NGL-34-002-032

ON THE EXTENSION OF THE DAVIDON-BROYDEN CLASS OF RANK ONE, QUASI-NEWTON MINIMIZATION METHODS TO AN INFINITE DIMENSIONAL HILBERT SPACE WITH APPLICATIONS TO OPTIMAL CONTROL PROBLEMS

By

TERRY ANTHONY STRAETER

A thesis submitted to the Graduate Faculty of
North Carolina State University at Raleigh
in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICS

RALEIGH

1971

APPROVED BY:

____________________ ____________________

____________________ ____________________
Chairman

ABSTRACT

STRAETER, TERRY ANTHONY. On the extension of the Davidon-Broyden class of rank one, quasi-Newton minimization methods to an infinite dimensional Hilbert space with applications to optimal control problems. (Under the direction of HANS SAGAN).

The various elements of the class of rank one, quasi-Newton minimization methods are distinguished by the manner in which a particular parameter is chosen at each iteration. For various choices of this parameter, conditions are found which guarantee that the algorithm's iterates converge to the location of the minimum of a quadratic functional. Also, conditions are found under which the iterates generated by the Davidon-Fletcher-Powell method, the method of conjugate gradients, and the rank one, quasi-Newton method with a particular choice of the parameter are the same. An idea for minimizing a function by a rank one, quasi-Newton method due to Powell is extended to infinite dimensional Hilbert spaces. Also considered is a modification of the rank one, quasi-Newton methods in order to minimize a functional subject to linear constraints. Conditions are found which guarantee the convergence to the location of the constrained minimum of a quadratic functional. The application of these rank one, quasi-Newton algorithms to various classes of optimal control problems is investigated. Also, the algorithms are applied to a sample optimal control problem. The results are compared with the results for the same problem using other known first-order minimization techniques.

BIOGRAPHY

Terry Anthony Straeter was born in [redacted], [redacted], on [redacted]. He received his primary and secondary education in the public school system of Missouri and graduated from Ritenour High School, Overland, Missouri, in 1960.

In September 1964, after graduating in June with a bachelor of arts degree in mathematics from William Jewell College in Liberty, Missouri, the author entered the College of William and Mary as a graduate assistant in the Department of Mathematics. After receiving a master of arts from the College of William and Mary in 1966, Mr. Straeter returned to William Jewell for one year as a visiting instructor in mathematics. In June 1967, he began work for the National Aeronautics and Space Administration at Langley Research Center, Hampton, Virginia.

As an employee of NASA he commenced full-time graduate study at North Carolina State University in Raleigh, North Carolina, in September of 1968. In the spring of 1970 at North Carolina State, the author held a graduate assistantship and later a research assistantship. In the fall of 1970, upon completion of graduate study, he returned to the Langley Research Center where he is presently employed in the Trajectory Application Section of the Analysis and Computation Division.

In August of 1964 he married Miss Virginia JoAnn [redacted] of Kansas City, Missouri. They have two daughters, Kelly Jeanette, born [redacted] and Kristen Joy, born [redacted]. The Straeters presently reside at 846 Isham Place in Newport News, Virginia.

ACKNOWLEDGMENTS

To North Carolina State University, its Mathematics Department and notably the author's Advisory Committee: Doctors R. A. Struble, E. E. Burniston, and M. N. Özisik, the author extends his sincere graditude for their aid throughout his graduate studies. The author wishes to make a special expression of appreciation to his committee chairman, Doctor Hans Sagan for his counsel, interest, and aid not only with regard to this dissertation, but also during the course of his studies at North Carolina State.

The author also wishes to express his thanks to the National Aeronautics and Space Administration for its cooperation and financial support during his graduate study at North Carolina State University.[1] In particular, thanks are extended to his supervisor, John E. Hogge, for his encouragement, guidance, and most patient understanding. To the typists of this manuscript, the East Stenographic Unit, the author offers not only his thanks for an outstanding job, but also because of the nature of the manuscript and the author's untidy hand, his sympathy. Also his thanks are extended to Athena T. Markos for her help in the computer programing of the example problem.

To my wife Jinny, the author says thank you for the encouragement and understanding you have shown for the past 3 years of graduate study. This has been shown in the face of two children and six household moves. You are truly a good wife.

[1] Part of the author's research was supported by NASA Grant NGL 34-002-032.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

1.1 Background and Preview

In the past few years the problem of finding the location of the minimum value of a real valued function of n real variables by numerical methods has been the subject of a great deal of research [7,10,11]. Several iterative procedures have been developed to solve the problem. Much of the work has been directed toward developing algorithms which use the function value and its gradient to locate the minimum by iteration. This type of algorithm is usually referred to as a gradient or first-order method. Historically the method of steepest descent was the first such method. In order to accelerate convergence the method of conjugate gradients was developed later by Hestenes and Stiefel [19] and then was applied to the minimization problem by Fletcher and Reeves [11]. Later first-order methods were developed which were inspired by Newton's second-order method.

Two of the most effective of these techniques are due to Davidon. In 1959 [7] Davidon proposed two techniques for solving the problem. The first method, hereafter denoted by D1, was given in the main body of his report. In 1964 Fletcher and Powell [10] modified D1 and established that for any real valued function the method is stable, that is, does not diverge. (This modified D1 we will denote by DFP.) Moreover, they showed that for a real valued quadratic function of n variables, the DFP algorithm converges in a finite number of steps. In fact, at most $n + 1$ steps are needed. In 1968 Myers [27] showed the relationship between the search directions of the DFP method and those of the conjugate gradient method if the function to be minimized is a quadratic function of n variables. Also in 1968 Horwitz and Sarachik [20] extended the DFP method from an n dimensional Euclidean vector space to an infinite dimensional,real Hilbert space and established convergence of the iterates when the functional to be minimized is quadratic. The result due to Myers was also extended to any real Hilbert space. In 1970 Tokumaru, Adachi, and Goto [36] also extended the DFP algorithm to an infinite dimensional, real Hilbert space and gave a comparison of the DFP method, steepest descent and the conjugate gradient method on some sample optimal control problems.

The second method due to Davidon, denoted herein by D2, was outlined in the appendix to the 1959 report [7]. Later in 1968 [8] he published a modification of the second method and established conditions insuring its convergence to the minimum of a quadratic function of n variables in a finite number of steps and insuring the stability of the method. In 1969 [9] Davidon proposed a second modification of the second method. In 1967 Broyden [4] proposed a family of methods based on a parameter α the choice of which was left unspecified. If $\alpha = 1$, then under certain conditions, Broyden's method and the second Davidon method, D2, are the same. In 1969 Goldfarb [13] established convergence of the iterates of the Broyden algorithm for a class of real functions of n variables when α is chosen by means of a linear minimization technique (i.e., a one-dimensional search).

The purpose of this paper is to extend the Davidon-Broyden family of algorithms to an infinite dimensional real Hilbert space, to establish conditions guaranteeing convergence of the iterates for various algorithms in the family, and to apply the family of algorithms to optimal control problems.

In chapter 2 of this paper, the Davidon-Broyden family of algorithms is extended from an n-dimensional Euclidean vector space to an infinite dimensional real Hilbert space. In the case of a quadratic functional defined on a real Hilbert space, conditions are given which guarantee convergence of the iterates to the location of the minimum for Goldfarb's method of choosing the parameter and for a far more general choice of the parameter. In this approach the need for a linear minimization is eliminated.

In chapter 3, the relationship between the Davidon-Broyden algorithm with Goldfarb's method for choosing the parameter, the DFP method, and the method of conjugate gradients is examined. Also conditions are given which insure that all three methods generate the same search directions. Since the step size is chosen the same way for each method, the same sequence of iterates is generated.

In chapter 4, a modification in the method of choosing the search directions in the extended Davidon-Broyden algorithm is examined. This modification was suggested by Powell in 1970 [30] in an article reviewing the state-of-the-art for finite dimensional optimization. For this modified method, conditions insuring convergence of the iterates to the location of the minimum of a quadratic functional are given.

In chapter 5, the basic algorithm as given in chapter 2 is modified so that it can be applied to a constrained minimization problem. The constrained problem is to find the location of the minimum of a functional $J(x)$ defined on a real Hilbert space H, finite or infinite dimensional, subject to the constraint that $Ax = b$, where b is a fixed element of another real Hilbert space $\bar{H}$ and $A:H \to \bar{H}$, is a bounded linear operator.

The mechanics of applying the algorithm to various classes of optimal control problems are examined and discussed in chapter 6. In many optimal control problems, only controls lying in a subset of the Hilbert space are considered. For example, those $L_p^2[0,1]$ functions whose range is contained in U, a compact, convex subset of R^p. However, the basic algorithm discussed in chapter 2 updates the new estimate of the location of the minimum based only upon the functional's value and its gradient at the old estimate. The new estimate can then lie anywhere in the Hilbert space. Because of this, to apply the basic algorithm to an optimal control problem, its control region U must be an Euclidean space. Park [28] has examined various classes of optimal control problems with a compact, convex control region and by means of certain transformations has reformulated these problems so that their new control region is an Euclidean space. The equations necessary to apply this basic algorithm to these transformed problems are also derived in chapter 6.

In chapter 7, the basic algorithm and its modification are applied to one of the sample control problems given by Tokumaru _et al_. The results are summarized and compared. The results given by

Tokumaru. et al. [36] comparing the conjugate gradient, steepest descent, and DFP methods for the same problem are presented. The Tokumaru et al. results show the DFP method superior in terms of rate of convergence. The DFP method is then compared with our rank one algorithm.

1.2 Outline of Known Methods

Let H denote a real Hilbert space with the inner product $(\ , \ )$. Let R denote the real numbers. A functional $J:H \to R$ is said to be differentiable at x if there exists a linear functional $u_x:H \to R$ such that for $h \in H$

$$J(x + h) - J(x) = u_x(h) + \epsilon_1(h) \tag{1}$$

where $\frac{\epsilon_1(h)}{\|h\|} \to 0$ $\|h\| \to 0$ (Frechét differential). If such a functional u_x exists, then it is unique [33]. Moreover, by the Riesz representation theorem there exists a $g(x) \epsilon H$ such that $(g(x),h) = u_x(h)$ for all $h \in H$ and $g(x)$ is given by

$$\left.\frac{dJ(x + th)}{dt}\right|_{t=0} = (g(x),h) \tag{2}$$

We call $g(x)$ the gradient of the functional J.

Suppose we wish to find the location of the minimum value of a differentiable functional $J:H \to R$ with gradient $g(x)$ at each point x. The three iterative techniques, steepest descent, conjugate gradients, and DFP, could be applied to finding the location of the minimum of J. These algorithms are all descent methods and are only distinguished from each other by the manner in which the search direction is computed. If $x_o \in H$ is the initial estimate of the minimum and $i = 0$, the algorithms are as follows:

Step 1: Compute $J(x_i)$ and $g(x_i)$; if $\|g(x_i)\| = 0$ stop, otherwise,

Step 2: Let $x_{i+1} = x_i + \alpha_i s_i$ where α_i, called the step size, is a real number and $s_i \in H$ is called the search direction. α_i is chosen so that $J(x_i + \alpha_i s_i) \leq J(x_i + \lambda s_i)$ for all $\lambda \in R$. The search direction s_i for the above-mentioned methods is chosen in one of the following three ways.

If $s_i = - g(x_i)$, then the algorithm is the classical method of steepest descent [25].

If $s_i = - g(x_i) + \beta_{i-1}s_{i-1}$ where $\beta_{i-1} = \dfrac{(g(x_i),g(x_i))}{(g(x_{i-1}),g(x_{i-1}))}$ and $s_o = - g(x_o)$ then the algorithm is called the method of conjugate gradients [11, 18, 19, 23, 25, 34].

Finally we have the DFP method, if $s_i = - H^{(i)}g(x_i)$ where the $H^{(i)}: H \rightarrow H$ $i = 0,1,2 \dots$ are a sequence of linear operators defined iteratively as follows: $H^{(o)}$ is a strongly positive, linear, self-adjoint operator on H and $H^{(i+1)} = H^{(i)} + A^{(i)} + C^{(i)}$ where $A^{(i)}$ and $C^{(i)}: H \rightarrow H$ are so that if $x \in H$

$$A^{(i)}x = \frac{(H^{(i)}y_i,x)}{(y_i,H^iy_i)} H^{(i)}y_i$$

where

$$y_i = g(x_{i+1}) - g(x_i)$$

and

$$C^{(i)}x = \frac{(\sigma_i, x)}{(\sigma_i, y_i)} \sigma_i$$

where

$$\sigma_i = x_{i+1} - x_i$$

We set $i + 1 = i$, return to step 1, and continue.

A summary of the results known concerning the application of these three techniques to quadratic functionals will be given at the end of the next section.

1.3 Quadratic Functionals

Let $A:H \rightarrow H$ be a linear, self-adjoint operator such that

$$m \| x \|^2 \leq (x, Ax) \leq M \| x \|^2 \qquad (3)$$

where

$$M = \sup_{x \neq \theta} \frac{(x, Ax)}{\| x \|^2}, \quad m = \inf_{x \neq \theta} \frac{(x, Ax)}{\| x \|^2} \qquad (4)$$

and where we assume that $0 < m \leq M$. Hence, $\| A \| = M$ [2]. Since $m > 0$, A^{-1} exists [26] and A^{-1} is also self-adjoint. Moreover, we have

$$\frac{1}{M} \| x \|^2 \leq (x, A^{-1}x) \leq \frac{1}{m} \| x \|^2 \qquad (5)$$

We call the functional $J:H \to R$ given by

$$J(x) = J_o + (x,b) + \frac{1}{2}(x,Ax) \tag{6}$$

a quadratic functional on H where b is a fixed element in H and $J_o \in R$. Using (5) we can compute the gradient $g(x)$ of the quadratic functional given by (6) as follows:

$$\frac{dJ(x + th)}{dt} = \frac{d(J_o + (x + th,b) + \frac{1}{2}(x + th,A(x + th)))}{dt}$$

$$= \frac{d(J_o)}{dt} + \frac{(h,b)d(t)}{dt} + \frac{d(x,h)}{dt}$$

$$+ \frac{\frac{1}{2}d\left[(x,Ax) + 2t(h,Ax) + t\ (h,Ah)\right]}{dt}$$

$$= (h,b) + (h,Ax) + t(h,Ah)$$

and we have

$$\left.\frac{dJ(x + th)}{dt}\right|_{t=0} = (h,b + Ax).$$

Therefore, by (2), the gradient $g(x)$ of the quadratic functional $J(x)$ is given by

$$g(x) = b + Ax. \tag{7}$$

The following well known theorem states a necessary and sufficient condition for $\tilde{x}$ to minimize $J(x)$:

Theorem 1.1: A necessary and sufficient condition that $\tilde{x}$ minimizes $J(x)$ as given by (6) is that $g(\tilde{x}) = \theta$ where θ denotes the zero element of H.

Proof: Suppose $g(\tilde{x}) = \theta$ then $A\tilde{x} + b = \theta$ by (7) so that $b = - A\tilde{x}$, hence if $x \neq \tilde{x}$

$$J(\tilde{x}) - J(x) = J_0 + (\tilde{x},b) + \frac{1}{2}(\tilde{x},A\tilde{x}) - J_0 - (x,b) - \frac{1}{2}(x,Ax)$$

$$= (\tilde{x}, - A\tilde{x}) + \frac{1}{2}(\tilde{x},A\tilde{x}) + (x,A\tilde{x}) - \frac{1}{2}(x,Ax)$$

since

$$b = - A\tilde{x}$$

Therefore, $J(\tilde{x}) - J(x) = - \frac{1}{2}(\tilde{x} - x,A(\tilde{x} - x))$ since $A = A^*$ and $((\tilde{x} - x),A(\tilde{x} -x)) > m\|\tilde{x} - x\|^2 > 0$ by (3). Hence, $J(\tilde{x}) - J(x) \leq -\frac{1}{2} m\|\tilde{x} - x\|^2 < 0$. So $\tilde{x}$ is the location of the minimum of J.

Conversely let us suppose that $J(\tilde{x}) \leq J(x)$ for all $x \in H$. If we let $h \in H$, h fixed, then for $t \in R$ we have

$$J(\tilde{x} + th) - J(\tilde{x}) \geq 0. \qquad (8)$$

Hence,

$$0 \leq J_o + (\tilde{x} + th,b) + \frac{1}{2}(\tilde{x} + th,A(\tilde{x} + th)) - J_o - (\tilde{x},b) - \frac{1}{2}(\tilde{x},A\tilde{x})$$

$$= t(h,b) + t(h,A\tilde{x}) + \frac{1}{2}t^2(h,Ah)$$

$$= t\left[(h,g(\tilde{x})) + \frac{1}{2}t(h,Ah)\right] \leq t(h,g(\tilde{x})) + \frac{t^2M}{2}\| h \|^2$$

Now suppose $(h,g(\tilde{x})) < 0$; then since $M, \|h\|^2$ and $(g(\tilde{x}),h)$ are constants and $M > 0, \|h\|^2 < 0$, we can force $t(h,g(\tilde{x})) + \frac{t^2}{2}M\|h\|^2 < 0$ by letting $t \to 0^+$. So this would imply $J(\tilde{x} + th) - J(\tilde{x}) < 0$ which contradicts (8). Similarly, if $(h,g(\tilde{x})) > 0$ by letting $t \to 0^-$ we have $t(h,g(\tilde{x})) + \frac{t^2}{2}M\|h\|^2 < 0$ which leads to a contradiction to (8). Hence, it must be true that $(h,g(\tilde{x})) = 0$, and since h was an arbitrary element in H it follows that $g(\tilde{x}) = \theta$.

Theorem 1.2: If $\tilde{x}$ denotes the location of the minimum of the quadratic functional J given by (6) then

$$\tilde{x} = - A^{-1}b. \qquad (9)$$

Moreover, if $x,h \in H$ are such that $x + h = \tilde{x}$ then

$$h = - A^{-1}g(x) \qquad (10)$$

Proof: By theorem 1.1 and (7) $\theta = g(\tilde{x}) = A\tilde{x} + b$, so that $\tilde{x} = - A^{-1}b$ since A^{-1} exists. If $x + h = \tilde{x}$, then $g(x + h) = g(\tilde{x}) = \theta$. So, $A(x + h) - b = \theta$, $Ax + Ah = -b$. Hence, $Ah = - (Ax + b) = g(x)$ by (7). Therefore, $h = - A^{-1}g(x)$.

Of course, the equation $h = - A^{-1}g(x)$ is the basis for the well known Newton-Raphson method for minimizing a functional on a Hilbert space [22] .

Other useful results due to the fact that J is a quadratic functional are the following: If $x,x^* \in H$, then

$$A^{-1}(g(x) - g(x^*)) = A^{-1}(Ax + b - Ax^* - b) = x - x^*. \qquad (11)$$

Hence, if we let $y = g(x) - g(x^*)$ and $\sigma = x - x^*$, we have

$$A^{-1}y = \sigma. \tag{12}$$

Moreover, we can see that $s \in H$ and $\alpha \in R$ are such that

$$x^* = x + \alpha s \tag{13}$$

then by (7) and (13) $g(x^*) = Ax^* + b = Ax^* + b + \alpha As.$. So that by (7) we have again

$$g(x^*) = g(x) + \alpha As. \tag{14}$$

Also, for all $x_0 \in H$ the smallest closed, convex set containing the points $x \in H$ at which $J(x) \leq J(x_0)$ is bounded [25] . We denote this set by $S'_{x_0} = \overline{\text{conv}}\{x \in H : J(x) \leq J(x_0)\}$.

It is known [20] that if a quadratic functional is minimized by the conjugate gradient method the ith search direction is given by

$$s_i = - \|g(x_i)\|^2 \sum_{l=0}^{i} \frac{g(x_l)}{\|g(x_l)\|^2}. \tag{15}$$

Horwitz and Sarachik have shown for a quadratic functional that the ith search direction of the DFP method is given by

$$s_i = - H^{(o)}(g(x_i), H^{(i)}g(x_i)) \sum_{l=0}^{i} \frac{g(x_l)}{(g(x_l), H^{(o)}g(x_l))}. \tag{16}$$

If $H^{(o)}$ is the identity denoted by I, then (15) and (16) are the same directions. Since the step size is picked in the same fashion for both methods they will generate the same sequence of iterates [20] .

The convergence of the iterates to the location of the minimum by the method of steepest descent, method of conjugate gradients, and the DFP algorithm has been established [20], [36] for the case where the functional to be minimized is quadratic.

A note concerning the notation to be used throughout this paper would appear to be in order. It shall be our practice that if reference is made to an equation, identity or relation in the same chapter, only the number at the right-hand side of the page will be enclosed in parenthesis. However, if the reference is to an equation, etc., in another chapter, then the chapter number followed by a period and then by the reference number will be given. Theorems will be numbered sequentially with a chapter prefix, that is, as theorem 1.1, and will be referenced in that fashion. The numbers enclosed in square brackets refer to the references in chapter 8.

