}{x}, W^{(k)}, \underset{\sim}{\eta}^{(k)})$ to determine $\underset{\sim}{x}_k$, $\underset{\sim}{u}_k$.

(iii) IF $\sum_{i=1}^{m} u_i^k \, g_i(\underset{\sim}{x}_k) < \text{TOL}$ THEN STOP.

(iv) FOR I = 1 STEP 1 UNTIL M DO

IF $\text{ABS}(g_i(\underset{\sim}{x}_k)) < \text{DECR} * \text{ABS}(g_i(\underset{\sim}{x}_{k-1}))$

THEN $W_{ii}^{(k+1)} = -1 / \dfrac{\partial \phi}{\partial g_i}$,

$$\eta_i^{(k+1)} = \eta_i^{(k)} + g_i(\underset{\sim}{x}_k)$$

ELSE $W_{ii}^{(k+1)} = \text{DECR} * W_{ii}^{k}$,

$$\eta_i^{(k+1)} = \left(\frac{\partial \phi}{\partial g_i}\right)^{-1} \left(-\frac{1}{W_{ii}^{(k+1)}}\right) .$$

(v) K := K+1.

(vi) GO TO (ii).

Remark: As in the previous algorithm $(\frac{\partial \phi}{\partial g})^{-1}$ denotes the inverse function to $\frac{\partial \phi}{\partial g}$. For example, if $\phi = -\log g$ then $\eta = W$.

Consider now another modified penalty function for the ECP

$$S(\underset{\sim}{x}) = f(\underset{\sim}{x}) - \underset{\sim}{u}(\underset{\sim}{x})^T \underset{\sim}{h}(\underset{\sim}{x}) + \underset{\sim}{h}(\underset{\sim}{x})^T W \underset{\sim}{h}(\underset{\sim}{x}) \tag{6.12}$$

where the matrix W is positive definite.

Lemma 6.1: If the second order sufficiency conditions hold for $\underset{\sim}{x} = \underset{\sim}{x}^*$, and the $\nabla h_i(\underset{\sim}{x}^*)$, $i \in I_2$, are linearly independent then $S(\underset{\sim}{x})$ has a

local minimum at $\underset{\sim}{x} = \underset{\sim}{x}^*$ provided $\underset{\sim}{u}(\underset{\sim}{x}) \to \underset{\sim}{u}^*$ as $\underset{\sim}{x} \to \underset{\sim}{x}^*$ and the smallest eigenvalue of W is large enough.

<u>Proof</u>: We have

$$\begin{aligned}\nabla S(\underset{\sim}{x}^*) &= \nabla f(\underset{\sim}{x}^*) - \underset{\sim}{u}(\underset{\sim}{x}^*)^T \nabla h(\underset{\sim}{x}^*) \\ &\quad - \underset{\sim}{h}(\underset{\sim}{x}^*)^T(\nabla \underset{\sim}{u}(\underset{\sim}{x}^*) - 2W\nabla h(\underset{\sim}{x}^*)) \\ &= \nabla f(\underset{\sim}{x}^*) - \underset{\sim}{u}^{*T}\nabla \underset{\sim}{h}(\underset{\sim}{x}^*) \\ &= 0\end{aligned}$$

as $\underset{\sim}{u}^*$ is the vector of Lagrange multipliers for the ECP. Thus S has a stationary point for $\underset{\sim}{x} = \underset{\sim}{x}^*$. Now

$$\begin{aligned}\nabla^2 S(\underset{\sim}{x}^*) &= \nabla^2 \mathcal{L}(\underset{\sim}{x}^*, \underset{\sim}{u}^*) - \nabla \underset{\sim}{u}(\underset{\sim}{x}^*)^T \nabla h(\underset{\sim}{x}^*) \\ &\quad - \nabla h(\underset{\sim}{x}^*)^T \nabla \underset{\sim}{u}(\underset{\sim}{x}^*) \\ &\quad + 2\nabla h(\underset{\sim}{x}^*)^T W \nabla h(\underset{\sim}{x}^*) \qquad (6.13)\end{aligned}$$

where terms which vanish at $\underset{\sim}{x}^*$ have been ignored. Corollary 3.2 can now be applied to show $\nabla^2 S(\underset{\sim}{x}^*)$ is positive definite. We set $V = \nabla \underset{\sim}{h}(\underset{\sim}{x}^*)^T$, $U = \nabla^2 \mathcal{L}(\underset{\sim}{x}^*, \underset{\sim}{u}^*) - \nabla \underset{\sim}{u}(\underset{\sim}{x}^*)^T \nabla \underset{\sim}{h}(\underset{\sim}{x}^*) - \nabla \underset{\sim}{h}(\underset{\sim}{x}^*)^T \nabla \underset{\sim}{u}(\underset{\sim}{x}^*)$, and note that