Also herein we shall denote by $L_r^2[t_o,t_1]$ the real Hilbert space of Lebesque measurable functions $u = u(t)$ defined on $[t_o,t_1]$ with range in R^r (Euclidean r space) such that

$$\sum_{i=1}^{r} \int_{t_o}^{t_1} [u_i(t)]^2 dt < \infty$$

where $u,$ (t), $u_2(t)$, ... $u_r(t)$ are the components of u.

2. THE CLASS OF DAVIDON-BROYDEN ALGORITHMS

In this chapter, we shall discuss the extension to an infinite dimensional Hilbert space of the Davidon-Broyden minimization algorithms alluded to in chapter 1. We shall also relate conditions insuring the convergence of the iterates of various members of this family of algorithms in the case where the functional to be minimized is quadratic. In the case of a finite dimensional Hilbert space, Broyden [4] called this family of algorithms "quasi-Newton methods." Special cases of this family have been called "optimal variance algorithm" by Goldfarb [13] and "rank one variance algorithm" by Davidon [9]. The author's contribution is to show the relationship of these methods to each other, to extend their applicability to infinite dimensional real Hilbert spaces, and to establish conditions insuring convergence of the iterates. For the latter purpose, new proofs of convergence of the algorithm's various manifestations were necessary.

2.1 Outline of the Class of Algorithms

Let $J:H \rightarrow R$ be a differentiable functional with gradient $g(x)$. Let $x_o \in H$ be the initial estimate of the location of the minimum of J, and let $V^{(o)}$ be a self-adjoint, strongly positive linear operator from H onto H. Let $M_o \geq m_o > 0$ be such that $m_o I \leq V^{(o)} \leq M_o I$. If J, the functional to be minimized, is quadratic as in chapter 1, then $V^{(o)}$ is an estimate of A^{-1}. We compute $J(x_o)$ and $g(x_o)$ and and obtain the first iteration as follows:

Step 1: Let

$$x^* = x_n - \alpha_n V^{(n)} g_n \tag{1}$$

where g_n denotes $g(x_n)$ and α_n is a scalar, the choice of which is discussed later. Let

$$s_n = - V^{(n)} g_n \tag{2}$$

and compute $J(x^*)$ and $g(x^*)$ denoted by g^*; if $\|g^*\| = 0$, a necessary condition for x^* to be the location of the minimum, we stop. If J is a quadratic functional and $g^* = \theta$, then by theorem 1.1 x^* is the location of the minimum.

Step 2: Compute the residual vector

$$r_n = V^{(n)} g^* - V^{(n)} g_n + \alpha_n V^{(n)} g_n \tag{3}$$

that is,

$$r_n = V^{(n)} (g^* - (1 - \alpha_n) g_n) \tag{4}$$

or

$$r_n = V^{(n)} y_n - \alpha_n s_n \tag{5}$$

where $y_n = g^* - g_n$. If $r_n = \theta$, then set $\alpha_n = 1$ and return to step 1.

Step 3: Define scalars

$$\rho_n = (g^*,r_n) \tag{6}$$

and

$$\gamma_n = - \frac{(g_n,r_n)}{\rho_n} \tag{7}$$

and let

$$\lambda_n = \begin{cases} \dfrac{\gamma_n}{1 + \gamma_n} & \text{if } \gamma_n \neq - 1 \\ \\ 1 & \text{if } \gamma_n = - 1 \end{cases} \tag{8}$$

Step 4: Let

$$V^{(n+1)} = V^{(n)} + \frac{(\lambda_n - 1)}{\rho_n} B^{(n)} \tag{9}$$

where $B^{(n)}:H \to H$ is defined such that for all $x \in H$

$$B^{(n)}x = (x,r_n)r_n. \tag{10}$$

Step 5: If $J(x^*) < J(x_n)$, let $x_{n+1} = x^*$ and, consequently, $J(x_{n+1}) = J(x^*)$ and $g_{n+1} = g^*$; otherwise, let $x_{n+1} = x_n$ so that $J(x_{n+1}) = J(x_n)$ and $g_{n+1} = g_n$. Set $n = n + 1$ and go to step 1.

The elements of the class of algorithms outlined above are distinguished by the manner in which the parameter α_n is chosen with each iteration. Davidon [8], Broyden [4], and Goldfarb [13] proposed techniques for choosing α_n in the finite dimensional case. For Davidon's rank one variance algorithm $\alpha_n = 1$ for all n, however, the scalar λ_n given by (8) is chosen so that certain inequality constraints are satisfied. These constraints insure that Davidon's $V^{(n)}$ remain positive definite. Goldfarb's optimal variance algorithm required that α_n be chosen so that $J(x_n + \alpha s_n)$ be minimized with respect to α. The Broyden quasi-Newton method requires only that α_n be chosen so that $\left(V^{(n)}\right)^{-1}$ exists. For a quadratic functional theorem 2.7 proved later shows that for $V^{(o)} \geq A^{-1}$ or $V^{(o)} \leq A^{-1}$ either Davidon's or Goldfarb's method of choosing α_n satisfy Broyden's criteria.

For the remainder of chapter 2, we shall assume that the functional to be minimized is quadratic as defined in section 2 of chapter 1. We shall make note of any results which are independent of the type of functional to be minimized.

2.2 Theorems That Are Independent of the Choice of α_n.

Theorem 2.1: $B^{(n)}$ as given in (10) is a self-adjoint positive operator for all n, for any choice of α_n.

Proof: If $x \in H$, then

$$(x,B^{(n)}x) = (x,(x,r_n)r_n) = (x,r_n)^2 \geq 0$$

and if $x,y \in H$, then

$$(x,B^{(n)}y) = (x,(y,r_n)r_n) = (y,r_n)(x,r_n) = (y,(x,r_n)r_n) = (y,B^{(n)}x)$$

Theorem 2.2: $V^{(n)}$ is self-adjoint for all n, for any choice of α_n.

Proof: $V^{(n)} = V^{(o)} + \sum_{i=0}^{n-1} \frac{(\lambda_i - 1)}{\rho_i} B^{(i)}$ by (9), and $V^{(o)}$ is self-adjoint by definition. By the above theorem, the $B^{(i)}$'s are self-adjoint and the finite sum of self-adjoint operators is self-adjoint.

Notice that the two theorems proved above are independent of the type of functional that is to be minimized.

We have seen in chapter 1 that the location of the minimum $\tilde{x}$ of a quadratic functional is given by $\tilde{x} = x_n - A^{-1}g_n$. Also recall from chapter 1 that the change in x from one iteration to the next for the Newton Raphson method is given by $-A^{-1}g_n$. In the algorithm outlined in section 1, the change is $-\alpha_n V^{(n)} g_n$ hence, the name quasi-Newton was given to the finite dimensional form of these algorithms by Broyden [4]. The search directions for the algorithm outlined in section 1 are given by $-V^{(n)} g_n$ and we want $V^{(n)}$ to play the role of A^{-1}. Hence, it is desirable that the sequence of operators $V^{(n)}$ retain from one iteration to the next the following property: if for some $u \in H$, $A^{-1}u = V^{(n)}u$ then $A^{-1}u = V^{(n+1)}u$. By the definition of the vector r_n we have the following general result.

Theorem 2.3: If $u \in H$ is such that $A^{-1}u = V^{(n)}u$ and $B:H \to H$ is a linear operator such that $B - V^{(n)} = \mu B^{(n)}$ for some real μ then $A^{-1}u = Bu$.

Proof: Since $A^{-1}(g^* - g_n) = x^* - x_n$ by (1.12) and $x^* = x_n - \alpha_n V^{(n)} g_n$ by definition, then

$$x^* - x_n = -\alpha_n V^{(n)} g_n = A^{-1}(g^* - g_n) \tag{11}$$

and by (3)

$$r_n = V^{(n)}(g^* - g_n) + \alpha_n V^{(n)} g_n.$$

Therefore, $r_n = V^{(n)}(g^* - g_n) - A^{-1}(g^* - g_n)$ by (11) and hence

$$r_n = \left(V^{(n)} - A^{-1}\right)(g^* - g_n). \tag{12}$$

Since $A^{-1}u = V^{(n)}u$ we have

$$\left(V^{(n)} - A^{-1}\right)u = \theta \tag{13}$$

So

$$(r_n, u) = \left(\left(V^{(n)} - A^{-1}\right)(g^* - g_n), u\right) = \left((g^* - g_n), \left(V^{(n)} - A^{-1}\right)u\right)$$

$$= (g^* - g_n, \theta) = 0$$

by theorem 2.2 and equations (12) and (13). Hence, the hypothesis $(B - A^{-1})u = \mu B^{(n)}u$ implies

$$\mu B^{(n)}u = \mu(u, r_n)r_n = \mu \cdot 0 \cdot r_n = \theta$$

Therefore, $Bu = A^{-1}u$.

Since $V^{(n+1)} = V^{(n)} + \frac{(\lambda_n - 1)}{\rho_n} B^{(n)}$ we have the following:

Corollary 1: If $V^{(n)}u = A^{-1}u$ for some $u \in H$, then $V^{(n+1)}u = A^{-1}u$.

In chapter 1 we showed that for a quadratic functional

$$A^{-1}(g^* - g_n) = x^* - x_n.$$

The following theorem gives the fundamental reason for our choice of $V^{(n+1)}$, that is, so that when $\gamma_n \neq -1$, then $V^{(n+1)}$ and A^{-1} will agree on the space spanned by $g^* - g_n$.

Theorem 2.4: (Basic theorem). If $\gamma_n \neq -1$, then

$$V^{(n+1)}(g^* - g_n) = x^* - x_n$$

that is,

$$V^{(n+1)}y_n = \alpha_n s_n$$

Proof: If $r_n = \theta$ then by (10) $B^{(n)}$ is the zero operator, therefore $V^{(n+1)} = V^{(n)}$, so that $V^{(n+1)}y_n = \alpha_n s_n$. Otherwise, consider

$$V^{(n+1)}y_n - \alpha_n s_n = V^{(n)}y_n + \frac{(\lambda_n - 1)}{\rho_n}(r_n, y_n)r_n - \alpha_n s_n \qquad \text{(by (9))}$$

$$= r_n - \alpha_n V^{(n)}g_n + \frac{(\lambda_n - 1)}{\rho_n}(r_n, y_n)r_n - \alpha_n s_n \qquad \text{(by (3))}$$

$$= r_n\left\{1 + \frac{(\lambda_n - 1)}{\rho_n}(r_n, y_n)\right\} \qquad \text{(by (2))}$$

$$= r_n\left\{1 + \frac{(\lambda_n - 1)}{\rho_n}(\rho_n + \gamma_n\rho_n)\right\} \qquad \text{(by (6),(7))}$$

$$= r_n\, 0 = \theta \qquad \text{(by (8))}$$

Notice that the basic theorem is independent of the fact that J is a quadratic functional. The following corollary combines theorems 2.3 and 2.4 to show that each iteration, if $\gamma_n \neq -1$, raises the dimension of the subspace, on which $V^{(n)}$ and A^{-1} agree, by one. Hence, some authors [10] have called the finite dimensional form of this algorithm a rank one method.

Corollary 1: (Fundamental property of $V^{(n)}$) $V^{(n)}y_i = \alpha_i s_i$ for all $i < n$ if $\gamma_j \neq -1$ for $j = 0,1,\ldots,n$

Proof: (By mathematical induction)

$$V^{(1)}y_o = \alpha_o s_o \qquad \text{(by theorem 2.4)}$$

Assume $V^{(n)}y_i = \alpha_i s_i$ for all $i < n$. Consider $V^{(n+1)}y_i$ for $i = n$. Then by theorem 2.4, $V^{(n+1)}y_n = \alpha_n s_n$. Otherwise, for $i < n$, since $A^{-1}y_i = \alpha_i s_i$ by (1.12) and $V^{(n)}y_i = \alpha_i s_i$, A^{-1} and $V^{(n)}$ agree on y_i. The corollary to theorem 2.3 implies $V^{(n+1)}y_i = \alpha_i s_i$.

The above corollary is most useful in later convergence arguments and, hence, we have named it "the fundamental property of $V^{(n)}$."

In order to facilitate the proof of some later results, we shall now find another way of expressing $(\lambda_n - 1)/\rho_n$.

Theorem 2.5: If $\gamma_i \neq -1$, then

$$\frac{(\lambda_i - 1)}{\rho_i} = -(V^{(i)}y_i - \alpha_i s_i, y_i)^{-1} = -(r_i, y_i)^{-1}.$$

Proof: $(\lambda_i - 1)/\rho_i = (\gamma_i/(\gamma_i+1) - 1)/\rho_i = -(\rho_i(\gamma_i + 1))^{-1}$
$= -(\rho_i - (r_i,g_i))^{-1} = -((r_i,g^*) - (r_i,g_i))^{-1} = -(r_i,y_i)^{-1}$
$= -(V^{(i)}y_i - \alpha_i s_i, y_i)^{-1}$

In view of this, (9) can be written as

$$V^{(n+1)} = V^{(n)} - \frac{B^{(n)}}{\left(V^{(n)}y_n - \alpha_n s_n, y_n\right)} \tag{14}$$

and since $\alpha_n s_n = A^{-1}y_n$, we have

$$V^{(n+1)} = V^{(n)} - \frac{B^{(n)}}{\left((V^{(n)} - A^{-1}\right)y_n, y_n\right)}, \tag{15}$$

which yields the following theorem:

Theorem 2.6: If $V^{(o)} \geq A^{-1}$, then $V^{(n)} \geq A^{-1}$ for all n and similarly, if $V^{(o)} \leq A^{-1}$, then $V^{(n)} \leq A^{-1}$ for all n.

Proof: We proceed by induction and assume that $V^{(n)} \geq A^{-1}$. If $V^{(n+1)} = V^{(n)}$, i.e., $\gamma_n = -1$, the result is trivial. Otherwise, by (15) and (10),

$$\left(x,(V^{(n+1)} - A^{-1}\right)x\right) = \left(x,\left(V^{(n)} - A^{-1}\right)x\right) - \frac{(x,r_n)^2}{\left(y_n,\left(V^{(n)} - A^{-1}\right)y_n\right)}$$

From the C.B.S. inequality [2] : $\left(x,\left(V^{(n)} - A^{-1}\right)x\right) - \frac{\left(x,\left(V^{(n)} - A^{-1}\right)y_n\right)^2}{\left(y_n,\left(V^{(n)} - A^{-1}\right)y_n\right)} \geq$

The second part of the theorem is obtained by merely considering $\left(x,\left(A^{-1} - V^{(n+1)}\right)x\right)$ instead.

The following theorem gives a condition under which the $V^{(n)}$'s form a monotone sequence of self-adjoint bounded linear operators.

<u>Theorem 2.7:</u> If $V^{(o)} \geq A^{-1}$, then $V^{(n)} \leq V^{(n-1)} \leq \ldots \leq V^{(o)}$ for all n. Similarly, if $V^{(o)} \leq A^{-1}$, then $V^{(n)} \geq V^{(n-1)} \geq \ldots V^{(o)}$ for all n.

<u>Proof:</u> By theorem 2.6, if $V^{(o)} \geq A^{-1}$, then $V^{(n)} \geq A^{-1}$ for all n. If $V^{(n+1)} = V^{(n)}$, <u>i.e.</u>, $\gamma_n = -1$, then the assertion is obvious. Otherwise, we have

$$\left(x,\left(V^{(n+1)} - V^{(n)}\right)x\right) = -\frac{\left(x,B^{(n)}x\right)}{\left(y_n,\left(V^{(n)} - A^{-1}\right)y_n\right)} \leq 0$$

by (15). The inequality holds since theorem 2.1 gives $\left(x,B^{(n)}x\right) \geq 0$ and from theorem 2.6 $V^{(n)} - A^{-1} \geq 0$. The second part of the theorem follows by considering $V^{(n)} - V^{(n+1)}$ instead.

<u>Corollary 1:</u> If $V^{(o)} \leq A^{-1}$ or $V^{(o)} \geq A^{-1}$, then the $V^{(n)}$'s form a monotone sequence of strongly positive self-adjoint linear operators bounded by $V^{(o)}$ and A^{-1}. Moreover, there exists a strongly positive self-adjoint operator V such that $\lim_{n \to \infty} V^{(n)}x = Vx$ for all $x \in H$.

<u>Proof:</u> The $V^{(n)}$'s form a bounded monotone sequence of strongly positive, self-adjoint operators by theorems 2.2 and 2.7. That is, if $V^{(o)} < A^{-1}$, we have $V^{(o)} \leq V^{(1)} \leq V^{(2)} \leq \ldots \leq V^{(n)} \leq \ldots \leq A^{-1}$. This

implies the existence of a strongly positive, self-adjoint linear operator V such that $V^{(n)}$ converges to V pointwise [1].

Theorem 2.8: If $V^{(o)} \leq A^{-1}$ or $V^{(o)} \geq A^{-1}$ and $\gamma_n \neq -1$ for all n and if S is the closure of the space spanned by $\{y_i\}$, then $\lim_{n\to\infty} V^{(n)}x = A^{-1}x$ for all $x \in S$ independent of the choice of the α_n's. (By closure of the space spanned by a set M, we mean the smallest topologically closed subspace containing M.)

Proof: For any $x \in S$ there exist $\beta_i \in R$ such that

$$x = \sum_{i=0}^{\infty} \beta_i y_i. \tag{16}$$

Consider

$$\|A^{-1}x - V^{(n)}x\| = \|(A^{-1} - V^{(n)})x\| = \left\|(A^{-1} - V^{(n)}) \sum_{i=0}^{\infty} \beta_i y_i\right\|$$

$$\leq \left\|(A^{-1} - V^{(n)}) \sum_{i=0}^{n-1} \beta_i y_i\right\| + \left\|(A^{-1} - V^{(n)}) \sum_{i=n}^{\infty} \beta_i y_i\right\|$$

By the corollary to theorem 2.4, $(A^{-1} - V^{(n)}) \sum_{i=0}^{n-1} \beta_i y_i = \theta$. Since $V^{(o)} \geq A^{-1}$ or $V^{(o)} \leq A^{-1}$ by theorem 2.7 and its corollary, it must be that $\|V^{(n)}\| \leq \|A^{-1}\|$ or $\leq \|V^{(o)}\|$. So $\|A^{-1} - V^{(n)}\|$ is bounded for all n, and by (16) it follows that the remainder must go to zero, i.e., $\left\|\sum_{i=n}^{\infty} \beta_i y_i\right\| \to 0$ as $n \to \infty$. So we have $\lim_{h\to\infty} \|A^{-1}x - V^{(n)}x\| = 0$.

Corollary: If $V^{(o)} \leq A^{-1}$ or $V^{(o)} \geq A^{-1}$ and $\gamma_n \neq - 1$ for all n and the y_i form a basis for H, then $V^{(n)} \to A^{-1}$ point-wise independent of the choice of α_n.

Notice that all these results have been established without regard to the choice of α_n. We called r_n as defined in (3) and (4) a residual vector. The reason for this terminology will now be explained.

Suppose $r_n = \theta$ for some n. Then $V^{(n)}y_n = \alpha_n s_n$, and if $\left(V^{(n)}\right)^{-1}$ exists, $y_n = \alpha_n\left(V^{(n)}\right)^{-1}V^{(n)}g_n = - \alpha_n g_n$ by (2) and by (5) we have $y_n = g^* - g_n = - \alpha_n g_n$. By (1.14) $g^* = g_n + \alpha_n A s_n$. Therefore, $\alpha_n A s_n = - \alpha_n g_n$, so that $s_n = - A^{-1}g_n$. Hence, since $s_n = -V^{(n)}g_n$ we have $V^{(n)}g_n = A^{-1}g_n$.

As we have seen in chapter 1 (theorem 1.2), the minimum of J is attained by $\tilde{x} = x_n - A^{-1}g_n$. In the basic algorithm outlined in section 1 of this chapter, step 2 says if $r_n = \theta$ we let $\alpha_n = 1$ and repeat step 1. Then the new x^* is $x^* = x_n - V^{(n)}g_n$ and we have shown above that $s_n = - A^{-1}g_n$, hence, $V^n g_n = A^{-1}g_n$. Therefore, by theorem 1.2 x^* is the location of the minimum of J. This explains the reason for step 2, and we have proved the following:

Theorem 2.9: If $r_n = \theta$ and $\left(V^{(n)}\right)^{-1}$ exists, then by applying step 2 of the basic algorithm we let $\alpha_n = 1$ and we find that the resulting x^* given by $x^* = x_n - V^{(n)}g_n$ is the location of the minimum of J.

2.3 Convergence if α_n is Chosen by a One Dimensional Minimization Process

There are two rather obvious ways to choose α_n at each step: (1) let $\alpha_n = 1$ for all n, and (2) let α_n be such that $J(x_n + \alpha_n s_n) \leq J(x_n + \lambda s_n)$ for all real λ. Both cases have been investigated by Davidon and Goldfarb and convergence has been established in the case of a quadratic functional on a finite dimensional Hilbert space.

We shall now demonstrate the convergence of the algorithm of section 1 to the location of the minimum of a quadratic functional on an infinite dimensional Hilbert space when α_n is chosen for every n so that

$$J(x_n + \alpha_n s_n) \leq J(x_n + \lambda s_n) \tag{17}$$

for all real λ. This, of course, implies that $x_{n+1} = x^*$ in step 5 of the algorithm given in section 1. If α_n is chosen in this manner, then, by necessity,

$$\frac{dJ(x_n + \lambda s_n)}{d\lambda} = 0 \tag{18}$$

at $\lambda = \alpha_n$.

That is, $(g^*, s_n) = (g(x_n + \alpha_n s_n), s_n) = 0$ so that from (1.7) we have

$$\alpha_n = \frac{(s_n, g_n)}{(s_n, As_n)} \tag{19}$$

Therefore,

$$\begin{aligned}
J(x_1) &= J_o + (b,x_1) + \frac{1}{2}(x_1,Ax_1) && \text{(by def.)} \\
&= J_o + (b,x_o + \alpha_o s_o) + \frac{1}{2}(x_o + \alpha_o s_o, A(x_o + \alpha_o s_o)) \\
&= J_o + (b,x_o) + \frac{1}{2}(x_o,Ax_o) + \alpha_o\big[(s_o,b) + (s_o,Ax_o)\big] + \frac{\alpha_o^2}{2}(s_o,As_o) \\
&= J(x_o) + \alpha_o(s_o,g_o) + \frac{\alpha_o^2}{2}(s_o,As_o) && \text{(by (1.6)} \\
&= J(x_o) - \frac{1}{2}\,\frac{(s_o,g_o)^2}{(s_o,As_o)}\,. && \text{(by 19)}
\end{aligned}$$

In general,

$$J(x_{n+1}) = J(x_o) - \sum_{i=0}^{n} \frac{(s_i,\ g_i)^2}{2(s_i,As_i)}\,.$$

Since, $\inf J > -\infty$ and $J(x_{n+1}) \leq J(x_n)$, it must be that

$$\lim_{n\to\infty} J(x_{n+1}) = J(x_o) - \lim_{n\to\infty} \sum_{i=0}^{n} \frac{(s_i,g_i)^2}{2(s_i,As_i)} > -\infty$$

so that

$$\sum_{i=0}^{\infty} \frac{(s_i,g_i)^2}{2(s_i,As_i)} < \infty$$

which implies that by necessity

$$\lim_{i\to\infty} \frac{(s_i,g_i)^2}{(s_i,As_i)} = 0 \tag{20}$$

Since its derivation in no way depended on (2), (20) must be true for any descent method. This result and the following lemma are given by Horwitz and Sarachik [20]. They used them to prove convergence of Davidon's first method, steepest descent, and the conjugate gradient method in an infinite dimensional real Hilbert space for the problem under consideration.

Lemma 2.1: If $g_n \to \theta$ as $n \to \infty$, then x_n converges in norm to the location of the minimum $\tilde{x} = -A^{-1}b$.

Proof: $0 \leq (x_n + A^{-1}b, A(x_n + A^{-1}b)) = (x_n + A^{-1}b, g_n)$
$\leq \|x_n + A^{-1}b\| \cdot \|g_n\| \to 0$

Now $\|x_n + A^{-1}b\|$ is bounded for all n, since for all n, x_n is contained in a bounded set, namely $S'_{x_0} = \overline{\text{conv}}\{x \in H: J(x) \leq J(x_0)\}$ as in chapter 1. Hence, $\lim_{n\to\infty} (x_n + A^{-1}b, A(x_n + A^{-1}b)) = 0$ and since A is strongly positive, we have $\lim_{n\to\infty} x_n + A^{-1}b = \theta$.

We can now prove a general convergence theorem for this case.

Theorem 2.10: If there exist positive reals α, β such that $\alpha I \leq V^{(n)} \leq \beta I$ for all n larger than some N and if α_n is chosen as in (17), then $\lim_{n\to\infty} \|x_n + A^{-1}b\| = 0$, that is, x_n converges in norm to the location of the minimum.

Proof: Since for all $u \in H$, $m\|u\|^2 \leq (u,Au) \leq M\|u\|^2$ we have

$$\frac{1}{M\|u\|^2} \leq \frac{1}{(u,Au)} \leq \frac{1}{m\|u\|^2}$$

and since $\alpha\|u\|^2 \leq (u,V^{(n)}u) \leq \beta\|u\|^2$ for all n,

$$\frac{1}{\beta\|u\|^2} \leq \frac{1}{(u,V^{(n)}u)} \leq \frac{1}{\alpha\|u\|^2} .$$

Since $V^{(n)}$ is self-adjoint, we have $\|V^{(n)}u\| \leq \beta\|u\|$ [2]. Therefore,

$$\frac{(s_k,g_k)^2}{(s_k,As_k)} \geq \frac{(s_k,g_k)^2}{M\|s_k\|^2} = \frac{(g_k,V^{(k)}g_k)^2}{M\|V^{(k)}g_k\|^2} \geq \frac{(g_k,V^{(k)}g_k)^2}{M\beta^2\|g_k\|^2}.$$

$$\geq \frac{\alpha}{M\beta^2}\,\frac{(\|g_k\|^2)^2}{\|g_k\|^2} = \frac{\alpha}{M\beta^2}\,\|g_k\|^2 \geq 0$$

and by (20) $(s_k,g_k)^2/(s_k,As_k) \to 0$. Therefore, $\|g_k\|^2 \to 0$ as $k \to \infty$ and by lemma 2.1 $x_k \to - A^{-1}b$ in norm.

Corollary 1: If $V^{(o)} \leq A^{-1}$ or $V^{(o)} \geq A^{-1}$ and α_n is chosen as in (17), then $J(x_n)$ converges to the minimum of $J(x)$, and moreover x_n converges in norm to the location of the minimum.

Proof: If $V^{(o)} \leq A^{-1}$, then by theorems 2.6 and 2.7 we have $V^{(o)} \leq V^{(n)} \leq A^{-1}$ for all n. Hence, $M_oI \leq V^{(n)} \leq \frac{1}{m} I$ for all n.

2.4 Convergence with a More General Choice of α_n

Let $\{\alpha_k\}$ denote a sequence of real numbers. We then apply the algorithm outlined in section 1 using these $\{\alpha_k\}$'s in step 1 to minimize the quadratic function discussed in chapter 1, section 3. Select a subsequence $K = \{\alpha_{k_n}\}$ so that $J(x^*) < J(x_{k_n})$ for all

$n = 0,1,2,\ldots$ To simplify the notation, let us write n for k_n. Then we have

$$g_n = g_o + (g_1 - g_o) + (g_2 - g_1) + \ldots + (g_n - g_{n-1})$$

or

$$g_n = g_o + \sum_{i=0}^{n-1} y_i$$

since

$$y_i = g_{i+1} - g_i.$$

Then

$$V^{(n)}g_n = V^{(n)}g_o + \sum_{i=0}^{n-1} V^{(n)}y_i. \tag{21}$$

From the corollary to theorem 2.4, $V^{(n)}y_i = \alpha_i s_i \triangleq \sigma_i$. Further, from step 5 we have $x_1 = x_o + \sigma_o$ and $x_2 = x_1 + \sigma_1 = x_o + \sigma_o + \sigma_1$, etc., so that

$$x_n = x_o + \sum_{i=0}^{n-1} \sigma_i \tag{22}$$

and so on. From equations (1), (21), and (22), we have

$$x^* = x_n - \alpha_n V^{(n)} g_n$$

$$= x_o + \sum_{i=0}^{n-1} \sigma_i - \alpha_n\left(V^{(n)} g_o + \sum_{i=0}^{n-1} \sigma_i\right).$$

Hence,

$$x^* = x_o - \alpha_n V^{(n)} g_o + (1 - \alpha_n) \sum_{i=0}^{n-1} \sigma_i . \tag{23}$$

Now let us consider

$$\|x^* - (- A^{-1}b)\| = \left\| A^{-1}b + x_o - \alpha_n V^{(n)} g_o + (1 - \alpha_n) \sum_{i=0}^{n-1} \sigma_i \right\|$$

$$= \left\| A^{-1}b + A^{-1}Ax_o - \alpha_n V^{(n)} g_o + (1 - \alpha_n) \sum_{i=0}^{n-1} \sigma_i \right\| .$$

Hence

$$\|x^* + A^{-1}b\| = \left\| (A^{-1} - \alpha_n V^{(n)}) g_o + (1 - \alpha_n) \sum_{i=0}^{n-1} \sigma_i \right\| . \tag{24}$$

In order to establish convergence, we must show that $\|x^* + A^{-1}b\|$ can be made small as $n \to \infty$. Let $S'_{x_o} = \overline{\text{conv}} \{x \in H : J(x) \leq J(x_o)\}$ as in chapter 1. Since it is known that S'_{x_o} is bounded, [25], we can prove the following:

<u>Lemma 2.2:</u> If $n(\alpha_n - 1) \to 0$ as $n \to \infty$ and there exist $\alpha, \beta > 0$ such th $\alpha I \leq V^{(n)} \leq \beta I$ and $\gamma_n \neq -1$ for all n, then $\left\| (1 - \alpha_n) \sum_{i=0}^{n-1} \sigma_i \right\| \to 0$ as $n \to \infty$.

<u>Proof:</u> $\|\sigma_i\| = \|\alpha_i s_i\| = \|- \alpha_i V^{(i)} g_i\|$ (by definition).

So,

$$\|\sigma_i\| = |\alpha_i|\|V^{(i)}(Ax_i + b)\| \le |\alpha_i|\left\{\|V^{(i)}\|\|A\|\|x_i\| + \|V^{(i)}\|\cdot\|b\|\right\} \qquad (25)$$

Since $x_i \in S'_{x_o}$ is a bounded set, $\|x_i\|$ is bounded and since $\alpha_i \to 1$ as $i \to \infty$, α_i is bounded. By hypothesis $\|V^{(i)}\| \le \beta$ and $\|A\| \le M$, so everything on the right side of (25) is independent of i and $\|\sigma_i\| \le L$ for some $L \ge 0$ and all i. Hence,

$$\left\|(1 - \alpha_n)\sum_{i=0}^{n-1}\sigma_i\right\| \le |(1 - \alpha_n)| \cdot L \cdot n \to 0$$

since $(\alpha_n - 1)n \to 0$ as $n \to \infty$.

Lemma 2.3: If g_o is an element of the smallest closed subspace containing the y_i's denoted by $\overline{S(y_i)}$, if the $V^{(n)}$'s are uniformly bounded, $\alpha_n \to 1$ as $n \to \infty$ and if $\gamma_n \neq -1$ for all n, then $\left\|\left(A^{-1} - \alpha_n V^{(n)}\right) g_o\right\| \to 0$ as $n \to \infty$.

Proof: By hypothesis there exist scalars β_k such that

$$g_o = \sum_{i=0}^{\infty}\beta_i y_i \quad \text{and so} \quad A^{-1}g_o = \sum_{i=0}^{\infty}\beta_i A^{-1}y_i = \sum_{i=0}^{\infty}\beta_i\sigma_i \quad \text{by (1.12).}$$

Consider

$$\begin{aligned}\|A^{-1}g_o - \alpha_n V^{(n)}g_o\| &= \left\|\left(A^{-1} - \alpha_n V^{(n)}\right)\sum_{i=0}^{\infty}\beta_i y_i\right\| \\ &\le \left\|\left(A^{-1} - \alpha_n V^{(n)}\right)\sum_{i=0}^{n-1}\beta_i y_i\right\| + \left\|\left(A^{-1} - \alpha_n V^{(n)}\right)\sum_{i=n}^{\infty}\beta_i\right. \\ &\le |1 - \alpha_n|\left\|\sum_{i=0}^{n-1}\beta_i\sigma_i\right\| + \left\|A^{-1} - \alpha_n V^{(n)}\right\| \cdot \left\|\sum_{i=n}^{\infty}\beta_i y_i\right\|\end{aligned}$$

Since $A^{-1}g_o = \sum_{i=0}^{\infty} \beta_i\sigma_i$ and $g_o = \sum_{i=0}^{\infty} \beta_i y_i$ we know $\left\| \sum_{i=0}^{n-1} \beta_i\sigma_i \right\|$ is bounded for all n and $\left\| \sum_{i=n}^{\infty} \beta_i y_i \right\| \to 0$ as $n \to \infty$. Since $\alpha_n \to 1$, we know that $\left|1 - \alpha_n\right| \to 0$. Hence, $\left\|A^{-1}g_o - \alpha_n V^{(n)} g_o\right\| \to 0$ as $n \to \infty$.

We can now assert the following:

Theorem 2.11: If $g_o \in \overline{S(y_i)}$, $\gamma_n \neq -1$ for all n, if the $V^{(n)}$ are uniformly bounded, and $(\alpha_n - 1)n \to 0$ as $n \to \infty$, then $\|x^* + A^{-1}b\| \to 0$ as $n \to \infty$.

Proof: By (24) we have

$$\|x^* + A^{-1}b\| = \left\|\left(A^{-1} - \alpha_n V^{(n)}\right) g_o + (1 - \alpha_n) \sum_{i=0}^{n-1} \sigma_i\right\|$$

$$\leq \left\|\left(A^{-1} - \alpha_n V^{(n)}\right) g_o\right\| + \left\|(1 - \alpha_n) \sum_{i=0}^{n-1} \sigma_i\right\|$$

and by lemma 2.3 the first term goes to zero. By lemma 2.2 the second term goes to zero.

In this chapter, we have established conditions under which two variations of the basic algorithm converge to the location of the minimum of a quadratic functional. These are given in theorems 2.10 and 2.11. In both of these theorems we are most interested in the convergence question for an infinite dimensional Hilbert space. In a finite dimensional space of dimension n, we see that for almost any collection of α_n's the algorithm converges to the location of the minimum in a finite number of steps. The conditions on the α_n's and the proof are given in the following theorem.

Theorem 2.12: If $\gamma_j \neq -1$ and $\alpha_j \neq 0$ for all $j = 0,1,\ldots,$ and if $\left(V^{(j)}\right)^{-1}$ exists for all j, then after at most $n + 1$ steps $x^* = -A^{-1}b$, where $n = \dim H$.

Proof: First we show that the y_i's form a linearly independent set if $r_i \neq \theta$. Assume that $\{y_i\}_{i=0}^{l}$ is linearly dependent for some l. Therefore, there exist scalars β_i such that

$$y_l = \sum_{i=0}^{l-1} \beta_i y_i \tag{26}$$

By (12) and theorem 2.9

$$\left(A^{-1} - V^{(j)}\right)y_j = r_j \neq \theta \quad j=0,1,2,\ldots,l - 1 \tag{27}$$

Moreover, by the fundamental property of $V^{(j)}$

$$\left(A^{-1} - V^{(j)}\right) y_i = \theta \quad \text{for} \quad i < j \tag{28}$$

By operating on (26) by $\left(A^{-1} - V^{(l)}\right)$ and applying (27) and (28) we have

$$r_l = \left(A^{-1} - V^{(l)}\right) y_l = \sum_{i=0}^{l-1} \beta_i\left(A^{-1} - V^{(l)}\right)y_i = \sum_{i=0}^{l-1} \beta_i \theta = \theta.$$

If $\{y_i\}_{i=0}^{l}$ are linearly dependent then $r_l = \theta$. Therefore, by step 4 of the algorithm α_l is reset to 1 and by theorem 2.9

the resulting x^* is the location of the minimum. Hence, the theorem is true, if $\{y_i\}_{i=0}^{l}$ are linearly dependent for $l < n$.

Since H is finite dimensional of dimension n, we have at most n linearly independent y's. Now, if we apply the algorithm n times and the resulting $r_n \neq \theta$, we have generated n linearly independent y's and they must form a basis for H. Moreover, by the fundamental property of $V^{(n)}$, i.e., theorem 2.4 and its corollary, we have $V^{(n)}y_i = A^{-1}y_i$ $i=0,1,2,\dots,n-1$. Since the two linear operators $V^{(n)}$ and A^{-1} agree on the y_i's, a basis for the space, it must be that

$$V^{(n)} = A^{-1} \text{ on the whole space.} \tag{29}$$

Hence, by definition of x^*, (29) and (1.10) we have

$$x^* = x_n - \alpha_n V^{(n)} g_n = x_n - \alpha_n A^{-1} g_n = x_n - \alpha_n x_n - \alpha_n A^{-1} b. \tag{30}$$

Now from (3)

$$\begin{aligned} r_n &= V^{(n)}(g^* - g_n) + \alpha_n V^{(n)} g_n \\ &= A^{-1}(g^* - g_n) + \alpha_n A^{-1} g_n && \text{by (29)} \\ &= x^* - x_n + \alpha_n A^{-1} g_n && \text{by (1.12)} \\ &= x^* - x_n + \alpha_n A^{-1}(Ax_n + b) && \text{by (1.6)} \\ &= x^* - x_n + \alpha_n x_n + \alpha_n A^{-1} b \\ &= \theta && \text{by (30)} \end{aligned}$$

So by step 4 of the algorithm α_n is reset to one and by theorem 2.9 x^* is the location of the minimum.

Many times in this chapter we have proved results dependent upon $\gamma_n \neq - 1$. We shall continue to do this in subsequent chapters. For this reason, we shall investigate the case of $\gamma_n \neq - 1$. From (6) and (7) we have $- 1 = \gamma_n = - \frac{(g_n, r_n)}{(g^*, r_n)}$ which implies that

$$(y_n, r_n) = 0. \tag{31}$$

Now we know from theorem 2.9 that if $(V^{(n)})^{-1}$ exists and (31) holds because $r_n = \theta$ that convergence is achieved on the next iteration with $\alpha_n = 1$. Also if (31) holds because $y_n = \theta$ then $g^* = g_n$ and by (1.7) then $Ax^* + b = Ax_n + b$ or $x^* = x_n$. But if $V^{(n)} > 0$ this contradicts (1) since $g_n \neq \theta$.

Now by (5) and (1.17) $r_n = (V^{(n)} - A^{-1})y_n$, hence, (31) can be written as

$$\left(y_n, \left(V^{(n)} - A^{-1}\right)y_n\right) = 0 \tag{32}$$

and, if $V^{(n)} > A^{-1}$ or $V^{(n)} < A^{-1}$ then (32) is impossible for $y_n \neq \theta$. Theorem 2.6 states that if $V^{(o)} \geq A^{-1}$ or $V^{(o)} \leq A^{-1}$ then $V^{(n)} \leq A^{-1}$ or $V^{(n)} \geq A^{-1}$ for all n.

Moreover, the convergence of the iterates to the location of the minimum of a quadratic functional assured by theorem 2.10 and its corollary is independent of γ_n. Hence, if $\gamma_n = - 1$ then α_n should be computed by (17).

3. COMPARISON WITH OTHER CONJUGATE GRADIENT TECHNIQUES

If the functional to be minimized is quadratic as discussed in chapter 1, then Myers [27] and Horwitz and Sarachik [20] have shown that whenever $H^{(o)} = I$ the DFP technique generates the same search directions as those given by the conjugate gradient method. Here, we shall examine the relationship between these two methods mentioned above and the method discussed in chapter 2 with α_n chosen as in (2.17)assuming that the functional to be minimized is quadratic. That is, throughout chapter 3 we shall assume that α_n satisfies $J(x_n + \alpha_n s_n) \leq J(x_n + \lambda s_n)$ for all real λ, and that $J(x) = J_o + (b,x) + \frac{1}{2}(x,Ax)$, as in chapter 1.

Theorem 3.1: If $\gamma_i \neq -1$ for all i, then the $\{\sigma_i\}$ generated by the algorithm outlined in chapter 2 are A conjugate and the $\{y_i\}$ are A^{-1} conjugate, i.e.,

$$(\sigma_i, A\sigma_j) = (y_i, A^{-1}y_j) = (\sigma_j, y_i) = 0 \quad \text{if} \quad i \neq j, \qquad (1)$$

$$(g_k, s_i) = \begin{cases} 0 & \text{if} \quad 0 \leq i < k \\ (g_i, s_i) & \text{if} \quad 0 \leq k \leq i, \end{cases} \qquad (2)$$

and also

$$V^{(k)} A\sigma_i = \sigma_i \qquad \text{holds for all} \qquad i < k. \qquad (3)$$

Proof: (By mathematical induction) By 2.4 $r_n = V^{(n)}y_n - \alpha_n s_n$ so that

$$\sigma_n = V^{(n)}y_n - r_n \qquad (4)$$

where $\sigma_n = \alpha_n s_n$.