$$\min_{V^T\underset{\sim}{t}=0,\ \|\underset{\sim}{t}\|=1} \underset{\sim}{t}^T U \underset{\sim}{t} = \min_{V^T\underset{\sim}{t}=0,\ \|\underset{\sim}{t}\|=1} \underset{\sim}{t}^T \nabla^2 \mathcal{L}(\underset{\sim}{x}^*, \underset{\sim}{u}^*)\underset{\sim}{t} = m > 0$$

as the second order sufficiency conditions hold at $\underset{\sim}{x}^*$. □

Theorem 6.1: Let the conditions of Lemma 6.1 hold at x^* and set

$$\underset{\sim}{u}(\underset{\sim}{x}) = B(\underset{\sim}{x})^+ \nabla f(\underset{\sim}{x})^T \tag{6.14}$$

where $B(\underset{\sim}{x}) = \nabla \underset{\sim}{h}(\underset{\sim}{x})^T$. Then (6.12) has a local minimum at $\underset{\sim}{x}^*$ provided the smallest eigenvalue of W is large enough.

Proof: This result is an immediate consequence of Lemma 6.1. As the $\nabla h_i(\underset{\sim}{x}^*)$, $i \in I_2$, are linearly independent, $B(\underset{\sim}{x}^*)^+$ is a bounded operator for $\|\underset{\sim}{x} - \underset{\sim}{x}^*\|$ small enough. Thus $\underset{\sim}{u}(\underset{\sim}{x}) \to \underset{\sim}{u}^*$ as $\underset{\sim}{x} \to \underset{\sim}{x}^*$. □

Remark: (i) By using (6.14) we can construct a penalty function which is differentiable in a neighborhood of $\underset{\sim}{x}^*$ (contrast with (5.13)) and which has a local minimum at $\underset{\sim}{x} = \underset{\sim}{x}^*$ for sufficiently large but finite values of the penalty parameter. However, (6.14) requires first derivatives of the problem functions so that minimization of (6.12) with a method that requires first derivatives of S will require second derivatives of the problem functions. Two cases have been considered (Fletcher).

$$\text{(i)} \quad S(\underset{\sim}{x}) = f(\underset{\sim}{x}) - \underset{\sim}{h}(\underset{\sim}{x})^T B(\underset{\sim}{x})^+ \nabla f(\underset{\sim}{x})^T + \sigma \|\underset{\sim}{h}(\underset{\sim}{x})\|^2 \quad , \text{ and} \tag{6.15}$$

$$\text{(ii)} \quad S(\underset{\sim}{x}) = f(\underset{\sim}{x}) - \underset{\sim}{h}(\underset{\sim}{x})^T B(\underset{\sim}{x})^+ \nabla f(\underset{\sim}{x})^T + \sigma \|(B(\underset{\sim}{x})^+)^T \underset{\sim}{h}(\underset{\sim}{x})\|^2 \quad , \tag{6.16}$$

where σ is a penalty parameter.

(ii) There is a close connection between the penalty function (6.15) and the algorithm based on (6.1) in the case $\psi(h) = h^2$. At a minimum of P we have (as $\nabla P = 0$)

$$2W(\underset{\sim}{h} + \underset{\sim}{\theta}) = - B^+(\nabla f)^T \quad . \tag{6.17}$$

Thus the correction formula corresponds to updating the Lagrange multiplier estimate by (6.14) at the end of each unconstrained minimization rather than continuously which the use of S requires.

(iii) Note that $S(\underset{\sim}{x})$ can be interpreted as a Lagrangian. For example, in the case $S(\underset{\sim}{x})$ is given by (6.16),

$$S(\underset{\sim}{x}) = \mathcal{L}(\underset{\sim}{x},\underset{\sim}{w}(\underset{\sim}{x})) \tag{6.18}$$

where

$$\underset{\sim}{w}(\underset{\sim}{x}) = B(\underset{\sim}{x})^{+}\nabla f(\underset{\sim}{x})^{T} - \sigma B(\underset{\sim}{x})^{+}(B(\underset{\sim}{x})^{+})^{T}\underset{\sim}{h}(\underset{\sim}{x}) \quad . \tag{6.19}$$