By (1.12) $A\sigma_o = y_o$, so that

$$V^{(1)}A\sigma_o = V^{(1)}y_o$$

$$V^{(1)}A\sigma_o = V^{(o)}y_o - \frac{(r_o,y_o)r_o}{(r_o,y_o)} \qquad \text{(by (2.9) and theorem 2.5)}$$

$$= \sigma_o \qquad \text{(by (4) with } n = 0\text{)}$$

Hence, $(\sigma_o,A\sigma_1) = (\sigma_o,A(-\alpha_1V^{(1)}g_1) = -\alpha_1(V^{(1)}A\sigma_o,g_1)$ since A and $V^{(1)}$ are self-adjoint. Therefore, $(\sigma_o,A\sigma_1) = -\alpha_1(\sigma_o,g_1)$ since $V^{(1)}A\sigma_o = \sigma_o$. Hence, $(\sigma_o,A\sigma_1) = -\alpha_1.0$ since α_1 was chosen to be the minimum in the direction s_n, $(\sigma_o,g_1) = 0$ by (2.18). Hence, the theorem is true for $k = 1$. We shall now assume that $(\sigma_j,A\sigma_i) = 0$ if $0 \leq j < i \leq k$ and $V^{(k)}A\sigma_i = \sigma_i$ if $0 \leq i < k$. By (1.7) and (2.1),

$$g_k = b + Ax_k = b + A(x_{k-1} + \sigma_{k-1})$$

$$= b + A(x_{i+1} + \sigma_{i+1} + \dots + \sigma_{k-1})$$

$$= g_{i+1} + A\sigma_{i+1} + \dots + A\sigma_{k-1}.$$

Therefore,

$$(\sigma_i, g_k) = (\sigma_i, g_{i+1}) + (\sigma_i, A\sigma_{i+1}) + \dots + (\sigma_i, A\sigma_{k-1})$$

$$= 0 + 0 \dots + 0 = 0$$

that is,

$$(\sigma_i, g_k) = 0 \qquad (5)$$

by choice of α_n since $(\sigma_i, g_{i+1}) = 0$, and the other terms are zero by the induction hypothesis. So we have established the first part of (2) for $i < k$.

Now for $i < k$ we can see that

$$(\sigma_i, A\sigma_k) = (\sigma_1, -\alpha_k A V^{(k)} g_k) \qquad \text{(by (2.1))}$$

$$(\sigma_i, A\sigma_k) = \alpha_k (V^{(k)} A\sigma_i, g_k) \qquad \text{(since } A \text{ and } V^{(k)} \text{ are self-adjoint)}$$

$$(\sigma_i, A\sigma_k) = -\alpha_k (\sigma_i, g_k) = 0 \qquad (6)$$

by the induction bypothesis and (5). For a quadratic functional, $A\sigma_i = y_i$ by (1.12), hence by substitution into (6) we have proved (1).

We consider for $i < k$

$$V^{(k+1)} A\sigma_i = V^{(k)} A\sigma_i - \frac{(r_k, A\sigma_i) r_k}{(r_k, y_k)} \qquad \text{(by def. of } V^{(k+1)}\text{)}$$

$$V^{(k+1)} A\sigma_i = \sigma_i - \frac{(V^{(k)} y_k - \sigma_k, A\sigma_i) r_k}{(r_k, y_k)} \qquad \text{(by def. of } r_k\text{)}$$

$$= \sigma_i - \frac{(V^{(k)}y_k, y_i)r_k}{(r_k, y_k)} \qquad \text{(since } (\sigma_k, A\sigma_i) = 0\text{)}$$

$$= \sigma - \frac{(y_k, V^{(k)}y_i)r_k}{(r_k, y_k)} \qquad \text{(by theorem 2.1)}$$

$$= \sigma - \frac{(y_k, A^{-1}y_i)r_k}{(r_k, y_k)} \qquad \text{(by the corollary to theorem 2.4)}$$

$$= \sigma_i - \frac{(A\sigma_k, \sigma_i)r_k}{(r_k, y_k)} = \sigma_i \qquad \text{(since } A\sigma_k = y_k \text{ and } (A\sigma_k, \sigma_i) = 0 \text{ for } i < k\text{)}$$

Moreover, by (2.9) and (4)

$$V^{(k+1)}A\sigma_k = V^{(k+1)}y_k = V^{(k)}y_k - \frac{(r_k, y_k)r_k}{(r_k, y_k)}$$

$$= V^{(k)}y_k - r_k = \sigma_k.$$

Hence we have established 1, 2, and 3, and the first half of 4. We know that $x_k = x_{k+1} - \sigma_k$ and hence, $x_k = x_i - \sigma_{i-1} - \dots - \sigma_k$ for $i > k$. Then $g_k = Ax_k + b = Ax_i + b - A(\sigma_i + \dots + \sigma_k)$. Hence,

$$g_k = g_i - A(\sigma_{i-1} + \dots + \sigma_k), \text{ so}$$

$$(g_k, s_i) = (g_i, s_i) - \frac{1}{\alpha_i} \sum_{l=0}^{i-k} (A\sigma_{i-j}, \sigma_i)$$

$$= (g_i, s_i) - \frac{1}{\alpha_i} \sum_{j=0}^{i-k} 0 = (g_i, s_i) \qquad \text{for} \quad k \leq i.$$

We see from the preceeding theorem that this method is a conjugate direction method. In light of the remarks at the beginning of this chapter, the question arises as to how our method is related to the conjugate gradient and DFP techniques. Since our method is a conjugate direction method we must have, if $\gamma_n \neq - 1$ for all n, that the σ_n's are linearly independent. For if the $\{\sigma_n\}_{n=0}^{l}$ are linearly dependent then there exist scalars such that

$$\sum_{i-0}^{l} \beta_i \sigma_i = \theta \,. \tag{7}$$

So if $j < l$ we have from (7) that $0 = \sum_{i=0}^{l} (\sigma_j, A\sigma_i)$, which implies that $\beta_j(\sigma_j, A\sigma_j) = 0$. Hence, $\beta_j = 0$ since $\sigma_j \neq 0$ and since A is strongly positive. Since $\sigma_n = \alpha_n s_n$ the s_n's are linearly independent.

Notice also that $V^{(o)}g_o = V^{(o)} \cdot 1 \cdot g_o$. If we choose $c_{oo} = 1$ then $V^{(o)}g_o = V^{(o)}\{c_{oo}g_o\}$ and

$$V^{(1)}g_1 = V^{(o)}g_1 - \frac{(r_o, g_1)}{(r_o, y_o)}V^{(o)}(g_1 - (1 - \alpha_o)g_o)$$

$$V^{(1)}g_1 = V^{(o)} \left\{\left(1 - \frac{(r_o, g_1)}{(r_o, y_o)}\right)g_1 + \frac{(1 - \alpha_o)(r_o, g_1)}{(r_o, y_o)} g_o\right\}$$

that is, $V^{(1)}g_1 = V^{(o)} \sum_{i=0}^{1} c_{i1}, g_i$ for scalars $c_{o1} = \dfrac{(1 - \alpha_o)(r_o, g_1)}{(r_o, y_o)}$ and $c_{11} = 1 - \dfrac{(r_o, g_1)}{(r_o, y_o)}$. The above suggests that for every n there exist scalars c_{in}, $i = 0, 1, \ldots , n$, such that

$$V^{(n)}g_n = V^{(o)} \sum_{i=0}^{n} c_{in}g_i \tag{8}$$

We shall now establish (8) and find a convenient way to express the c_{in}'s.

Theorem 3.2: If $\gamma_j \neq -1$ for $j = 0,1,2,\ldots,$ then for every integer k, there exist scalars a_{ik}, b_{ik}, c_{ik} $i = 0,1,\ldots,k$ such that

$$V^{(k)}y_k = V^{(o)}\left[\sum_{i=0}^{k} a_{ik}y_i + \sum_{i=0}^{k-1} b_{ik}g_i\right] \tag{9}$$

and

$$V^{(k)}g_k = V^{(o)} \sum_{i=0}^{k} c_{ik}g_i \tag{10}$$

where $y_k = g_{k+1} - g_k$.

Proof: (By mathematical induction)

$$V^{(o)}y_o = V^{(o)}(1)y_o, \qquad \text{so} \qquad a_{oo} = 1.$$

$$V^{(1)}y_1 = V^{(o)}y_1 - \frac{(r_o,y_1)r_o}{(r_o,y_o)} \qquad \text{by (2.10)}$$

$$= V^{(o)}y_1 - \frac{(r_o,y_1)}{(r_o,y_o)}\left[V^{(o)}y_o + \alpha_o V^{(1)}g_o\right] \quad \text{by (2.5)}$$

$$= V^{(o)}\left[\sum_{i=0}^{1} a_{i1}y_i + \sum_{i=0}^{0} b_{i1}g_i\right]$$

where $a_{11} = 1$, $a_{01} = -\frac{(r_o,y_1)}{(r_o,y_o)}$, and $b_{01} = -\alpha_o\frac{(r_o,y_1)}{(r_o,y_o)}$. Moreover,

$$V^{(1)}g_1 = \sum_{i=0}^{1} c_{i1}g_i$$

where

$$c_{11} = \left[\frac{1 - (r_o,g_1)}{(r_o,y_o)}\right], c_{01} = \frac{(1 - \alpha_o)(r_o,g_1)}{(r_o,y_o)}$$

as shown in the previous paragraph. The induction assumption is: there exist a_{ji}, b_{ji}, and c_{ji}, $i = 0,1,2, \ldots,k$ $j = 0,1,2, \ldots,i$. Such that,

$$V^{(i)}y_i = V^{(o)}\left[\sum_{j=0}^{i} a_{ji}y_j + \sum_{j=0}^{i-1} b_{ji}g_j\right] \tag{11}$$

and

$$V^{(i)}g_i = V^{(o)}\left[\sum_{j=0}^{i} c_{ji}g_j\right]. \tag{12}$$

Since

$$V^{(k+1)}y_{k+1} = V^{(k)}y_{k+1} - \frac{(r_k,y_{k+1})}{(r_k,y_k)}(V^{(k)}y_k + \alpha_k V^{(k)}g_k) \qquad \text{(by (2.5) and (2.10))}$$

$$= V^{(o)}y_{k+1} - \sum_{i=0}^{k-1} \frac{(r_i,y_{k+1})}{(r_i,y_i)}(V^{(i)}y_i + \alpha_i V^{(i)}g_i)$$

$$- \frac{(r_k,y_{k+1})}{(r_k,y_k)}(V^{(k)}y_k + \alpha_k V^{(k)}g_k) \qquad \text{(by (2.9))}$$

We have

$$\begin{aligned}
V^{(k+1)}y_{k+1} = V^{(o)}y_{k+1} &- \sum_{i=0}^{k-1} \frac{(r_i,y_{k+1})}{(r_i,y_i)} \left[V^{(o)} \sum_{j=0}^{i} a_{ji}y_i + \sum_{j=0}^{i-1} b_{ji}g_i \right] \\
&- \sum_{i=0}^{k-1} \frac{(r_i,y_{k+1})}{(r_i,y_i)} \alpha_i V^{(o)} \left[\sum_{j=0}^{i} c_{ji}g_i \right] \\
&- \frac{(r_k,y_{k+1})}{(r_k,y_k)} V^{(o)} \left[\sum_{i=0}^{k} a_{ik}y_i + \sum_{i=0}^{k-1} b_{ik}g_i \right] \\
&- \alpha_k \frac{(r_k,y_{k+1})}{(r_k,y_k)} V^{(o)} \left[\sum_{i=0}^{k} c_{ik}g_i \right] \qquad \text{(by (11) and (12))}
\end{aligned}$$

Hence,

$$\begin{aligned}
V^{(k+1)}y_{k+1} &= V^{(o)} \left[y_{k+1} - \sum_{i=0}^{k} \frac{(r_i,y_{k+1})}{(r_i,y_i)} \sum_{j=0}^{i} a_{ji}y_j \right. \\
&\qquad \left. - \sum_{i=0}^{k} \frac{(r_i,y_{k+1})}{(r_i,y_i)} \sum_{j=0}^{i-1} b_{ji}g_i \right] \\
&= V^{(o)} \left[\sum_{i=0}^{k+1} a_{ik+1}y_i + \sum_{i=0}^{k} b_{ik+1}g_i \right].
\end{aligned}$$

Therefore, (9) is established for $k + 1$, if (11) and (12) hold for k.

Also

$$V^{(k+1)}g_{k+1} = V^{(k+1)}y_k + V^{(k+1)}g_k \qquad (\text{Since } \; y_k = g_{k+1} - g_k)$$

$$= V^{(k)}y_k - \frac{(r_k,y_k)}{(r_k,y_k)}\left[V^{(k)}y_k + \alpha_k V^{(k)}g_k\right] + V^{(k+1)}g_k \quad (\text{by } (2.10))$$

$$V^{(k+1)}g_{k+1} = \alpha_k V^{(k)}g_k + V^{(k+1)}g_k. \qquad (13)$$

Now let us consider

$$V^{(k+1)}g_k = V^{(k)}g_k - \frac{(r_k,g_k)}{(r_k,y_k)}\left(V^{(k)}y_k + \alpha_k V^{(k)}g_k\right) \qquad (\text{by } (2.10))$$

$$V^{(k+1)}g_k = V^{(o)}\left[\sum_{i=0}^{k} c_{ik}g_i - \frac{(r_k,g_k)}{(r_k,y_k)}\left(\sum_{i=0}^{k} a_{ik}y_i + \sum_{i=0}^{k-1} b_{ik}g_i + \alpha_k \sum_{i=0}^{k} c_{ik}g_i\right)\right] \qquad (14)$$

(by (11) and (12))

Using $y_i = g_{i+1} - g_i$ in (14) and substituting that back into (13) and applying (12), we have

$$V^{(k+1)}g_{k+1} = V^{(o)}\left[\alpha_k \sum_{i=0}^{k} c_{ik}g_i + \sum_{i=0}^{k} c_{ik}g_i - \frac{(r_k,g_k)}{(r_k.y_k)}\left(\sum_{i=0}^{k} a_{ik}g_{i+1} + \sum_{i=0}^{k} (a_{ik} + \alpha_k c_{ik})g_i + \sum_{i=0}^{k-1} b_{ik}g_i\right)\right].$$

Hence, (10) is established for $k + 1$, and the theorem is provided.

In order to establish the relationship between the three conjugate direction methods we wish to find an expression for c_{ik} in terms of the g_i's and $V^{(o)}$. From (2), $(g_k, s_i) = 0$ if $i < k$. Hence. $-(g_k, V^{(i)}g_i) = 0$ if $i < k$. From (8) we have

$$(g_k, V^{(i)}g_i) = \sum_{l=0}^{i} c_{li}(g_k, V^{(o)}g_l) = 0: \tag{15}$$

Let us fix $k \geq 1$ and notice that if $i = 0$, and since (15) $-s_o = V^{(o)}g_o = V^o(1)g_o$, we have $c_{oo} = 1$. Hence by (15) with $i = 0$ we have that $(g_k, V^{(o)}g_o) = 0$, but this is also true from (2), since $V^o g_o = - s_o = - \frac{1}{\alpha_o} \sigma_o$. We consider (15) with $i = 1$ and have

$$0 = c_{01}(g_k, V^{(o)}g_o) + c_{11}(g_k, V^{(o)}g_1) = c_{11}(g_k, V^{(o)}g_1)$$

since $(g_k, V^o g_o) = 0$ by (1). Now if $c_{11} = 0$, then $s_1 = - c_{01}V^{(o)}g_o = c_{01}s_o$, but we observed before that s_o, s_1 are linearly independent. So it must be true that $(g_k, V^{(o)}g_1) = 0$. Moreover from (8), we have

$$- s_1 = c_{01}s_o + c_{11}V^{(o)}g_1, \text{ so } V^{(o)}g_1 \in S(s_o, s_1).$$

By induction

$$V^{(o)}g_o, V^{(o)}g_1, \ldots, V^{(o)}g_{n-1} \in S(s_o, s_1, \ldots, s_{n-1}) \tag{16}$$

where $S(s_o, s_1, \ldots, s_{n-1})$ denotes the subspace spanned by the s_i's. Let us assume that $(g_k, V^{(o)}g_l) = 0$ for all $l = 1,2,\ldots,n - 1$ for $n < k$. By (2) and the induction hypothesis we have

$$0 = (g_k, V^{(n)} g_n) = \sum_{l=0}^{n} c_{ln}(g_k, V^{(o)} g_l) = c_{nn}(g_k, V^{(o)} g_n).$$

If $c_{nn} = 0$ we must have from (9) that $-s_n = \sum_{l=0}^{n-1} V^{(o)} c_{ln} g_l$ so that $s_n \in S\left(\left\{V^{(o)} g_i\right\}_{i=0}^{n-1}\right) \subseteq S(s_o, s_1, \ldots s_{n-1})$ by (16). But this implies that $\left\{s_i\right\}_{i=0}^{n}$ is a linearly dependent set of vectors which contradicts the remarks following theorem 3.1. Hence, $(g_k, V^{(o)} g_l) = 0$ for all $l = 0,1,\ldots,k - 1$ and we have from (2), if $0 \leq l \leq i$, that $(V^{(i)} g_i, g_l) = -(g_l, s_i) = (g_i, V^{(i)} g_i)$. Therefore

$$(V^{(i)} g_i, g_l) = \sum_{j=0}^{i} c_{ji}(g_j, V^{(o)} g_l)$$

and for all $j \neq l$ $(g_j, V^{(o)} g_l) = 0$. So we have

$$(g_i, V^{(i)} g_i) = c_{li}(g_l, V^{(o)} g_l).$$

Hence, $c_{li} = \dfrac{(g_i, V^{(i)} g_i)}{(g_l, V^{(o)} g_l)}$ which implies that $-s_i = V^{(i)} g_i$ $= V^{(o)} \sum_{l=0}^{i} \dfrac{(g_i, V^{(i)} g_i)}{(g_l, V^{(o)} g_l)} g_l$. Therefore

$$- s_i = (g_i, V^{(i)} g_i) V^{(o)} \sum_{l=0}^{i} \frac{g_l}{(g_l, V^{(o)} g_l)} \tag{17}$$

Hence, we can state the following theorem.

Theorem 3.3: If $\gamma_n \neq -1$ for all n, α_n is chosen as in (2.17) and $V^{(o)} = H^{(o)}$ of the DFP method, then the search directions of the DFP and the Davidon-Broyden method with α_n chosen by (2.17) are the same. Moreover, if $V^{(o)} = H^{(o)} = I$, then these search directions are the same as those of the conjugate gradient method.

Proof: Horwitz and Sarachik [20] have shown that for the DFP method, the ith search direction is given by

$$- H^{(o)} (g_i, H^{(i)} g_i) \sum_{l=0}^{i} \frac{g_l}{(g_l, H^{(o)} g_l)}$$

If $H^{(o)} = V^{(o)}$ it follows from (17) that the directions are the same. In [19] it was shown that for the method of conjugate gradients, the ith search direction is

$$- \| g_i \|^2 \sum_{l=0}^{i} \frac{g_l}{\| g_l \|^2}.$$

At each point x_n the three methods generate a direction s_n then the stepsize is chosen so that the function $J(x_n + \lambda s_n)$ is minimized with respect to λ. Since the directions are the same and the stepsize is chosen in the same fashion for each method, the sequences of iterates generated by these methods $x_o, x_1, x_2, \ldots,$

will be the same. Again, we restate that throughout chapter 3, the functional to be minimized is quadratic as outlined in chapter 1.

It is well known [22] that the rate of convergence to the minimum of a quadratic functional for the method of steepest descent is given by

$$(J(x_i) - J(-A^{-1}b)) \leq \left(\frac{M - m}{M + m}\right)^i (J(x_o) - J(-A^{-1}b)) \ ,i = 1,2,\ldots \quad (18)$$

where m and M are given by (1.3). Daniel [6] has established that the rate of convergence for the conjugate gradient algorithm is given by

$$(J(x_i) - J(-A^{-1}b)) \leq 4\left(\frac{\left(1 - \sqrt{\frac{m}{M}}\right)^2}{\left(1 + \sqrt{\frac{m}{M}}\right)^2}\right)^i (J(x_o) - J(-A^{-1}b)), \ i= 1,2,.(19)$$

(17) is obviously a faster rate of convergence than (18).

Under the conditions of theorem 3.3 with $V^{(o)} = I$, we know that the iterates generated by our algorithm and those of the method of conjugate gradients are the same. Hence, the rate of convergence of our algorithm to the minimum is given by (19) and we have the following theorem:

Theorem 3.4: If for each n, α_n is chosen by (2.17), $\gamma_n \neq -1$ and $V^{(o)} = I$, then the rate of convergence for the algorithm outlined in chapter 2 is given by (19).

4. EXTENSION OF POWELL'S IDEA

In this chapter we shall extend an idea of Powell [30], concerning the basic algorithm as outlined in chapter 2; to a separable infinite dimensional Hilbert space. The idea is to use the rank one algorithm of chapter 2, but with search directions which are independent of the gradient. Specifically, we wish to compute the location of the minimum of a differentiable functional $J:H \to R$. We let $V^{(o)}$ be a strongly positive, self-adjoint, bounded linear operator, as in chapter 2, and let $x_o \in H$ be the initial estimate of the location of the minimum. Further, let p be an arbitrary integer. If the dimension of H is finite, it is advantageous to let $p = \dim H$. Let $\sum = \{\sigma_i\} \subseteq H$ represent a basis for H. Compute $J(x_o)$ and g_o, and proceed as follows.

<u>Step 1:</u> Let

$$x^* = x_n + \sigma_n, \tag{1}$$

and compute $J(x^*)$ and g^*. If $\|g^*\| = 0$ then x^* satisfies the necessary condition for a minimum, and we stop. Otherwise,

<u>Step 2:</u> Compute the residual vector as in chapter 2. Let $r_n = V^{(n)}y_n - \sigma_n$ where $y_n = g^* - g_n$ and compute the scalars

$$\left.\begin{aligned} \rho_n &\doteq (g^*,r_n) \\ \gamma_n &= -\frac{(g_n,r_n)}{\rho_n} \\ \lambda_n &= \begin{cases} \dfrac{\gamma_n}{(\gamma_n + 1)} & \text{if } \gamma_n \neq -1 \\ 1 & \text{if } \gamma_n = -1 \end{cases} \end{aligned}\right\} \tag{2}$$

Step 3: If $V^{(n)}y_n = \sigma_n$, let $V^{(n+1)} = V^{(n)}$, otherwise let

$$V^{(n+1)} = V^{(n)} + \frac{(\lambda_n - 1)}{\rho_n} B^{(n)} \tag{3}$$

where $B^{(n)}:H \to H$ is defined such that for all $x \in H$

$$B^{(n)}x = (x,r_n)r_n. \tag{4}$$

Step 4: If $J(x^*) \leq J(x_n)$, let $x_{n+1} = x^*$. Otherwise, let $x_{n+1} = x_n$. If $n = pk$ for some integer k, then let

$$z_k = x^o - V^{(n)}g_o. \tag{5}$$

Evaluate $J(z_k)$ and $g(z_k)$ and if $\|g(z_k)\| = 0$ stop. Otherwise, return to step 1.

We shall show that z_k converges in norm as $k \to \infty$ to the location of the minimum of a quadratic functional. For an infinite dimensional Hilbert space, we determine the frequency with which we apply the Newton-like iteration $z_k = x_o - V^{(k)}g_o$ with $pk = n$. With this modification of the basic algorithm, we can prove many theorems which are analogous to those of chapter 2. Henceforth, as in chapter 2, we shall assume that the functional to be minimized is quadratic. That is,

$$J(x) = J_o + (b,x) + \frac{1}{2}(x,Ax) \tag{6}$$

where A is as in (1.3). Theorems 4.1, 4.2 and 4.4 are independent of the type of functional being minimized.

Theorem 4.1: $B^{(n)}$, as defined in (4), is a self-adjoint positive operator for all n.

Proof: As in chapter 2.

Theorem 4.2: $V^{(n)}$ is self-adjoint for all n.

Proof: As in chapter 2.

With the next two theorems we see that the properties of $V^{(n)}$ given in theorems 2.3, 2.4, and their corollaries hold even though σ_n is a prescribed vector independent of $V^{(n)}$, α_n, and g_n.

Theorem 4.3: If $A^{-1}u = V^{(n)}u$ for some $u \in H$ and $B:H \to H$ is such that there exists some scalar μ such that $B - V^{(n)} = \mu B^{(n)}$ then $A^{-1}u = Bu$.

Proof: By (1.12) we know that $A^{-1}y_n = x^* - x_n = \sigma_n$ and by def. $r_n = V^{(n)}y_n - \sigma_n = \left(V^{(n)} - A^{-1}\right)y_n$. If $\left(V^{(n)} - A^{-1}\right)u = \theta$, then $(r_n,u) = \left(\left(V^{(n)} - A^{-1}\right)y_n,u\right) = \left(y_n,\left(V^{(n)} - A^{-1}\right)u\right) = (y_n,\theta) = 0$. Hence, if $B - V^{(n)} = \mu B^{(n)}$, $\left(B - V^{(n)}\right)u = \mu(r_n,u)r_n = \mu \cdot 0 \cdot r_n = \theta$ Since, by hypothesis $V^{(n)}u = A^{-1}u$, we have $Bu = A^{-1}u$.

Corollary: If $V^{(n)}u = A^{-1}u$ then $V^{(n+1)}u = A^{-1}u$.

Theorem 4.4: $V^{(n+1)}y_n = \sigma_n$, if $\gamma_n \neq -1$.

Proof: If $V^{(n)}y_n = \sigma_n$, then $V^{(n+1)} = V^{(n)}$ by step 3 and the theorem is obvious. Otherwise, using (5) and (6) we have

$$V^{(n+1)}y_n - \sigma_n = V^{(n)}y_n + \frac{(\lambda_n - 1)}{\rho_n}(r_n,y_n)r_n - \sigma_n.$$

Hence, using (1), (2), and (4)

$$V^{(n+1)}y_n - \sigma_n = r_n\left(1 + \frac{(\lambda_n - 1)}{\rho_n}(r_n,y_n)\right) = r_n \cdot 0 = \theta$$

Corollary: $V^{(n)}y_i = \sigma_i$ for $i < n$, if $\gamma_i \neq - 1$.

Theorem 4.5: If $\gamma_n \neq - 1$ then $(\lambda_n - 1)/\rho_n = - (r_n,y_n)^{-1}$.

Proof: Formally the same as the proof of the corresponding theorem in chapter 2 in spite of the change in the definition of σ_n.

Theorem 4.6: If $V^{(o)} \geq A^{-1}$, then $V^{(o)} \geq V^{(1)}, \ldots, \geq V^{(n)} \geq, \ldots, \geq A^{-1}$ and similarly, if $V^{(o)} \leq A^{-1}$, then $V^{(o)} \leq V^{(1)} \leq, \ldots, \leq V^{(n)} \leq, \ldots, \leq A^{-1}$.

Proof: Formally follows the proofs of theorems 2.6 and 2.7 and is based on theorem 4.5, $A^{-1}y_n = \sigma_n$, that is, (1.12) and the Schwarz inequality [2]. If $x \in H$ and $V^{(o)} \geq A^{-1}$ and $V^{(n)} \geq A^{-1}$, then by (2.10), (2.15), and the Schwarz inequality

$$\left(x,\left(V^{(n+1)} - A^{-1}\right)x\right) = \left(x,\left(V^{(n)} - A^{-1}\right)x\right) - \frac{\left(x,\left(V^{(n)} - A^{-1}\right)y_n\right)}{\left(y_n,\left(V^{(n)} - A^{-1}\right)y_n\right)} \geq 0.$$

Also from (2.10) $\left(x,\left(V^{(n+1)} - V^{(n)}\right)x\right) = - \dfrac{(x,r_n)^2}{\left(y_n,\left(V^{(n)} - A^{-1}\right)y_n\right)} \leq 0.$

We now wish to establish a convergence theorem for this modification of the basic algorithm. Since the set $\sum$ is a basis for H, for each $x \in H$, there exist scalars $c_i \in R$, $i = 0,1,\ldots$ such that

$$A^{-1}x = \sum_{i=0}^{\infty} c_i\sigma_i \tag{7}$$

or $x = \sum_{i=0}^{\infty} c_i A\sigma_i$. Since it is known that $A\sigma_i = y_i$ (1.12) where $y_i = g^* - g_i$, we have

$$x = \sum_{i=0}^{\infty} c_i y_i. \tag{8}$$

Then

$$V^{(n)}x = V^{(n)} \sum_{i=0}^{\infty} c_i y_i. \tag{9}$$

By the corollary to theorem 4.4, if $\gamma_j \neq -1$, $j = 0,1,2,\ldots,n$, we have $V^{(n)}y_i = \sigma_i$ for all $i < n$. So (9) becomes

$$V^{(n)}x = \sum_{i=0}^{n-1} c_i\sigma_i + V^{(n)} \sum_{i=n}^{\infty} c_i y_i. \tag{10}$$

Therefore, by (7) and (10) we have

$$\begin{aligned} \| A^{-1}x - V^{(n)}x\| &= \left\| \sum_{i=0}^{\infty} c_i\sigma_i - \sum_{i=0}^{n-1} c_i\sigma_i - V^{(n)}\left[\sum_{i=n}^{\infty} c_i y_i\right]\right\| \\ &= \left\| \sum_{i=n}^{\infty} c_i\sigma_i - V^{(n)} \sum_{i=n}^{\infty} c_i y_i\right\| \\ &= \left\| A^{-1}\left[\sum_{i=n}^{\infty} c_i y_i\right] - V^{(n)}\left[\sum_{i=n}^{\infty} c_i y_i\right]\right\| \\ &\leq \|A^{-1} - V^{(n)}\| \left\| \sum_{i=n}^{\infty} c_i y_i\right\| \end{aligned} \tag{11}$$

If the $V^{(n)}$ are uniformly bounded, then $\|A^{-1} - V^{(n)}\|$ is bounded. By (8) $\left\|\sum_{i=n}^{\infty} c_i y_i\right\| \to 0$ as $n \to \infty$ for this is the nth remainder of the expansion of x in terms of the y_i's. Therefore, we have $\|A^{-1}x - V^{(n)}x\| \to 0$ as $n \to \infty$. Hence, we have the following:

<u>Theorem 4.1:</u> If $V^{(n)}$ are uniformly bounded and $\gamma_n \neq -1$ for all n then $V^{(n)} \to A^{-1}$ pointwise.

<u>Corollary:</u> If $z_k = x_o - V^{(n)}g_o$ where $pk = n$, then z_k converges to the location of the minimum as $k \to \infty$.

<u>Proof:</u> In chapter 1 it was shown that the location of the minimum of the quadratic functional is $-A^{-1}b$. Hence

$$\begin{aligned} \|z_k + A^{-1}b\| &= \|x_o - V^{(n)}g_o + A^{-1}b\| \\ &= \|x_o - V^{(n)}(Ax_o + b) + A^{-1}b\| \\ &\leq \|x_o - V^{(n)}Ax_o\| + \|A^{-1}b - V^{(n)}b\| \qquad (12) \end{aligned}$$

(by (1.7))

By theorem 4.1 $V^{(n)}(Ax_o) \to A^{-1}(Ax_o) = x_o$ and $V^{(n)}b \to A^{-1}b$. Hence, $z_k \to -A^{-1}b$ as $k \to \infty$.

The above theorem and its corollary establish the convergence to the location of the minimum of the quadratic functional for this modification of the algorithm. As noted earlier, the search directions here are prescribed and are independent of α_n, g_n, and $V^{(n)}$. The rate of convergence could perhaps be improved by letting $z_k = x_n - V^{(n)}g_n$ where $pk = n$.

Notice, if H is finite dimensional then z_1 is the location of the minimum. This follows since by theorems 4.3 and 4.4 and their corollaries A^{-1} and $V^{(p)}$ agree on $\sigma_0, \sigma_1, \sigma_2, \dots, \sigma_{p-1}$, a basis for H. Hence $V^{(p)} = A^{-1}$. Therefore,

$$\begin{aligned} z_1 &= x_o - V^{(p)} g_o && \text{(by definition)} \\ &= x_o - A^{-1} g_o \\ &= x_o - A^{-1}(Ax_o + b) \\ &= -A^{-1}b && \text{(by 1.7))} \end{aligned}$$

and by theorem 1.2 $-A^{-1}b$ is the location of the minimum of the quadratic functional J defined in chapter 1. This is the idea due to Powell as mentioned in the opening sentence of this chapter.

5. CONSTRAINTS

In this chapter we shall consider the problem of computing the location of the minimum of a differentiable functional defined on a real Hilbert space, H, subject to linear equality constraints. It is shown how this problem can be attacked by a modification of the rank one, quasi-Newton algorithm outlined in section 1 of chapter 2.

5.1 Minimization on a Closed Linear Subspace

We shall assume that $J:H \to R$ is a differentiable functional and that D is a closed linear subspace of H. We wish to find $\tilde{x} \in D$ such that $J(\tilde{x}) \leq J(x)$ for all $x \in D$. Let D^* denote the orthogonal complement of D so that $H = D \oplus D^*$. Then for any $x \in H$ there exist unique $x_D \in D$ and $x_{D^*} \in D^*$ such that $x = x_D + x_{D^*}$. Therefore, we can define an operator

$$P:H \to D \tag{1}$$

such that $P(x) = x_D$ for each $x \in H$. P is called the projection operator of H onto D. It is known [1] that P is linear, self-adjoint, bounded and

$$P^2 = P. \tag{2}$$

Moreover, by (2), for all $z \in H$,

$$(z,Pz) = (z,P^2z) = (Px,Pz) = \|Pz\|^2 \tag{3}$$

Lemma 5.1: If we apply the basic algorithm outlined in chapter 2 with $V^{(o)} = P$, the projection operator defined in (1), then $V^{(k)} = V^{(o)}V^{(k)}V^{(o)}$ and $V^{(o)}r_k = r_k$ for all k, where $r_k = V^{(k)}(y_k + \alpha_k g_k)$.

Proof: (By mathematical induction) since $V^{(o)} = P$ we have

$$V^{(o)} \cdot V^{(o)} = V^{(o)} \tag{4}$$

Hence,

$$\begin{aligned} V^{(o)}r_1 &= V^{(o)}(V^{(o)}(y_1 + \alpha_o g_o)) && \text{(by (2.3))} \\ &= V^o(y_1 - \alpha_o g_o) && \text{(by (4))} \\ &= r_1 && \text{(by (2.3))} \end{aligned}$$

Also, $V^{(o)}(V^{(o)})V^{(o)} = V^{(o)}$ by (4). Hence, the theorem is true for $k = 0$. Assume that

$$V^{(o)}V^{(k)}V^{(o)} = V^{(k)} \tag{5}$$

By applying (2.3), (5), and (4), we have

$$\begin{aligned} V^{(o)}r_k &= V^{(o)}(V^{(k)}(y_k + a_k g_k)) \\ &= V^{(o)}V^{(o)}V^{(k)} \cdot (V^{(o)}(y_k + a_k g_k)) \\ &= V^{(o)}V^{(k)}V^{(o)}(y_k + a_k g_k) \\ &= V^{(k)}(y_k + a_k g_k) \\ &= r_k \end{aligned} \tag{6}$$

If $V^{(k)} = V^{(k+1)}$ the theorem is true. Otherwise

$$V^{(k+1)} = V^{(k)} + \frac{(\lambda_k - 1)}{\rho_k} r_k >< r_k \quad (\text{by } (2.9)) \qquad (7)$$

where the operator $B^{(k)}$ given in (2.10) is written in dyadic notation [12]. Hence,

$$V^{(o)}V^{(k+1)}V^{(o)} = V^{(o)}V^{(k)}V^{(o)} + \frac{(\lambda_k - 1)}{\rho_k} V^{(o)}r_k >< V^{(o)}r_k. \qquad (8)$$

By applying (5) and (6) to the right hand side of (8) we have

$$V^{(o)}V^{(k+1)}V^{(o)} = V^k + \frac{(\lambda_k - 1)}{\rho_k} r_k >< r_k = V^{(k+1)}. \quad (\text{by } (2.9))$$

Lemma 5.2: If $V^{(o)} = P$, the projection operator defined in (1), then for any $z \in H$, we have $V^{(o)}V^{(k)}z = V^{(k)}V^{(o)}z = V^{(k)}z$.

Proof: By lemma 5.1 and (4), we have for any $z \in H$

$$V^{(o)}V^{(k)}z = V^{(o)}(V^{(o)}V^{(k)}V^{(o)})z = V^{(o)}V^{(k)}V^{(o)}z = V^{(k)}z.$$

Notice that the proof of the two lemmas above required only that $V^{(o)} \cdot V^{(o)} = V^{(o)}$.

Theorem 5.1: If the initial estimate x_o of the location of the constrained minimum of J is an element of D and $V^{(o)} = P$, the projection operator on D defined in (1), then the iterates $x_1, x_2, \ldots, x_n \ldots$ generated by the basic algorithm outlined in section 1 of chapter 2 are all elements of D.

Proof: (By mathematical induction) since the x_1 generated by the basic algorithm is either x_o or $x^* = x_o - \alpha_o V^{(o)} g_o$ by (2.1) and $x_o \in D$ by hypothesis, we only need to show that $x^* \in D$ in order to establish the theorem for $k = 1$. But, since $V^{(o)}$ is the projection operator onto D, we have $V^{(o)}g_o \in D$ and since D is a subspace, we obtain $x_o - \alpha_o V^{(o)} g_o \in D$ for any $\alpha_o \in R$.

Assume $x_k \in D$ and consider $x^* = x_k - \alpha_k V^{(k)} g_k = x_k - \alpha_k V^{(o)} V^{(k)} g_k = x_k - \alpha_k V^{(o)}(V^{(k)} g_k)$ (lemma 5.2). Because $V^{(o)}$ is the projection operator, $V^{o}(V^{(k)}g_k) \in D$. Hence, $x^* = x_k - \alpha_k V^{(k)} g_k \in D$ for all $\alpha_k \in R$.

Notice that theorem 5.1 and the above lemmas are independent of the manner of choosing α_k and the functional J is only required to be differentiable. Further, notice that the theorem and lemmas hold if, in (2.9) $V^{(n+1)} = V^{(n)} + \mu B^{(n)}$ for any real number μ.

Now suppose that the functional to be minimized is quadratic as discussed in chapter 1. The problem is, therefore, to find the location of the minimum value of $J(x) = J_o + (b,x) + 1/2(x,Ax)$ for all $x \in D$, a closed linear subspace of H. Now if P denotes the projection mapping of H onto D and we denote $I - P$ by C, C is bounded, and the problem becomes to minimize J subject to $Cx = \theta$. Notice the null space of C is exactly D. If we make the substitution $x = y - A^{-1}b$, then

$$\begin{aligned} J(x) &= J_o + (-A^{-1}b + y, b) + \tfrac{1}{2}(-A^{-1}b + y, A(-A^{-1}b + y)) \\ &= J_o - (A^{-1}b, b) + (y,b) + \frac{(A^{-1}b,b)}{2} + \tfrac{1}{2}(y, A(-A^{-1}b)) \\ &\quad + \tfrac{1}{2}(-A^{-1}b, Ay) + \tfrac{1}{2}(y,Ay) \\ &= J_o - \tfrac{1}{2}(A^{-1}b,b) + \tfrac{1}{2}(y,Ay) \end{aligned}$$

$$J(x) = J_0 - \tfrac{1}{2}(A^{-1}b,b) + \tfrac{1}{2}\tilde{J}(y)$$

where $\tilde{J}(y) = (y,Ay)$. If $Cx = \theta$ then $C(-A^{-1}b + y) = \theta$ or $Cy = CA^{-1}b$. If we let $CA^{-1}b = d$ then minimizing $J(x)$ subject to $Cx = \theta$ is equivalent to minimizing $\tilde{J}(y)$ subject to $Cy = d$. We shall examine the problem of minimizing $\tilde{J}(y)$ subject $Cy = d$ and then see what this tells us about the original problem, that is, to minimize $J(x)$ subject to $Cx = \theta$.

We shall define a functional $(\ , \ )_1 : H \times H \to R$ as $(x,y)_1 = (x,Ay)$ for all $x,y \in H$. Notice that for any $x \in H$, $(x,x)_1 = (x,Ax) \geq m\|x\|^2$ (1.3) so that if $x \neq \theta, (x,x)_1 > 0$ and $(x,x)_1 = 0$ if and only if $x = \theta$. Moreover, the inner product $(\ , \ )$ is linear in the first term by definition, hence, we know that the function $(\ , \ )_1$ is linear in the first term. Moreover, since $A = A^*$ (1.3) for every $x,y \in H$, we have

$$(x,y)_1 = (x,Ay) = (Ay,x) = (y,A^*x)$$

$$= (y,Ax) = (y,x)_1.$$

That is, $(\ , \ )_1$ is symmetric. Hence, $(\ , \ )_1$ is an inner product on the linear space H. We shall denote the space $(H,(\ , \ )_1)$ by H'.