<u>Lemma 6.2</u>: $\underset{\sim}{w}(\underset{\sim}{t})$ defined by (6.19) is the vector of Lagrange multipliers for the problem

$$\underset{\underset{\sim}{x}}{\text{minimize}}\ f(\underset{\sim}{t}) + \nabla f(\underset{\sim}{t})(\underset{\sim}{x}-\underset{\sim}{t}) + \frac{\sigma}{2}\|\underset{\sim}{x}-\underset{\sim}{t}\|^{2} \tag{6.20}$$

subject to the linear constraints

$$\underset{\sim}{h}(\underset{\sim}{t}) + \nabla\underset{\sim}{h}(\underset{\sim}{t})(\underset{\sim}{x}-\underset{\sim}{t}) = 0 \quad , \tag{6.21}$$

provided this minimum exists.

<u>Proof</u>: Any point satisfying the constraints (6.21) has the form

$$\underset{\sim}{x} = \underset{\sim}{t} - (B(\underset{\sim}{t})^{T})^{+}\underset{\sim}{h}(\underset{\sim}{t}) + A(\underset{\sim}{t})\underset{\sim}{z} \tag{6.22}$$

where $B(\underset{\sim}{t})^{T}A(\underset{\sim}{t}) = 0$. The multiplier relation for (6.20), (6.21) is

$$\nabla f(\underset{\sim}{t}) + \sigma(\underset{\sim}{x}-\underset{\sim}{t})^{T} = \underset{\sim}{u}^{T}B(\underset{\sim}{t})^{T} \tag{6.23}$$

so that $\underset{\sim}{u}$ can be taken as (substituting (6.22) into (6.23))

$$\underset{\sim}{u} = B(\underset{\sim}{t})^{+}\{\nabla f(\underset{\sim}{t})^{T} - \sigma(B(\underset{\sim}{t})^{T})^{+}\underset{\sim}{h}(\underset{\sim}{t})\} \quad . \quad \square$$

If (6.22) is substituted into (6.20) the problem becomes one of minimizing w.r.t. $\underset{\sim}{z}$

$$\nabla f A\underset{\sim}{z} + \frac{\sigma}{2} \underset{\sim}{z}^T A^T A \underset{\sim}{z}$$

whence

$$\underset{\sim}{z} = -\frac{1}{\sigma} (A^T A)^{-1} A^T \nabla f^T \quad . \tag{6.24}$$

Thus σ plays a role in ensuring that $\underset{\sim}{x}(\underset{\sim}{t})$, the minimum of (6.20), cannot deviate far from $\underset{\sim}{t}$ (cf. remark (ii) following MP Corollary 5.1).

<u>Example</u>: (i) The Lagrangian interpretation provides a method for generalizing the above discussion to inequality constraints. Consider the problem $\min_{\underset{\sim}{x}} \mathcal{L}(\underset{\sim}{x},\underset{\sim}{w}(\underset{\sim}{x}))$ where $\underset{\sim}{w}(\underset{\sim}{t})$ is the vector of multipliers for the problem

$$\min_{\underset{\sim}{x}} f(\underset{\sim}{t}) + \nabla f(\underset{\sim}{t})(\underset{\sim}{x}-\underset{\sim}{t}) + \frac{\sigma}{2} \|\underset{\sim}{x}-\underset{\sim}{t}\|^2$$

$$\text{subject to } \underset{\sim}{g}(\underset{\sim}{t}) + \nabla \underset{\sim}{g}(\underset{\sim}{t})(\underset{\sim}{x}-\underset{\sim}{t}) \geq 0 \quad .$$

Under what conditions does $\mathcal{L}$ have an unconstrained minimum at $\underset{\sim}{x}^*$. What role does strict complementarity play in this problem?

(ii) (6.12) can be generalized to other penalty functions and to barrier functions (cf. Remark (ii) above). How much of the above analysis goes through? What modifications are required? Evaluate the resulting algorithms.

Notes

1., 2. See Fiacco and McCormick's book. Also the paper 'Penalty function methods for mathematical programming problems', J. Math. Anal. and Applic. (1970), by Osborne and Ryan.

3. Fiacco and McCormick were the first to draw attention to the importance of these (as they were to much of the material in this section).

4. The log family is due to Osborne and Ryan. The importance of the conditioning of the Hessian to Walter Murray. Rate of convergence formulae have also been developed by F. A. Lootsma, (Thesis, also survey paper at Dundee conference).