We can see that H' is complete as follows: suppose that $(x_p - x_n, A(x_p - x_n)) \to 0$ as $p,n \to \infty$. Then, since for any p,n $(x_p - x_n, A(x_p - x_n)) \geq m(x_p - x_n, x_p - x_n) \geq 0$ by (1.3), $(x_p - x_n, x_p - x_n) \to 0$ and by the completeness of H there exists an $x \in H$ such that $(x_n - x, x_n - x) \to 0$ as $n \to \infty$. Since by (1.3), $M(x_n - x, x_n - x) \geq (x_n - x, A(x_n - x)) \to 0$, we have $(x_n - x, A(x_n - x)) \to 0$, Hence, H' is complete. Therefore, H' is a Hilbert space.

Now if we denote by M the closed linear subspace of H' which is the null space of C and if $\tilde{y} \in H'$ is such that $C\tilde{y} = d$ then the linear variety which satisfies $C\tilde{y} = d$ is given by $V = \tilde{y} + M$. By the projection theorem [1] there is a unique vector y_o in V of minimum norm with respect to the H' norm. Further, y_o is characterized by the fact that y_o is the only element of V orthogonal to M with respect to the $(\, , \,)$ inner product.

This means that

$$(y_o, y_o)_1 \leq (y, y)_1 \tag{9}$$

for all $y \in V = \tilde{y} + m$, that is for all y such that $Cy = d$. Moreover, for every y such that $Cy = \theta$, y_o is characterized by the fact that

$$(y, y_o)_1 = 0. \tag{10}$$

That is, in terms of the definition of $(\, , \,)_1$ we have from (9) and (10)

$$(y_o, Ay_o) \leq (y, Ay) \tag{11}$$

for all y such that $Cy = d$, and

$$(y, Ay_o) = 0 \tag{12}$$

for all y such that $Cy = \theta$. Hence, the solution to the problem of finding the minimum of $\tilde{J}(y)$ is characterized by (11) and (12). In terms of the original problem of minimizing $J(x)$ subject to $Cx = \theta$, this means that the problem has a unique solution $\tilde{x} = -A^{-1}b + y_o$ and if x satisfies $Cx = \theta$ then by (12)

$$(x, A(\tilde{x} + A^{-1}b)) = 0. \tag{13}$$

But by (1.7) $A\tilde{x} + b = g(\tilde{x})$. Hence, we have that at $\tilde{x}$, $(x, g(\tilde{x})) = 0$ for all x such that $Cx = \theta$. That is, $g(x_o)$ is orthogonal to the null space of C which is D. In other words the projection of $g(\tilde{x})$ onto D is zero.

Now let us use the modified rank one algorithm to locate $\tilde{x}$. Suppose that the scalar α_n is chosen so that

$$J(x_n + \alpha_n s_n) \leq J(x_n + \lambda s_n)$$

for all $\lambda \in R$, that is, α_n is chosen by (2.17). Therefore, the value of α_n is given by (2.19). We apply the modified basic algorithm as discussed in this chapter with the initial estimate $x_o \in D$ and $V^{(o)} = P$ as defined by (1).

We shall now establish conditions which will guarantee that the projection onto D of the gradient at the iterates tends to zero. As shown above, this is a necessary and sufficient condition for a minimum.

By (2.15), we have $(y_n, r_n) = (y_n, (V^{(n)} - A^{-1})y_n)^{-1}$. Then from (1.12) and (2.5) we have

$$(y_n, A^{-1}y_n) = (y_n, \sigma_n) = (g_{n+1} - g_n, \sigma_n)$$

$$= (g_{n+1}, \sigma_n) - (g_n, \sigma_n).$$

By the choice of α_n we know that $(g_{n+1}, \sigma_n) = 0$. Hence, by the definition of σ_n we have

$$(y_n, A^{-1}y_n) = \alpha_n(g_n, V^{(n)}g_n). \tag{13}$$

Also by (2.5)

$$(y_n, V^{(n)}y_n) = (g_{n+1}, V^{(n)}g_{n+1}) - (g_n, V^{(n)}g_{n+1}) - (g_{n+1}, V^{(n)}g_n) + (g_n, V^{(n)}g_n \tag{14}$$

Therefore, by theorem 2.2 and (2.18) $(g_n, V^{(n)}g_{n+1}) = 0$ and $(g_{n+1}, V^{(n)}g_n) = 0$. Hence, (14) becomes

$$(y_n, V^{(n)}y_n) = (g_{n+1}, V^{(n+1)}g_{n+1}) + (g_n, V^{(n)}g_n). \tag{15}$$

Hence, $(y_n, (V^{(n)} - A^{-1})y_n) = (g_{n+1}, V^{(n)}g_{n+1}) + (1 - \alpha_n)\ (g_n, V^{(n)}g_n)$.

Therefore we can say:

Lemma 5.3: If $V^{(n)}$ is a positive operator on D and $\alpha_n \leq 1$ then $(y_n, (V^{(n)} - A^{-1})y_n) = (y_n, r_n) \geq 0$.

Lemma 5.4: If $V^{(o)}$ is the projection operator onto D and the $V^{(i)}$ are positive uniformly bounded linear operators on D with bound $K > 0$, then

$$(g_i, V^{(i)}g_i) \geq \frac{\left\|V^{(i)}g_i\right\|}{K}. \tag{16}$$

Proof: Define $(\ , \ )_i : H \times H \to R$ such that $(x,y)_i = (x,V^{(i)}y)$ for all $x,y \in H$. By lemma 5.1 $V^{(i)} = V^{(o)}V^{(i)}V^{(o)}$, so if $x \in H$ then $V^{(o)} x \in D$, hence, $(x,x)_i = (x,V^{(i)}x) = ((V^{(o)}x),V^{(i)}(V^{(o)}x)) \geq 0$ since $V^{(i)}$ is positive on D. Therefore, the Schwarz inequality holds for each i, that is, $(x,y)_i \leq (x,x)_i(y,y)_i$ [2]. Hence,

$$\begin{aligned} \left\| V^{(i)}g_i \right\|^4 &= (V^{(i)}g_i,V^{(i)}g_i)^2 = (V^{(i)}g_i,g_i)_i \\ &\leq (V^{(i)}g_i,V^{(i)}g_i)_i \cdot (g_i,g_i)_i \\ &= (V^{(i)}g_i,V^{(i)}(V^{(i)}g_i)) \cdot (g_i,V^{(i)}g_i) \\ &\leq K \left\| V^{(i)}g_i \right\|^2 (g_i,V^{(i)}g_i). \end{aligned}$$

Therefore, if $\left\|V^{(i)}g_i\right\| \neq 0$ we have $K(g_i,V^{(i)}g_i) \geq \left\| V^{(i)}g_i \right\|^2$. Hence, $(g_i,V^{(i)}g_i) \geq \dfrac{\left\|V^{(i)}g_i\right\|^2}{K}$.

By our choice of α_n we know that (2.20) holds. Hence,

$$\lim_{n \to \infty} \frac{(s_n,g_n)^2}{(s_n,As_n)} = 0$$

Since

$$\begin{aligned} \frac{(s_i,g_i)^2}{(s\ ,As_i)} &\geq \frac{(g_i,V^{(i)}g_i)^2}{M\left\|V^{(i)}g_i\right\|^2} && \text{(by (1.3) and (2.2))} \\ &\geq \frac{\left\|V^{(i)}g_i\right\|^2}{KKM} && \text{(by (16))} \end{aligned} \qquad (17)$$

we have by (11) $\left\| V^{(i)} g_i \right\| \to 0$ as $i \to \infty$. Moreover, since by (12) $\dfrac{\left(g_i, V^{(i)} g_i\right)^2}{\left\| V^{(i)} g_i \right\|} \to 0$, and $\left\| V^{(i)} g_i \right\| \to 0$ as $i \to \infty$, we obtain

$$(g_i, V^{(i)} g_i) \to 0 \qquad \text{as} \quad i \to \infty \tag{18}$$

If $(r_l, y_l) \neq 0$ we have in view of (2.14) and (2.10) for any $x \in H$,

$$V^{(i)} x = V^{(o)} x - \sum_{l=0}^{i-1} \frac{(r_l, x)\, r_l}{(r_l, y_l)} \tag{19}$$

Hence,

$$(x, V^{(i)} x) = (x, V^{(o)} x) - \sum_{l=0}^{l-1} \frac{(r_l, x)^2}{(r_l, y_l)} \tag{20}$$

Recall from step 2 of the basic algorithm that if $(r_j, y_j) = 0$, for some j, then $V^{(j+1)} = V^{(j)}$ so that the term containing (r_j, y_j) in the sum given in (19) or (20) is not present. We shall assume that if $(r_j, y_j) = 0$ for some j we have not included that term in the sum in (19) or (20). Recall that from lemma 5.3, if $\alpha_l \leq 1$ for $l = 0,1,2 \ldots, i - 1$ then $(\gamma_l, y_l) \geq 0$. Hence we have

$$\begin{aligned} (x, V^{(i)} x) &\geq (x, V^{(o)} x) \\ &= (x, V^{(o)} V^{(o)} x) \quad \text{since} \quad V^{(o)} = V^{(o)} V^{(o)} \\ &= (V^{(o)} x, V^{(o)} x) \quad \text{since} \quad (V^{(o)})^* = V^{(o)} \\ &= \left\| V^{(o)} x \right\|^2 \end{aligned} \tag{21}$$

From (21) with $x = g_i$ we have the following.

Theorem 5.2: If $\alpha_i \leq 1$ for all i and the $V^{(i)}$ are uniformly bounded positive operators on D, then $V^{(o)}g_i \to \theta$ as $i \to \infty$.

Proof: If $\alpha_i \leq 1$ for all i, then by (21) with $x = g_i$ we have $(g_i, V^{(i)}g_i) \geq \left\| V^{(o)}g_i \right\|^2 \geq 0$ and $(g_i, V^{(i)}g_i) \to 0$ as $i \to \infty$ by (18). Hence, $\left\| V^{(o)}g_i \right\|^2 \to 0$.

We have now established conditions which guarantee that the projection on D of the gradient of the quadratic functional evaluated at the iterates tends to zero. Notice that if M as defined in (1.4) is such that $M \leq 1$, then since $\left\| P \right\| \leq 1$ [2],

$$\left\| Px \right\| \leq \left\| x \right\| \leq \frac{1}{M} \left\| x \right\| \leq \left\| A^{-1}x \right\| \leq \frac{1}{m} \left\| x \right\|,$$

that is, $P \leq A^{-1}$. Since $V^{(o)} = P$, $V^{(o)} \leq A^{-1}$ we have by theorem 2.6 that $V^{(n)} \leq A^{-1}$ for every n. Hence the $V^{(n)}$ are uniformly bounded.

5.2 Linear Equality Constraints of the Type $Cx = \omega$

Suppose the problem is to compute the location of the minimum of a differentiable function $J:H \to R$, with gradient $g:H \to H$, subject to the constraint that $Cx = \omega$, where C is a bounded linear operator from H into $\bar{H}$, where $\bar{H}$ is another Hilbert space, and ω is a fixed element of $\bar{H}$. That is, we wish to find $\tilde{x} \in H$ such that $C\tilde{x} = \omega$ and $J(x) \geq J(\tilde{x})$ for all $x \in H$ such that $Cx = \omega$. With a slight modification, we can apply the basic algorithm outlined in chapter 2 to this problem. Moreover, we can show that the sequence of iterates $x_1, x_2, \ldots, x_n, \ldots$ generated by this modification is such that for each k, $Cx_k = \omega$.

Let $V^{(o)}$ be the projection operator of H onto the null space of C (a closed subspace of H, since C is bounded). Let x_o be such that $Cx_o = \omega$ (22) and apply the algorithm. Now, $x_1 = x_o$ or x^* where $x^* = x_o - \alpha_o V^{(o)} g_o$. Consider

$$\begin{aligned} Cx^* &= Cx_o - C(\alpha_o V^{(o)} g_o) \\ &= \omega - \alpha_o C(V^{(o)} g_o) \end{aligned} \tag{22}$$

where $V^o g_o$ is in the null space of C by the choice of $V^{(o)}$. Hence, $Cx^* = \omega$ for all $\alpha_o \in R$. Therefore, $Cx_1 = \omega$ in either case.

Since either $x_{n+1} = x_n$ or $x_{n+1} = x_n - \alpha_n V^{(n)} g_n$ we know that if $Cx_n = \omega$ and $x_{n+1} = x_n$ then $Cx_{n+1} = \omega$. Otherwise, we consider $Cx_{n+1} = Cx_n - C\alpha_n V^{(n)} g_n$. Since the proof of lemma 5.2 depended only upon the fact that $V^{(o)} = V^{(o)} \cdot V^{(o)}$ and we know that this is true for the projection operator onto the null space of C, we have that $V^{(n)} g_n = V^{(o)} V^{(n)} g_n$. Hence, $V^{(n)} g_n$ is in the null space of C. Therefore, $C(\alpha_n V^{(n)} g_n) = \theta$. Hence, $Cx_{n+1} = Cx_n = \omega$. Therefore, by mathematical induction we have established the following theorem:

Theorem 5.3: If $V^{(o)}$ is the projection operator and the null space of C and $V^{(n)}$ is defined as in (2.10) and $Cx_o = \omega$, then $Cx_n = \omega$ for all n where the x_n's are the iterates generated by the algorithm outlined in chapter 2.

We shall now show that the problem considered in section 1 of this chapter is of the type examined in this section. The problem is that of finding $\tilde{x} \in D$, D a subspace of H, such that $J(\tilde{x}) \leq J(x)$ for all $x \in D$, where J is a differentiable function. Suppose we let P denote the projection operator of H onto D and we define the bounded linear operator from H into H by, $C = I - P$ where I the identity operator on H, then the problem can be seen as that of minimizing J subject to $Cx = \theta$. Therefore, the problem of section 1 is a special class of those problems considered in this section. Hence theorem 5.1 follows from 5.3.

6. APPLICATION TO OPTIMAL CONTROL THEORY

In this chapter the results of the first five chapters are used to develop a method of computing the solution of various types of optimal control problems. We shall consider fixed-time problems since by a simple transformation [3] the free-time problem can be transformed into a fixed-time problem. Moreover, Horwitz and Sarachik [21] have given several other schemes for solving the free-time problem using fixed-time techniques, and these schemes are applicable when the basic algorithm, outlined in chapter 2, is used. Also Leondes and Niemann [24] have proposed a computational scheme for handling the free-time problem by using fixed-time techniques.

6.1 A Quadratic Payoff With Linear Constraining Differential Equations

From the class $L_r^2[t_o, t_1]$ we wish to find that function $u^*(t)$ which minimizes

$$J[u] = \frac{1}{2} \int_{t_o}^{t_1} \{x^T(t)P(t)x(t) + u^T(t)R(t)u(t)\} dt \qquad (1)$$

subject to the constraints

$$\dot{x}(t) = G(t)\,x(t) + B(t)u(t) \qquad (2)$$

and $x(t_o) = x_o$. where x_o, t_o, and t_1 are fixed.

Hereby: x is an n-vector,

u is an r-vector,

$G(t)$ is an $n \times n$ matrix with components in $L^1[t_o,t_1]$,

$B(t)$ is an $n \times r$ matrix with components in $L^1[t_o,t_1]$, and bounded,

$P(t)$ is an $n \times n$ symmetric, positive semi-definite matrix the components of which are piece-wise continuous on $[t_o,t_1]$, and

$R(t)$ is an $r \times r$ symmetric uniformly positive definite matrix the components of which are piece-wise continuous on $[t_o,t_1]$.

Horwitz and Sarachik [20] have shown that this problem can be considered as that of finding the location of the minimum of a quadratic functional on $L_r^2[t_o,t_1]$. This can be seen by defining the following linear operators:

$$\begin{aligned} &P:L_n^2[t_o,t_1] \to L_n^2[t_o,t_1] \\ &R:L_r^2[t_o,t_1] \to L_r^2[t_o,t_1] \\ &E:L_n^2[t_o,t_1] \to L_n^2[t_o,t_1] \\ &F:L_r^2[t_o,t_1] \to L_n^2[t_o,t_1] \end{aligned} \tag{3}$$

where for $y \in L_n^2[t_o,t_f]$, $z \in L_r^2[t_o,t_1]$,

$$(Py)(t) = P(t)y(t)$$

$$(Rz)(t) = R(t)z(t)$$

$$(Ey)(t) = \Phi(t,t_o)y(t) \tag{4}$$

and

$$(Fz)(t) = \int_{t_o}^{t} \Phi(t,\tau)B(\tau)z(\tau)d\tau$$

where $\dot{\Phi} = G\Phi$ with $\Phi(t_o,t_o) = I$.

It is well known [5] that for any $u \in L_r^2[t_o,t_1]$, $x = Ex_o + Fu$, so that (1) becomes

$$\begin{aligned} J[u] = \tfrac{1}{2} &\langle Ex_o + Fu, P(Ex_o + Fu)\rangle \\ &+ \tfrac{1}{2}\langle u,Ru\rangle \end{aligned} \tag{5}$$

where $\langle\ ,\ \rangle$ is the usual inner product defined on $L_r^2[t_o,t_1]$.

Hence,

$$\begin{aligned} J[u] &= \tfrac{1}{2}\langle Ex_o,PEx_o\rangle + \tfrac{1}{2}\langle u,F^*PEx_o\rangle \\ &\tfrac{1}{2}\langle (PF)^*Ex_o,u\rangle + \tfrac{1}{2}\langle u,(F^*PF + R)u\rangle \end{aligned} \tag{6}$$

If we let

$$J = \frac{1}{2}\left\langle Ex_o, PEx_o \right\rangle$$

$$w = F^*PEx_o, \tag{7}$$

$$A = F^*PF + R,$$

(6) becomes

$$J\left[u\right] = J_o + \left\langle w,u \right\rangle + \frac{1}{2}\left\langle u,Au \right\rangle. \tag{8}$$

Moreover, since P is positive semi-definite and R is uniformly positive definite, A is a strongly positive linear operator. Hence, $J\left[u\right]$ as given by (8) is a quadratic functional on the real Hilbert space $L_r^2\left[t_o,t_1\right]$ of the type discussed in chapter 1. By (1.7) the gradient of J is given by

$$g(u) = Au + w \tag{9}$$

Moreover, this is exactly the type of function for which the conditions given in theorems 2.10, 2.11, and the corollary to theorem 4.1 guarantee the convergence of the various modifications of the basic algorithm.

Note that if we wish to find the location of the minimum of (1) subject to (2) but with $x(t_o) = \tilde{x}_o$ as initial condition, we can repeat the definitions given in (3) and (7). Then the vector w and the scalar J_o defined by (7) are changed to $\tilde{w}$ and $\tilde{J}_o$, say. However, the operator A also defined by (7) is unchanged.

From theorem 1.2 , the location of the minimum of the resulting quadratic functional $\tilde{J}\left[u\right] = \tilde{J}_o + \left\langle \tilde{w},u \right\rangle + \frac{1}{2}\left\langle u,Au \right\rangle$ is given by $-A^{-1}\tilde{w}$. Since the operator $V^{(n)}$ which we computed when solving for the minimum of (8), converges pointwise to A^{-1}, by theorem 2.8, we can use this $V^{(n)}$ as our new initial estimate of A^{-1}. In this fashion we can accelerate the convergence of the iterates for the second problem, that is, of computing the location of the minimum of $\tilde{J}$.

6.2 General Optimal Control Problems and the Gradient of the Payoff

In this section we shall describe a class of problems generally referred to as optimal control problems [29] or in the Calculus of Variations as Lagrange Problems [33] . Also we shall show how to apply the algorithms discussed in chapters 2 and 4 to compute solutions to these problems.

Suppose we have a system of n differential equations

$$\dot{x}(t) = f(x,u,t) \tag{10}$$

with $x(t_0) = x_0$ and $u \in R^r$. We wish to choose a function $u = \tilde{u}(t)$ which minimizes the value of $\int_{t_0}^{t_1} L(x(t),u(t),t)dt$.

We shall assume that $f(x,u,t)$ and $L(x,u,t)$ have continuous partial derivatives of at least second order in x and u and piecewise continuous in t. Also, we shall assume that there are no constraints on u or x, other than x must satisfy (10).