5. Fiacco and McCormick. The exact penalty function is due to Zangwill.

6. The algorithm for the ECP is due to Powell in the case $\psi = h^2$ (Harwell report, also Procedings of Keele Conference). The exact penalty function $S(\underset{\sim}{x})$ is due to Fletcher who has developed it together with his student Shirley Lill and described it in several Harwell reports. The extension to inequality constraints (example (i)) is also due to Fletcher.

References

A. V. Fiacco and G. P. McCormick (1968): Nonlinear Programming: Sequential Unconstrained Minimization Techniques, Wiley.

R. Fletcher and S. A. Lill (1970): A class of methods for non-linear programming, in Proceedings of Nonlinear Programming Symposium, Madison.

F. A. Lootsma (1970): Boundary properties of penalty functions for constrained minimization, Phillips Res. Repts. Suppl., No. 3.

W. Murray (1969): Constrained Optimization, Nat. Phys. Lab. Rep. Ma79.

M. R. Osborne and D. M. Ryan (1970): On penalty functions for non-linear programming problems, J. Math. Anal. Appl., 31, pp. 559-578.

M. J. Powell (1969): A Method for non-linear constraints in minimization problems, in Optimization (editor: R. Fletcher), Academic Press.

Bibliography

Abadie, J. (editor), (1967) Nonlinear Programming, (1970) Integer and Nonlinear Programming, North Holland.

Ablow, C. M. and Brigham, G. (1955). An analog solution of programming problems. Op. Res., 3, pp 388-394.

Akaike, H. (1959). On a successive transformation of probability distribution and its application to the analysis of the optimal gradient method. Ann. Inst. Stat. Math., Tokyo, 11, p 1.

Allran, R. R. and Johnsen, S. E. J. (1970). An algorithm for solving nonlinear programming problems subject to nonlinear inequality constraints. Comp. J., 13(2), pp 171-177.

Arrow, K. J. and Hurwicz, L. (1956). Reduction of constrained maxima to saddle point problems; in "Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability", ed. J. Neyman. University of California Press, pp 1-20.

Arrow, K. J., Hurwicz, L. and Uzawa, H. (1958). "Studies in Linear and Nonlinear Programming". Stanford University Press.

Arrow, K. J., Hurwicz, L. and Uzawa, H. (1961). Constraint qualifications in maximization problems. Nav. Res. Log. Q., 8, pp 175-191.

Bard, Y. (1968). On a numerical instability of Davidon-like methods. Maths. Comp., 22 (103), pp 665-666.

Bard, Y. (1970). Comparison of gradient methods for the solution of nonlinear parameter estimation problems. SIAM J. Numer. Anal., 7 (1), pp 157-186.

Bartels, R. H., Golub, G. H., and Saunders, M. A. (1970), Numerical techniques in Mathematical Programming, in Nonlinear Programming, Academic Press.

Bazaraa, M. S. Goode, J. J., Nashed, M. Z., and Shetty, C. M. (1971). Nonlinear programming without differentiability in Banach spaces: Necessary and sufficient constraint qualification. Applicable Analysis, to appear.

Beale, E. M. L. (1955). On minimizing a convex function subject to linear inequalities. J. Roy. Stat. Soc., Ser. B, 17 (2), pp 173-184.

Beale, E. M. L. (1959). Numerical methods; in "Nonlinear Programming", ed. J. Adadie. North-Holland Publishing Co., p 150.

Bellmore, M., Greenberg, H. J. and Jarvis, J. J. (1970). Generalized penalty function concepts in mathematical optimization. Op. Res., 18 (2), pp 229-252.

Beltrami, E. J. (1969). A constructive proof of the Kuhn-Tucker multiplier rule. J. Math. Anal. Appl., 26, pp 297-306.

Beltrami, E. J. (1970). "An Algorithmic Approach to Nonlinear Analysis and Optimization". Academic Press.

Beltrami, E. J. and McGill, R. (1966). A class of variational problems in search theory and the maximum principle. Op. Res., 14 (2), pp 267-278.

Box, M. J. (1966). A comparison of several current optimization methods and the use of transformations in constrained problems. Comp. J., 9, pp 67-77.

Box, M. J., Davies, D. and Swann, W. H. (1969). "Nonlinear Optimization Techniques". ICI Monograph, Oliver and Boyd.

Bracken, J. and McCormick, G. P. (1968). "Selected Applications of Nonlinear Programming". Wiley.

Broyden, C. G. (1965). A class of methods for solving nonlinear simultaneous equations. Math. Comp., 19 (92), pp 577-593.

Broyden, C. G. (1967). Quasi-Newton methods and their application to function minimization. Math. Comp., 21 (19), pp 368-381.

Broyden, C. G. (1970a). The convergence of a class of double-rank minimization algorithms; 1. general considerations. J. Inst. Math. Appl., 6 (1) pp 76-90.

Broyden, C. G. (1970b). The convergence of single-rank quasi-Newton methods. Math. Comp., 24 (110), pp 365-382.

Bui Trong Lieu and Huard, P. (1966). La méthode des centres dans un espace topologique. Numér. Math., 8 (1), pp 56-67.

Butler, T. and Martin, A. V. (1962). On a method of Courant for minimizing functionals. J. Math. Phys., 41, pp 291-299.

Camp. G. D. (1955). Inequality-constrained stationary value problems. J. Op. Res. Soc. Am., 3, pp 548-550.

Carrol, C. W. (1961). The created response surface technique for optimizing nonlinear restrained systems. Op. Res., 9 (2), pp 169-184.

Charnes, A. and Lemke, C. E. (1954). Minimization of nonlinear separable convex functionals. Nav. Res. Log. Q., 1, pp 301-302.

Cheney, E. W. and Goldstein, A. A. (1959). Newton's method for convex programming and Tchebycheff approximation. Numer. Math., 1, pp 254-268.

Colville, A. R. (1968). A comparative study on nonlinear programming codes. IBM New York Scientific Centre Tech. Rep. No. 320-2949.

Courant, R. (1943). Variational methods for the solution of problems of equilibrium and vibrations. Bull. Am. Math. Soc., 49, pp 1-23.

Crockett, J. B. and Chernoff, H. (1955). Gradient methods of maximization. Pac. J. Math., 5, pp 33-50.

Curry, H. B. (1944). The method of steepest descent for non-linear minimization problems. Quart. Appl. Math., 2 (3), pp 250-261.

Daniel, J. W. (1967a). The conjugate gradient method for linear and nonlinear operator equations. SIAM J. Numer. Anal., 4, pp 10-26.

Daniel, J. W. (1967b). Convergence of the conjugate gradient method with computationally convenient modifications. Numer. Math., 10, pp 125-131.

Dantzig, G. B. (1951). Maximization of a linear function of variables subject to linear inequalities: in "Activity Analysis of Production and Allocation", ed. T. C. Koopmans. Cowles Commission Monograph No. 13, Wiley.

Dantzig, G. B. (1963). Variable metric method for minimization. U.S.A.E.C. Research and Development Rep. ANL-5990, Argonne National Laboratory.

Davidon, W. C. (1968). Variance algorithm for minimization. Comp. J., 10, pp 406-410.

Dorn, W. S. (1960). Duality in quadratic programming. Quart. Appl. Math., 18, pp 155-162.

Edelbaum, T. N. (1962). Theory of maxima and minima: in "Optimization Techniques", ed. G. Leitmann. Academic Press.

El-Hodiri, Mohamed A. (1971). Constrained Extrema, Lecture Notes in Operations Research and Mathematical Systems 56, Springer-Verlag.

Evans, J. P. and Gould, F. J. (1970). Stability in non-linear programming. Op. Res., 18 (1), pp 107-118.

Faure, P. and Huard, P. (1966). Résultats nouveaux relatifs á la méthode des centres. Quatriéme Conférence de Recherche Opérationelle, Cambridge, Mass.

Fiacco, A. V. (1967). Sequential unconstrained minimization methods for nonlinear programming. Ph.D. Dissertation, Northwestern University, Ill.

Fiacco, A. V. (1970). Penalty methods for mathematical programming in E^n with general constraint sets. J. Opt. Th. Appl., 6 (3), pp 252-268.

Fiacco, A. V. and McCormick, G. P. (1963). Programming under non-linear constraints by unconstrained minimization: a primal-dual method. Tech. Paper RAC-TP-96, Research Analysis Corporation.

Fiacco, A. V. and McCormick, G. P. (1964a). The sequential unconstrained minimization technique for nonlinear programming: a primal-dual method. Man. Sci., 10 (2), pp 360-366.

Fiacco, A. V. and McCormick, G. P. (1964b). Computational algorithm for the sequential unconstrained minimization technique for nonlinear programming. Man. Sci., 10 (2), pp 601-617.