Moreover, we shall assume that L and f are such that corresponding to every $u = u(t) \in L_r^2 [t_0,t_1]$, a real Hilbert space, there exists a solution, $x = x(t)$, of (10) and that for this x and u

the integral, $\int_{t_o}^{t_1} L(x(t),u(t),t)dt$, exists. By a solution to (10) we mean, as is the usual case in ordinary differential equations, an absolutely continuous function $\varphi = \varphi(t)$ such that $\varphi(t_o) = x^o$ and $\dot{\varphi}(t) = f(\varphi(t),u(t),t)$ almost everywhere for some $u = u(t)$ [5]. By the continuity conditions on L and f, if we can restrict our attention to a compact subset of (t,x) space for all u, then standard results of differential equations theory concerning existence and uniqueness of solutions hold [5, 16, 17, 29]. An assumption on f and L which guarantees this is to assume that there exists C, a scalar, such that for all $t \in [t_o,t_1]$, x, and u

$$\left|(\tilde{x},\tilde{f}(\tilde{x},u,t))\right| \leq C\left[1 + \left|\tilde{x}\right|^2\right] \tag{11}$$

where $\tilde{f}$ and $\tilde{x}$ denote the vectors (L,f) and (x_o,x) respectively, with $\dot{x}_o = L(x,u,t)$ and $x_o(t_o) = 0$. This implies $(\dot{\tilde{x}},\tilde{x}) \leq C\left[1 + \left|\tilde{x}\right|^2\right]$ so that $\left|x(t)\right|^2 \leq \left[1 + \left|x^o\right|^2\right]e^{2Ct_1}$. The above inequality is shown by Hermes and LaSalle in [16]. Hence, we can define the functional $J:H \to R$ by

$$J[u] = \int_{t_o}^{t_1} L(x(t),u(t),t)dt$$

where $x(t)$ is a solution of (10) corresponding to u.

Therefore, our problem appears to be that of locating the minimum of a functional J on a real Hilbert space H. In order to apply the algorithms discussed in chapters 2 and 4, we must compute the gradient of J. The gradient of J is that part of $J[u + \delta u] - J[u]$ which is linear in δu. From (10) we have

$$\tilde{x}(t) = x^{o} + \int_{t_o}^{t} f(\tilde{x}(\tau), u(\tau) + \delta u, \tau)dt$$

$$x(t) = x^{o} + \int_{t_o}^{t} f(x(\tau), u(\tau), \tau)d\tau \qquad (12)$$

Therefore,

$$\tilde{x}(t) - x(t) = \int_{t_o}^{t} \{f(\tilde{x}(\tau), u(\tau) + \delta u, \tau) - f(x(\tau), u(\tau), \tau)\}dt.$$

If we let δx denote the linear part of $\tilde{x}(t) - x(t)$, then

$$\dot{\delta x} = f_x \delta x + f_u \delta u \qquad (13)$$

with $\delta x(t_o) = 0$ where f_x denotes the $n \times n$ matrix $\frac{\partial f}{\partial x}$ and $f_u = \frac{\partial f}{\partial u}$ an $n \times r$ matrix, evaluated at $(x(t), u(t), t)$.

Moreover,

$$J[u + \delta u] - J[u] = \int_{t_o}^{t_1} \{L(x, u + \delta u, t) - L(x, u, t)\}dt \qquad (14)$$

and if we let δJ denote that portion of (14) which is linear in δx and δu, then

$$\delta J = \int_{t_o}^{t_1} L_x \delta x + L_u \delta u \; dt \qquad (15)$$

where L_x denotes $\frac{\partial L}{\partial x}$ and $L_u = \frac{\partial L}{\partial u}$ evaluated at $(x(t), u(t), t)$.

We then let $\lambda(t)$ be an n-vector valued function satisfying

$$\dot{\lambda}(t) = -f_x^T\lambda - L_x^T \tag{16}$$

with $\lambda(t_1) = 0$. Then we have from (16)

$$\begin{aligned} \frac{d(\lambda^T\delta x)}{dt} &= \dot{\lambda}^T\delta x + \lambda^T(\dot{\delta x}) \\ &= -\lambda^T\{f_x\delta x\} - L_x\delta x + \lambda^T f_x\delta x + \lambda^T f_u\delta u. \end{aligned} \tag{17}$$

So that integrating (17) from t_o to t_1 we have

$$\lambda(t_1)\delta x(t_1) - \lambda(t_o)\delta x(t_o) = -\int_{t_o}^{t_1} L_x\delta x\, dt + \int_{t_o}^{t_1} \lambda^T f_u\delta u\, dt$$

and since $\lambda(t_1) = 0 = \delta x(t_o)$, we have

$$\int_{t_o}^{t_1} L_x\delta x\, dt = \int_{t_o}^{t_1} \lambda^T f_u\delta u\, dt. \tag{18}$$

So, substituting (18) in (15), we get

$$\delta J = \int_{t_o}^{t_1}\{\lambda^T f_u + L_u\}\delta u\, dt.$$

Hence, the gradient of J is given by

$$g(u) = \frac{\partial L^T}{\partial u}(x(t),u(t),t) + \frac{\partial f^T}{\partial u}(x(t),u(t),t)\lambda(t) \tag{19}$$

where λ given by (16) can be thought of as an integrating factor for (13). [25, 33].

It is seen from (19) that if we define the Hamiltonian to be

$$H(x,\lambda,t,u) = L(x,u,t) + \lambda^T f(x,u,t) \tag{20}$$

Then the gradient of J at u is given by

$$\nabla_u H = (L_u(x(t),u(t),t) + \lambda^T(t) f_u(x(t),u(t),t))^T$$

where (21)

$$\dot{x}(t) = f(x(t),u(t),t) = \frac{\partial H}{\partial \lambda}, x(to) = x^o$$

$$\dot{\lambda}(t) = -\frac{\partial H}{\partial x} \qquad , \lambda(t_1) = 0$$

The computational steps necessary to compute the gradient of J at $u = u_o(t)$ are: integrate $\dot{x} = f(x,u_o,t)$ with $x(t_o) = x_o$ forward to $t = t_1$, then at $t = t_1$ we integrate

$$\dot{\lambda} = -f_x^T(x,u_o,t)\lambda - L_x^T(x,u_o,t)$$

with $\lambda(t_1) = 0$ backward to $t = t_o$. Therefore, we can then compute the gradient as given in (19) using the control $u = u_o(t)$ and the values of $x(t)$ and $\lambda(t)$ computed above. If the gradient is computed according to (19), then $B^{(n)}$ and r_n can be computed as in (2.10) and (2.5) by following the algorithms outlined in chapters 2 and 4. Hence, these algorithms can be used directly to compute the optimal control.

6.3 Computing the Optimal Control for a Problem With a Compact, Convex Control Region Via the Algorithm of Chapter 2

The problem considered in section 2 of this chapter, which we shall call the first problem, is to find a function u such that

$$\int_{t_0}^{t_1} L(x(t),u(t),t)dt \rightarrow \min. \tag{22}$$

subject to $\dot{x} = f(x,u,t)$ and $x(t_0) = x_0$. This problem is not entirely typical of optimal control problems in that the range of u is unrestricte For a large class of those problems generally considered to be optimal control problems, the function u is a member of $L_r^1[t_o,t_1]$ and has its range in some subset U of R^r. U is called the control region of the problem [29]. For (22), we assume that f and L are as section 6.2 except (11) holds for f at every x,t and $u \in U$.

Problems for which U is a convex, compact subset of R^r and which can be transformed into control problems with no spacial restriction on U, were examined by Park [28]. He showed that an optimal control problem as (22) for which U is a convex and compact subset of a Euclidean space can be transformed into an "equivalent" problem with its associated control region - a Euclidean space of dimension p. Hence, the new control variables have no restriction on their range. We shall see that this "equivalent" problem can be seen as that of locating the minimum value of a functional defined on a Hilbert space. The algorithms which we have previously discussed can be used to compute the location of the minimum of this functional and the results then can be transformed back to the original problem.

A problem of the type investigated by Park is to find an $L_r^1[t_o,t_1]$ function $u = u(t)$ with range in $U \subseteq E^r$, U a compact convex set, such that (22) holds. Let

$$\psi : R^p \to R^r \tag{23}$$

be a map of the type discussed by Park, that is, ψ is continuous, onto U and there exists a compact subset Z of R^p such that

$$\psi(Z) = U \tag{24}$$

By Filippov's Lemma [16] for every admissible control u, that is, $u \in L_r^1[t_o,t_1]$ with range in U, there exists a bounded measurable function $z:[t_o,t_1] \to Z$ such that for every t,

$$u(t) = \psi \cdot z(t) \tag{25}$$

Let us suppose that the problem to be solved is as in (22) where $u \in L_r^1[t_o,t_1]$ has its range in U, a compact, convex subset of R^r. Let ψ and Z be as in (23) and (24). The "equivalent" problem which we will call problem 2 then is to find $y:[t_o,t_1] \to R^p$ such that

$$\int_{t_o}^{t_1} L(x(t),\psi(y(t)),t\,)dt \to \min \tag{26}$$

subject to $\dot{x} = f(x(t),\psi(y(t),t)$ and $x(t_o) = x_o$ where $y \in L_p^2[t_o,t_1]$.

In problem 2, we are minimizing over the Hilbert space $L_p^2[t_o,t_1]$, not a subset of $L_r^1[t_o,t_1]$ as in problem 1. This follows because for every $y \in L_p^2[t_o,t_1]$ y is measurable, and since ψ is a given continuous function, $\psi \cdot y$ is measurable. Moreover ψ has a bounded range U, hence, $\psi \cdot y$ is bounded and measurable on $[t_o, t_1]$. Therefore, for any $y \in L_p^2[t_o, t_1]$ $\psi \cdot y$ is an admissible control, that is, $L_r^1[t_o, t_1]$ with range in U. Conversely, for any admissible control u, the corresponding y given by Filippov's Lemma is measurable and has its range in Z, a compact set. Hence, y is bounded. Therefore, $y \in L_p^2[t_o, t_1]$. Hence we see that the space of admissible controls for problem 2 is all of $L_p^2[t_o, t_1]$, a Hilbert space, whereas the "equivalent" problem 1 had as its admissible controls a subset of $L_r^1[t_o, t_1]$.

Note that if the transformation ψ given in (23) has continuous derivatives of second order, then problem 2 is of the type discussed in section 2 of this chapter. Hence, the computation of the location of the minimum can be carried out by the algorithms given in chapters 2 and 4, and the gradient of the functional to be minimized in problem 2 is given by

$$\begin{aligned} g(y(t)) = (\lambda^T(t) f_u(x(t), \psi(y(t)), t) \; \psi_y(y(t)) \\ + L_u(x(t), \psi(y(t)), t) \psi_y(y(t)))^T, \end{aligned} \tag{27}$$

where $\dot{\lambda}(t) = -f_x^T(x(t), \psi(y(t)), t)\lambda(t) - L_x^T(x(t), \psi(y(t)), t)$. This gradient is found by applying to problem 2 the same techniques used to get (19).

Hence this transformation technique can be useful in computing the solution to a wide class of optimal control problems. It can also be used to apply the classical calculus of variations results to various types of optimal control problems [15,28] .

6.4 Optimal Control Problems With End-Point Constraints

Suppose we wish to solve the problem posed in section 2 of this chapter as outlined in (19) subject to the additional constraint that some of the components of $x(t_1)$ are to be fixed numbers. That is, suppose the first q-components of $x(t_1)$ are to be such that $x_0(t_1) = \bar{x}_i$ for $i = 1,2,\dots,q$ where $\bar{x}_i$ are given scalars.

One approach to computing the solution to this problem would be a "penalty function" technique [23]. This technique is the following: use any admissible control $u = u(t)$, integrate $\dot{x} = f(x(t), u(t), t)$ from x_o at t_o to t_1. At t_1 the components of x will probably not be the prescribed values $\bar{x}_i$, $i = 1,2,\dots,q$, so we will compute

$$x_i(t_1) - \bar{x}_i = \Delta_i[u], i = 1,2,\dots,q.$$

$\Delta_i[u]$ is the error in the ith component of $x(t_1)$ corresponding to the control $u = u(t)$. Then for an arbitrary but fixed set of positive scalars $k_1, k_2,\dots,k_q$, we compute the penalty associated with u as follows:

$$P[u] = \sum_{i=1}^{q} k_i\left(\Delta_i[u]\right)^2. \tag{28}$$

The functional of u which we seek to minimize by our algorithm is

$$\tilde{J}[u] = \int_{t_o}^{t_1} L(x(t),u(t),t)dt + P[u]$$

where $P[u]$ is given in (28).

It can be shown by analysis similar to that of section 2 of this chapter that $\tilde{g}[u]$, the gradient of $\tilde{J}$ in (29), is given by

$$\tilde{g}[u] = (\lambda^T f_u(x(t),u(t),t) + L_u(x(t),u(t),t))^T \tag{30}$$

where

$$\dot{\lambda}(t) = f_x^T(x(t),u(t),t)\lambda(t) - L_x^T(x(t),u(t),t), \lambda(t_1) = \left.\frac{\partial P}{\partial x}\right|_{t=t_1}$$

and $\dot{x}(t) = f(x(t),u(t),t)$.

While this technique appears to handle the problem of the end constraints very nicely, we are left with the problem of choosing the k_i's. Due to the finite number of significant figures on a digital computer, if the k_i's are too large the algorithm will try to satisfy the end conditions at the expense of minimizing $\int_{t_0}^{t_1} L(x,u,t)dt$, and if the k_i's are too small the algorithm may not be sensitive to violations of the end constraints. In some cases, Lasdon et al. [23] have remarked that the penalty function terms in (29) may "create a steep-sided valley in the control space." This would slow the convergence of the algorithm.

Another possible method of computing the optimal control for a problem with end-point constraints is the projection method. This technique is discussed by Rosen [32], Sinnott [34], and Luenberger [25] for various algorithms. The adaptation of this technique to our algorithm appears to be rather straightforward, but we shall not pursue it here.

In the next section we shall examine the optimal control problem with end-point constraints for the case where the state differential equations are linear in the control.

6.5 Optimal Control Problems With Linear Constraining Equations and End Conditions of the Type $Kx(t_1) = d$

Suppose our optimal control problem is to find that $L_r^2[t_o, t_1]$ function $u = \tilde{u}(t)$ which minimizes

$$\int_{t_o}^{t_1} L(x(t), u(t), t)dt \tag{31}$$

with

$$\dot{x}(t) = G(t)x(t) + F(t)u(t)$$

$$x(t_o) = x^o$$

and t_1 is fixed with $Kx(t_1) = d$.

We assume that L has continuous partial derivatives of at least second order in its x and u arguments and is piecewise continuous in t. G and F are matrix valued functions with $L^1[t_o, t_1]$ components and continuous components respectively. K is a $q \times n$ matrix of scalars and d is a q vector of scalara where $q \leq n$. Moreover, we assume that L is such that for any $u \in L_r^2[t_o, t_1]$ and its corresponding $x = x(t)$ (31) exists.

If we denote the principle matrix solution of the homogeneous system $\dot{x}(t) = G(t)x(t)$ by $\Phi(t, t_o)$ where $\Phi(t_o, t_o) = I$ then the state vector x corresponding to any admissible control u is given by

$$x(t;u) = \Phi(t,t_o)x^o + \int_{t_o}^{t_1} \Phi(t_1,s)F(s)u(s)ds. \tag{32}$$

Hence, by (32) we have for any admissible control $u = u(t)$,

$$Kx(t_1) = K\Phi(t_1,t_o)x^o + \int_{t_o}^{t_1} K\Phi(t_1,s)F(s)u(s)ds. \tag{33}$$

In order to satisfy $Kx(t_1) = d$, we see from (33) that

$$\int_{t_o}^{t} K\Phi(t,s)F(s)u(s)ds = \omega \tag{34}$$

where $\omega = d - K\Phi(t,t_o)x^o$ is a fixed q vector of scalars.

If we define a linear operator C from the space of admissible controls into the Hilbert space R^p such that

$$Cu = \int_{t_o}^{t_1} K\Phi(t_1,t)F(t)u(t)dt, \tag{35}$$

u must satisfy

$$Cu = \omega \tag{36}$$

in order to satisfy $Kx(t_1) = d$. It is known that if

$$\int_{t_o}^{t_1} \int_{t_o}^{t_1} \left\| F^T(s)\Phi(s,t_1)K^T K\Phi(t_1,t)F(t) \right\| dsdt < \infty \tag{37}$$

then C is a continuous linear operator. Since the components of K are scalars and $\Phi(t_1,t)$ and $F(t)$ have continuous components on

(t_o,t_1), it follows that the components of $K\Phi(t_1,t)$ $F(t)$ are bounded, hence (37) holds.

Hence we know that C as given by (35) is bounded. Then our optimal control problem as given by (31) becomes: from the set of admissible controls which satisfy $Cu = \omega$ given in (34), find that control which minimizes

$$J\left[u\right] = \int_{t_o}^{t_1} L(x(t;u),u(t),t)dt$$

where $x(t;u)$ is given by (32). That is, we wish to minimize the differentiable function $J\left[u\right]$ subject to the equality constraint $Cu = \omega$ for the bounded linear operator C given by (35). In section 2 of chapter 5, this type of problem was examined and the application of the basic algorithm to compute the solution was explained.

7. AN EXAMPLE, CONCLUSION, RECOMMENDATIONS, AND SUMMARY

7.1 Example

In order to exhibit the convergence characteristics of the algorithm, we formally applied the procedures of chapter 2 to a sample optimal control problem which others have used to display convergence characteristics of other algorithms [23, 34, 36]. The problem is the following: Find the function $u = u(t)$ which minimizes

$$J = \int_0^5 \left(x_1^2 + x_2^2 + u^2 \right) dt \tag{1}$$

subject to constraining differential equations described by the Van der Pol equation [35] with $\epsilon = 1$, that is

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 + \left(1 - x_1^2\right) x_2 + u \end{aligned} \tag{2}$$

with initial conditions

$$x_1(0) = 3.0$$
$$x_2(0) = 0.0.$$

By (6.19) the gradient g of J at u is given by

$$g(t) = 2u(t) + \lambda_2(t) \tag{3}$$

where

$$\begin{aligned} \dot{\lambda}_1 &= (1 + 2x_1x_2)\lambda_2 - 2x_1 \\ \dot{\lambda}_2 &= -\lambda_1 - (1 - x_1^2)\lambda_2 - 2x_2 \end{aligned} \qquad (4)$$

with

$$\lambda_1(5) = 0.0$$

$$\lambda_2(5) = 0.0$$

In order to compute the gradient $g(t)$ of J at some $u = \tilde{u}(t)$, we integrate (2) forward to $t = 5.0$ using $u = \tilde{u}(t)$. Next, (4) is integrated from $t = 5.0$ back to $t = 0.0$. Then using $u = \tilde{u}(t)$ and the computed value of λ_2, we can compute $g(t)$ given by (3).

Figures 1 and 2 depict the progress toward the minimum of J using the algorithm outlined in chapter 2 with four different methods of choosing α_n. These four methods of choosing α_n are:

Method 1: $\alpha_n = 1 - (n^3 + 2)^{-1/2}$ for all n

Method 2: $\alpha_n = 1$ for all n

Method 3: $\alpha_n = \min\left\{(-J(u_n) + \tilde{J}_0)/(s_n, g_n), 1.0\right\}$ where $\tilde{J}_0$ is the estimated minimum value of J, s_n is defined by (2.2) and g_n is the gradient of J at $u = u_n(t)$.

Method 4: α_n is the minimum with respect to α of $J(x_n + \alpha s_n)$ as computed by Davidon's one dimensional cubic minimization method [8].

Methods 1 and 2 of choosing α_n satisfy the condition that $(1 - \alpha_n)n \to 0$ as $n \to \infty$ given in theorem 2.11. As chosen by method

3, α_n is a rough estimate of the minimum of J along the line $u_n + s_n$. The form of α_n for method 3 follows by considering $\tilde{J}_o = J(u_n) + \alpha_n(s_n, g_n) + \text{h.o.t.}$, dropping the higher order terms, and solving for α_n.

Notice that methods 1, 2, and 3 of choosing α_n involve no extra functional and gradient evaluations. That is, for each iteration we must integrate (2) and (4) only once. For the fourth method of choosing α_n, although the one dimensional minimum is computed more accurately than by method 3, the fourth method involves at least one more functional evaluation per iteration. Hence, with the fourth method of choosing α_n, we have at least two functional and gradient evaluations per iteration.

In Figure 1, we have plotted $J(u_n)$ versus n (i.e., the iteration number) for the four different methods of choosing α_n. Figure 1 shows that the fastest convergence in terms of iterations is achieved by the algorithm with α_n chosen by method 4. Also, Figure 1 shows that after 12 iterations, all the methods have converged. Moreover, after eight iterations for all methods of choosing α_n the change in the value of J is too small to show up in the graph.

In Figure 2, we have plotted J versus the number of functional evaluations. Notice that in Figure 2, methods 3 and 1 converge faster with respect to function evaluations than method 4. Note also that after at most eight functional evaluations, the change in J is too small to be noticed in the graph.