Fiacco, A. V. and McCormick, G. P. (1966). Extensions of SUMT for nonlinear programming: equality constraints and extrapolation. Man. Sci., 12 (11), pp 816-829.

Fiacco, A. V. and McCormick, G. P. (1967). The slacked unconstrained minimization technique for convex programming. SIAM J. Appo. Math., 15 (3), pp 505-515.

Fiacco, A. V. and McCormick, G. P. (1968). "Nonlinear Programming: Sequential Unconstrained Minimization Techniques". Wiley.

Fletcher, R. (1965). Function minimization without evaluating derivatives - a review. Comp. J., 8, pp 33-41.

Fletcher, R. (1968). Programming under linear equality and inequality constraints. ICI Rep. No. MSDH/68/19.

Fletcher, R. (Ed.) (1969a). "Optimization". Academic Press.

Fletcher, R. (1969b). A review of methods for unconstrained optimization: in "Optimization", ed. R. Fletcher. Academic Press.

Fletcher, R. (1969c). A new approach to variable metric algorithms. AERE Tech. Rep. No. 383.

Fletcher, R. (1970). A general quadratic programming algorithm. AERE Tech. Rep. No. 401.

Fletcher, R. (1970). A new approach to variable metric algorithms, Computer J., 13, pp 317-322.

Fletcher, R. and McCann, A. P. (1969). Acceleration techniques for nonlinear programming: in "Optimization", ed. R. Fletcher. Academic Press.

Fletcher, R. and Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. Comp. J., 6, pp 163-168.

Fletcher, R. and Reeves, C. M. (1964). Function minimization by conjugate gradients. Comp. J., 7, pp 149- 154.

Forsythe, G. E. (1967). On the asymptotic directions of the s-dimensional optimum gradient method. Tech. Rep. No. CS61, Stanford University.

Forsythe, G. E. and Motzkin, T. S. (1951). Asymptotic properties of the optimum gradient method (abstract). Bull. Am. Math. Soc., 57, p 183.

Frisch, K. R. (1955). The logarithmic potential method of convex programming. Memorandum May 13, 1955, University Institute of Economics, Oslo.

Geoffrion, A. M. (1966). Strictly concave parametric programming, part 1: basic theory. Man. Sci., 13, pp 244-253.

Geoffrion, A. M. (1967). Strictly concave parametric programming, part 2: additional theory and computational considerations. Man. Sci., 13 (5), pp 359-370.

Gill, P. E. and Murray, W. (1970). Stable implementation of unconstrained optimization algorithms, NPL report Maths. 97.

Goldfarb, D. (1969a). Sufficient conditions for the convergence of a variable metric algorithm: in "Optimization", ed. R. Fletcher. Academic Press.

Goldfarb, D. (1969b). Extension of Davidon's variable metric method to maximization under linear inequality and equality constraints. SIAM J. Appl. Math., 17, pp 739-764.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. Math. Comp., 24, pp 23-26.

Goldfarb, D. and Lapidus, L. (1968). Conjugate gradient method for nonlinear programming problems with linear constraints. I. and E. C. Fundamentals, 7 (1), pp 142-151.

Goldstein, A. A. (1962). Cauchy's method of iminmization. Numer. Math., 4, pp 146-150.

Goldstein, A. A. (1965). On steepest descent, SIAM J. Control, 3, pp 147-151.

Goldstein, A. A. (1968). Constructive Real Analysis, Harper and Row.

Goldstein, A. A. and Price, J. (1967). An effective algorithm for minimization, Num. Math., 10, pp. 184-189.

Golub, G. H., and Saunders, M. A. (1969). Numerical methods for solving linear least squares problems, Computer Science Dept. Tech. Rep. CS134, Stanford University.

Gould, F. J. and Tolle, J. W. (1971). A necessary and sufficient constraint qualification, SIAM J. Appl. Maths. 20, pp 164-172.

Greenstadt, J. (1967). On the relative efficiencies of gradient methods. Math. Comp., 21, pp 360-367.

Greenstadt, J. (1970). Variations on variable-metric methods. Math. Comp., 24, pp 1-25.

Haarhoff, P. C. and Buys, J. D. (1970). A new method for the optimization of a nonlinear function subject to nonlinear constraints. Comp. J., 13 (2), pp 178-184.

Hadley, G. (1964). "Nonlinear and Dynamic Programming". Addison-Wesley.

Hartley, H. O. and Hocking, R. R. (1963). Convex programming by tangential approximation. Man. Sci., 9, pp 600-612.

Hestenes, M. R. (1966). "Calculus of Variations and Optimal Control Theory". Wiley.

Huang, H. Y. (1969). Unified approach to quadratically convergent algorithms for function minimization. J.O.T.A., 5, pp 405-425.

Huang, H. Y. and Levy, A. V. (1969). Numerical experiments on quadratically convergent algorithms for function minimization. Aero-Astronautics Rep. No. 66, Rice University.

Huard, P. (1964). Résolution de programmes mathématiques á constraintes non-linéaires par la méthode des centres. Note E.D.F., HR5690/317.

Huard, P. (1967). Resolution of mathematical programming with nonlinear constraints by the method of centres; in "Nonlinear Programming", ed. J. Adadie. North-Holland Pub. Co., Amsterdam.

Huard, P. (1968). Programmation mathématique convex. R.I.R.O., 7, pp 43-59.

John, F. (1948). Extremum problems with inequalities as subsidiary conditions: in "Studies and Essays". Courant Anniversary Volume, Interscience, pp 187-204.

Kelley, H. J. (1962). Methods of gradients: in "Optimization Techniques", ed. G. Leitmann. Academic Press.

Kelley, J. E., Jr. (1960). The cutting plane method for solving convex programs. J. Soc. Ind. Appl. Math., 8 (4), pp 703-712.

Kowalik, J. (1966). Nonlinear programming procedures and design optimization. Acta Polytechnica Scandinavica, 13, Trondheim.

Kowalik, J. and Osborne, M. R. (1968). "Methods for Unconstrained Problems". American Elsevier.

Kowalik, J., Osborne, M. R. and Ryan, D. M. (1969). A new method for constrained optimization problems. Op. Res., 17 (6), pp 973-983.

Kuhn, H. W. and Tucker, A. W. (1951). Non-linear Programming: in "Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability", ed. J. Neyman. University of California Press, pp 481-493.

Kunzi, H. P. and Oettli, W. (1969). Nichtlineare Optimierung, Lecture Notes in Operations Research and Mathematical Systems 16, Springer-Verlag.

Lemke, C. E. (1962). A method for solution of quadratic programming problems. Man. Sci., 8, pp 442-453.

Lootsma, F. A. (1967). Logarithmic programming: a method of solving nonlinear-programming problems. Philips Res. Rep., 22 (3), pp 329-344.

Lootsma, F. A. (1968a). Extrapolation in logarithmic programming. Philips Res. Res. Rep., 23, pp 108-116.

Lootsma, F. A. (1968b). Constrained optimization via penalty functions. Philips Res. Rep., 23, pp 408-423.

Lootsma, F. A. (1969). Hessian matrices of penalty functions for colving constrained-optimization problems. Philips Res. Resp., 24, pp 322-331.

Lootsma, F. A. (1970). Boundary properties of penalty functions for constrained minimization. Philips Res. Resp. Supp. No. 3.

Lootsma, F. A. (editor), (1972). Numerical Methods for Nonlinear Optimization, Academic Press.

Luenburger, D. G. (1969). "Optimization by Vector Space Methods". Wiley.

Mangasarian, O. L. (1969). Nonlinear Programming. McGraw-Hill.

Mangasarian, O. L. and Fromowitz, S. (1967). The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. J. Math. Anal. Appl., 17 (1), pp 37-47.

McCormick, G. P. (1969). The rate of convergence of the reset Davidon variable metric method. MRC Tech. Rep. No. 1012, University of Wisconsin.

McCormick, G. P. and Pearson, J. D. (1969). Variable metric methods and unconstrained optimization: in "Optimization", ed. R. Fletcher. Academic Press.

Meyers, G. E. (1968). Properties of the conjugate-gradient and Davidon methods. J. Opt. Th. Appl., 2 (4), pp 209-219.

Morrison, D. D. (1968). Optimization by least squares. SIAM J. Numer. Anal., 5 (1), pp 83-88.

Murray, W. (1969a). Indefinite quadratic programming. Nat. Phys. Lab. Rep. Ma 76.

Murray, W. (1969b). Behaviour of Hessian matrices of barrier and penalty functions arising in optimization. Nat. Phys. Lab. Rep. Ma 77.

Murray. W. (1969c). Constrained Optimization. Nat. Phys, Lab. Rep. Ma 79.

Murtagh, B. A. and Sargent, R. W. H. (1969). A constrained minimization method with quadratic convergence; in "Optimization, ed. R. Fletcher, Academic Press.

Murtagh, B. A. and Sargent, R. W. H. (1970). Computational experience with quadratically convergent minimization methods. Comp. J., 13 (2), pp 185-194.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. Comp. J., 7, pp 308-313.

Osborne, M. R. and Ryan, D. M. (1970a). On penalty function methods for nonlinear programming problems. J. Math. Anal. Appl., 31 (3), pp 559-578.

Osborne, M. R. and Ryan, D. M. (1970b). An algorithm for nonlinear programming. Tech. Rep. No. 35, Computer Centre, A.N.U.

Osborne, M. R. and Ryan, D. M. (1972). A hybrid method for nonlinear programming, in Numerical Methods for Nonlinear Optimization (ed. F. A. Lootsma), Academic Press.

Ostrowski, A. M. (1966). Solution of Equations and Systems of Wquations (2nd Edition), Academic Press.

Parisot, G. R. (1961). Résolution numérique approchée du probléme de programmation linéaire par application de la programmation logarithmique. Rev. Fr. Recherche Opérationelle, 20, pp 227-259.

Pearson, J. D. (1969). On variable metric methods of minimization. Comp. J., 12, pp 171-178.

Pietrzykowski, T. (1962). Application of the steepest descent method to concave programming; in "Proceedings of IFIPS Congress", Munich, North-Holland Pub. Co., pp 185-189.

Polak, E. and Ribiére, G. (1969). Note sur la convergence de méthodes de directions conjuguées. R.I.R.O., 3, pp 35-43.

Pomentale, T. (1965). A new method for solving conditioned maxima problems. J. Math. Anal. Appl., 10, pp 216-220.

Powell, M. J. D. (1967). A method for nonlinear constraints in minimization problems. AERE Tech. Rep. No. 310.

Powell, M. J. D. (1968). A survey of numerical methods for unconstrained optimization. AERE Tech. Rep. No. 340.

Powell, M. J. D. (1969a). Rank one methods for unconstrained optimization. AERE Tech. Rep. No. 372.

Powell, M. J. D. (1969b). On the convergence of the variable metric algorithm. AERE Tech. Rep. No. 382.

Powell, M. J. D. (1970). A new algorithm for unconstrained optimization. AERE Tech. Rep. No. 393.

Rosen, E. M. (1966). A review of quasi-Newton methods in nonlinear equation solving and unconstrained optimization; in "Proceedings 21st National Conference of the A.C.M.". Thompson Book Co., pp 37-41.

Rosen, J. B. (1960). The gradient projection method for nonlinear programming, part 1: linear constraints. J. Soc. Ind. Appl. Math., 8 (1), pp 181-217.

Rosen, J. B. and Suzuki, S. (1965). Construction of nonlinear programming test problems. Comm. A.C.M., 8, p 113.

Rosenbrock, H. H. (1960). An automatic method for finding the greatest or least value of a function. Comp. J., 3, pp 175-184.

Rudin, W. (1953). "Principles of Mathematical Analysis". McGraw-Hill, N. Y.

Schmit, L. A. and Fox R. L. (1965). Integrated approach to structural synthesis, AIAA J., 3 (6), pp 1104-1112.

Shah, B. V., Buehler, R. J. and Kempthorne, O. (1964). Some algorithms for minimizing a function of several variables. J. Soc. Ind. Appl. Math., 12, pp 74-92.

Spang, H. A. (1962). A review of minimization techniques for nonlinear functions. SIAM Review, 4 (4), pp 343-365.

Stoer, J. (1971). On the numerical solution of constrained least squares problems, SIAM J. Numerical Analysis, 8, pp 382-411.

Strong, R. E. (1965). A note on the sequential unconstrained minimization technique for non-linear programming. Man. Sci., 12 (1), pp 142-144.

Tremoliéres, R. (1968). La méthode des centre á troncature variable. Thése, Paris.

Whittle, P. (1971). "Optimization under Constraints", Wiley.

Wolfe, P. (1959). The simplex method for quadratic programming. Econometrica, 27 (3), pp 382-398.

Wolfe, P. (1961). A duality theorem for nonlinear programming. Quart. Appl. Math., 19 (3), pp 239-244.

Zangwill, W. I. (1967). Nonlinear programming via penalty functions. Man. Sci., 13 (5), pp 344-358.

Zangwill, W. I. (1969). "Nonlinear Programming: A Unified Approach". Prentice-Hall.

Zoutendijk, G. (1960). "Methods of Feasible Directions". Elsevier.