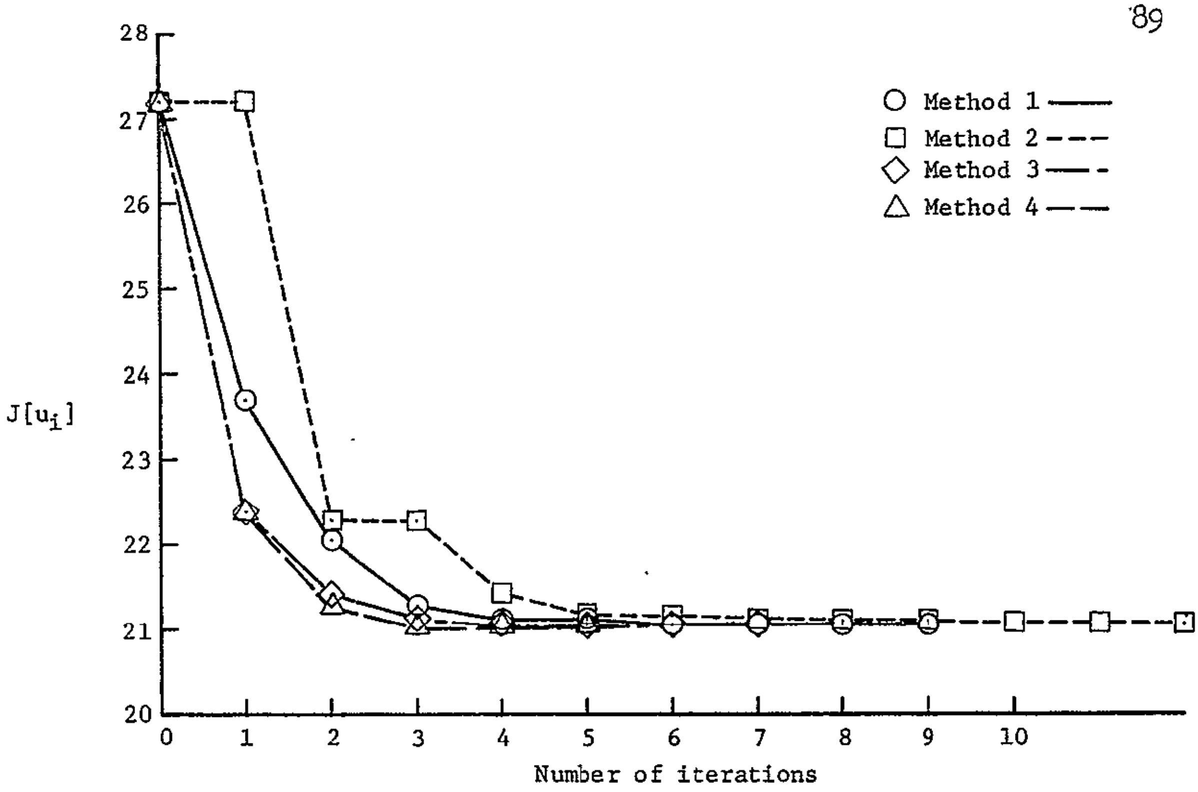

Figure 1. $J(u_i)$ versus i for the four methods of choosing α_i

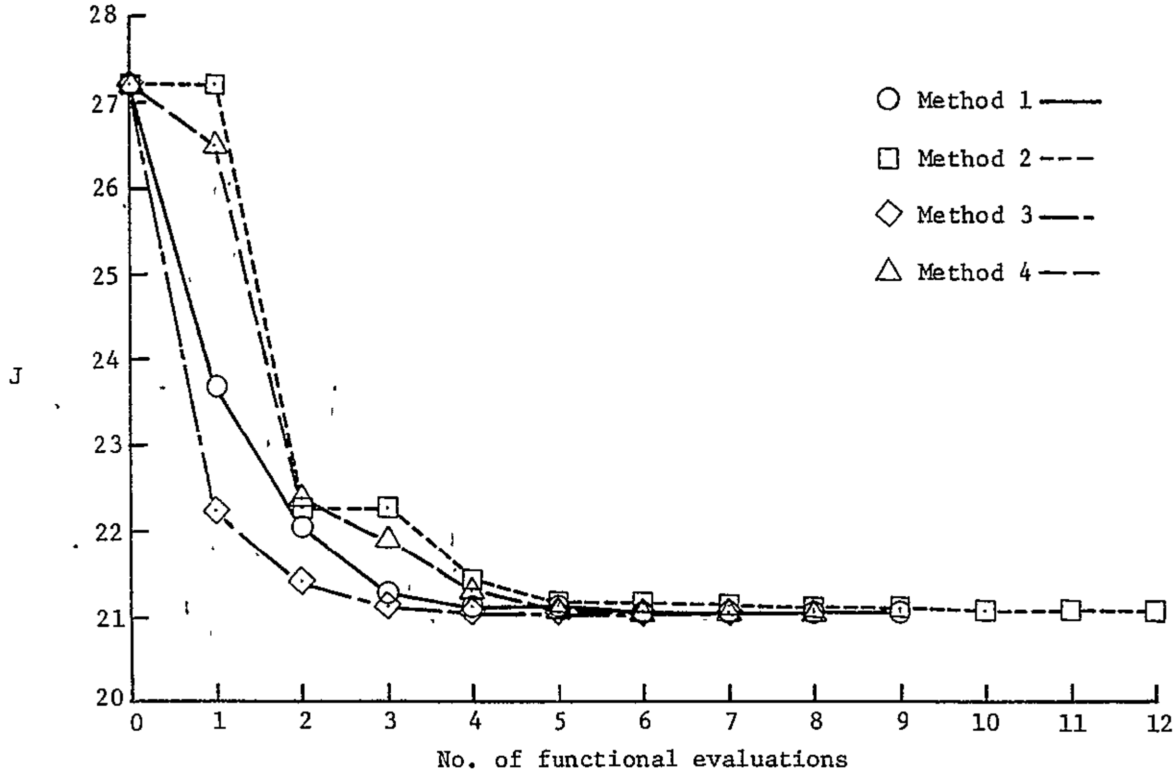

Figure 2. $J(u_i)$ versus the number of function evaluations for the four methods of choosing α_i

Figure 3 shows the rates of convergence to the minimum for the example problem for the three first-order methods given in chapter 1. These results were reported by Tokumaru, *et al.*, [36]. Note that the DFP algorithm shows the fastest rate of convergence

Using the same initial estimate of u that we used for the results shown in Figure 1, we applied the DFP method to the example problem. Our results for the DFP method were identical to those of the rank one algorithm with α_n chosen by method four. The reduction in the payoff and the iterates for the two methods were the same.

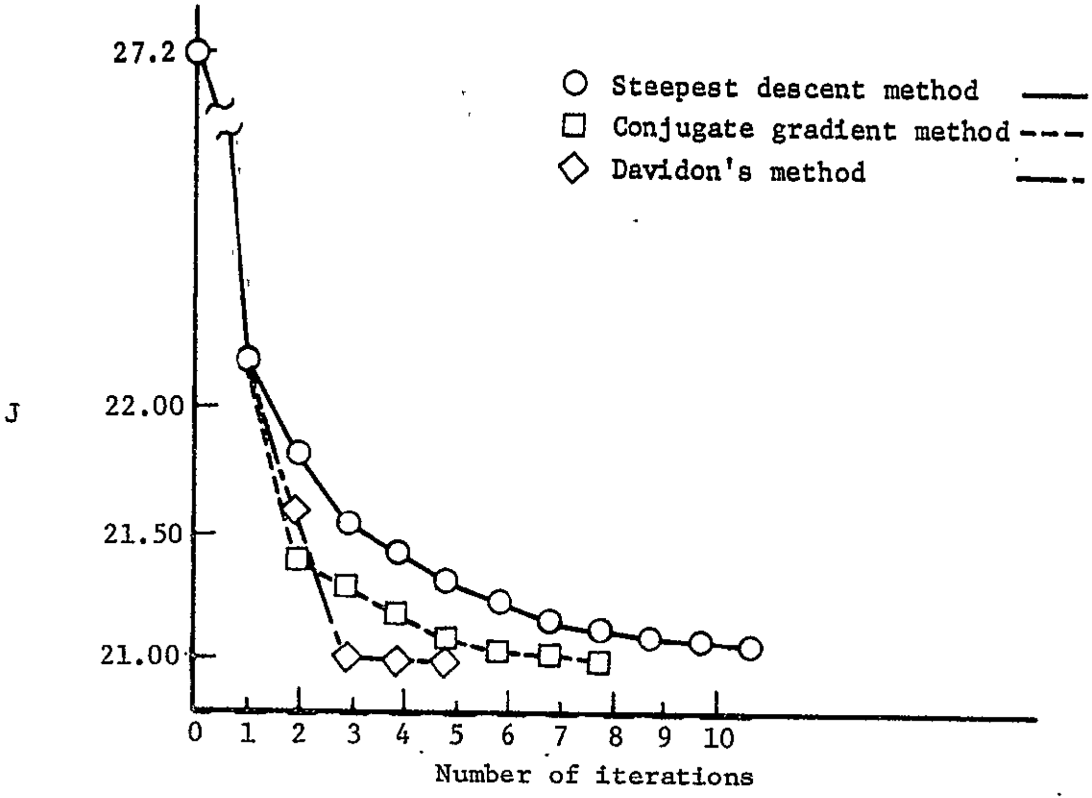

Figure 3. Comparison of first-order methods due to Tokumaru $J(u_i)$ versus i

In Figure 4, we have plotted the values of $J(u_i)$ versus the number of function evaluations for the DFP method and our algorithm when α_n is chosen by method 3. Notice that in terms of function evaluations, our method for this choice of α_n converges faster than the other algorithm. The linear minimizations for the DFP algorithm were carried out by method 4. This method was chosen because high accuracy for the linear minimization is necessary for the DFP method.

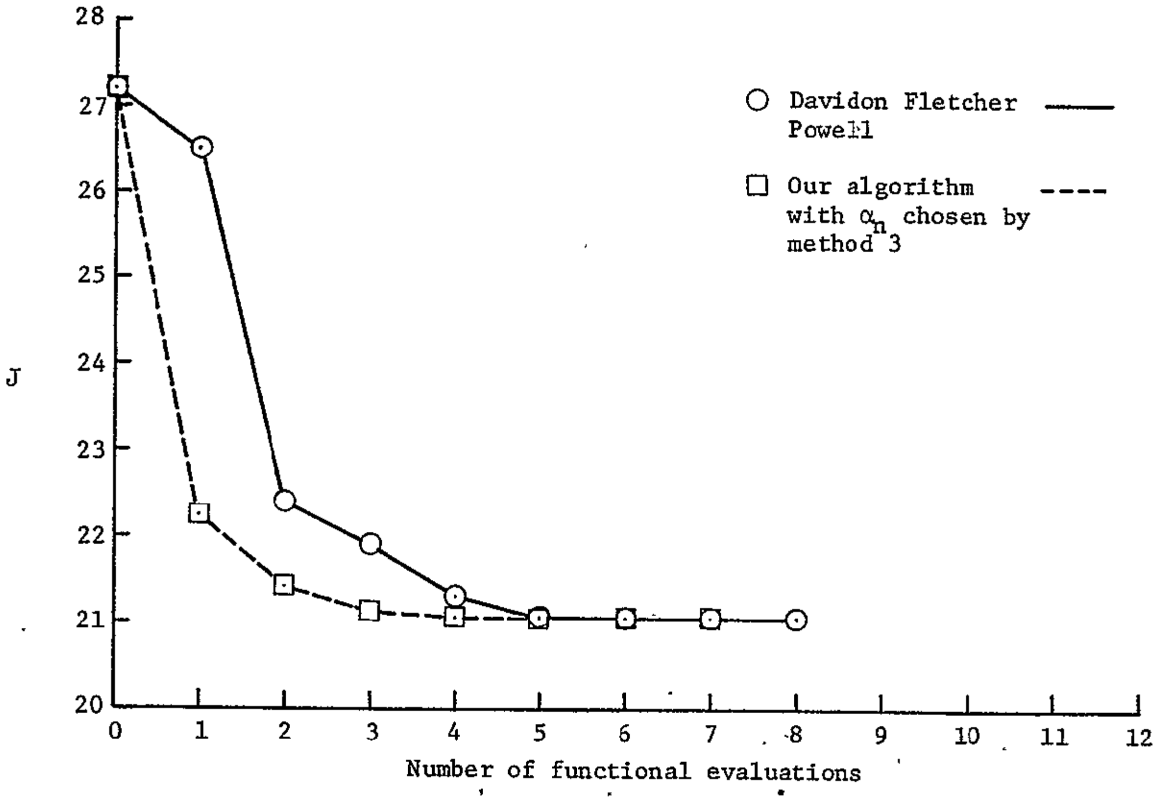

Figure 4. Comparison of Davidon-Fletcher-Powell method and Rank One method with α_n chosen by the third method with $J(u_i)$ versus function evaluations

In Figure 5, we have plotted the iterates of the control $u_i(t)$ for $i = 0,1,2,3$ of our algorithm. The integrations of (2) and (4) were carried out by the Adams-Bashforth predictor and Adams-Moulton corrector method on a CDC 6000 series computer with step size of 0.03125.

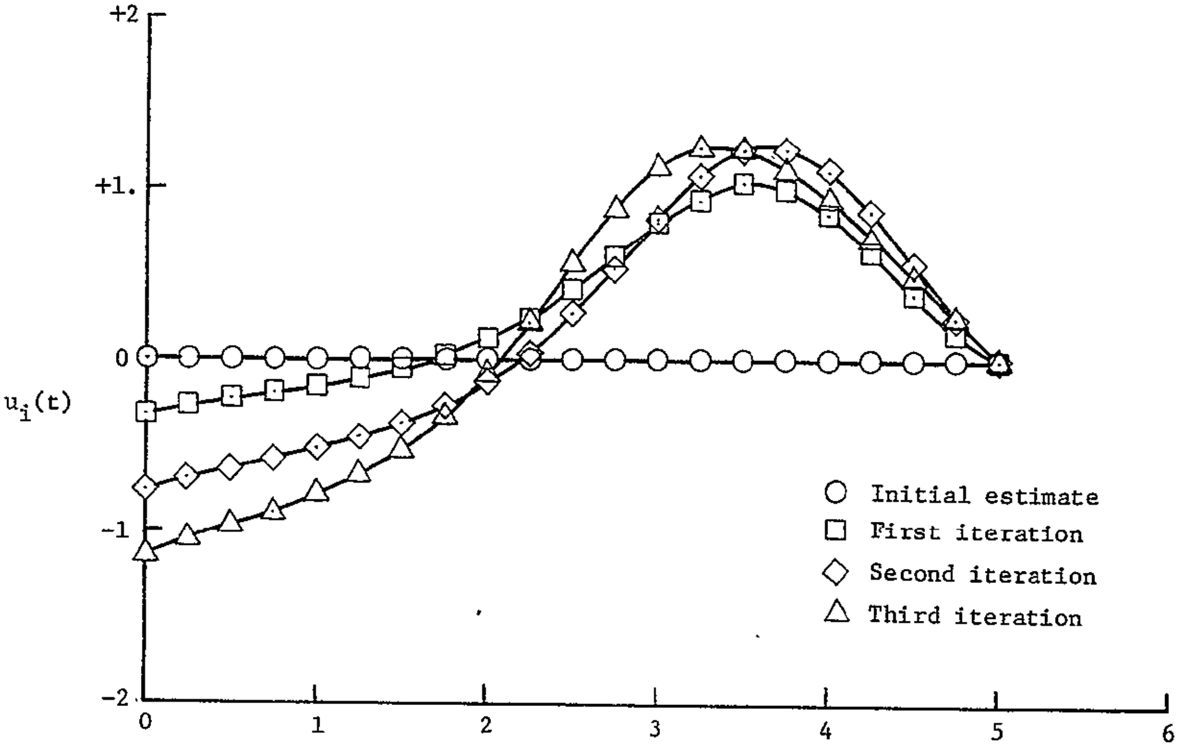

Figure 5. u_i versus t for $i = 0,1,2,3$ generated by Rank One algorithm with α_n chosen by method 4

7.2 Conclusion

The algorithm outlined in chapter 2, when applied to compute the location of the minimum of a quadratic functional, has several attractive properties. Theorem 3.3 shows that if α_n is chosen by (2.17), then our algorithm, the DFP and conjugate gradient methods generate the same iterates. Hence, the methods will have the same rates of convergence if the hypothesis of theorem 3.3 hold. Moreover, by theorem 2.8 $V^{(n)} \to A^{-1}$ pointwise where $V^{(n)}$ is given by (2.9) and A is given by (1.3). This property can be used to accelerate the convergence when many solutions corresponding to different initial conditions are desired. This was discussed in section 1 of chapter 6. This property is not available to the method of conjugate gradients. Theorem 3.3 shows that if α_n is chosen by the fourth method, then our algorithm, the DFP, and the conjugate gradient methods generate the same iterates, hence, the same rates of convergence. Moreover, our algorithm requires one-half the storage necessary for the DFP method. Also, it requires the computation of one operator per iteration versus the computation of two operators per DFP iteration.

The results of the example problem show that the algorithm can be applied with success when α_n is chosen in a variety of ways. It appears that method 3 of choosing α_n is best when the functional to be evaluated is very complex, its computation is time-consuming, and storage considerations are not as important. If storage considerations are pressing and the computation of the functional is not as time-consuming, then method 4 would seem to be the best choice for α_n.

7.3 Recommendations

Possible research topics related to this work are the following: (1) Research could be done on the application of the algorithm outlined in chapter 2 to the solution of the singular linear operator equation

$$Kx = d. \tag{5}$$

Hereby in (5), $x \in H$, a real Hilbert Space, $K:H \rightarrow \bar{H}$ is linear, bounded and has a closed range and d is fixed element of $\bar{H}$, another real Hilbert Space. Nashed [31] has discussed solving this problem using the method of steepest descent to compute at least squares solution. So it appears that the problem could be solved by our algorithm. By using theorem 2.8, perhaps it could be shown that $V^{(n)}K^*$ converges pointwise to the generalized inverse of K. In a finite dimensional space this could perhaps give another technique for computing the generalized inverse of K. (2) Research could be done to extend to an infinite dimensional real Hilbert Space the class of first-order algorithms recently proposed by Greenstadt [14] .

7.4 Summary

The various elements of the class of rank one, quasi-Newton minimization methods are distinguished by the manner in which a particular parameter is chosen at each iteration. In chapter 2, conditions were found which guarantee that the rank one, quasi-Newton algorithms generate iterates which converge to the location of the minimum of a quadratic functional for various choices of this parameter. In chapter 3, the iterates of the rank one, quasi-Newton algorithm with the parameter chosen by a linear minimization technique are compared with the iterates of the Davidon-Fletcher-Powell method and method of conjugate gradients. It is found that for a quadratic functional with the hypothesis of theorem 3.3 that the iterates of the three methods are the same. In chapter 4, an idea due to Powell is extended to infinite dimensional Hilbert spaces. In chapter 5, a modification of the rank one, quasi-Newton method is outlined in order to minimize a functional subject to linear constraints. Conditions are found which guarantee the convergence to the location of the constrained minimum of a quadratic functional. The application of these rank one, quasi-Newton minimization methods to various types of optimal control problems is investigated in chapter 6. In chapter 7, the rank one, quasi-Newton methods are applied to a sample optimal control problem. The results are compared with the results of other known first-order minimization techniques for the same sample problem. This comparison is in terms of speed of convergence with respect to iterations and number of functional evaluations. The rank one, quasi-Newton algorithms are shown to be superior.

8. LIST OF REFERENCES

1. Akhiezer, N. I., and Glazman, I. M. 1961. Theory of Linear Operators in Hilbert Space. Frederick Unger Publishing Co., New York.

2. Berberian, S. K. 1961. Introduction of Hilbert Space. Oxford University Press, New York.

3. Bliss, G. A. 1935. Calculus of Variations. Open Court Publishing Co., LaSalle, Illinois.

4. Broyden, C. G. 1967. Quasi-Newton Methods and Their Application to Function Minimization. Math. Comp. 21:368-381.

5. Coddington, E. E., and Levinson, N. 1955. Theory of Ordinary Differential Equations. McGraw-Hill Book Company, Inc., New York.

6. Daniel, J. W. 1967. The Conjugate Gradient Method for Linear and Nonlinear Operator Equations. J. SIAM Num. Anal. 4 (1):10-26.

7. Davidon, W. C. 1959. Variable Metric Method for Minimization. AEC Research and Development Report ANL 5990 (Rev. TID 4500, 14th edition).

8. Davidon, W. C. 1968. Variance Algorithm for Minimization. Computer J. 10:406-410.

9. Davidon, W. C. 1969. Variance Algorithms for Minimization. pp.13-20. In R. Fletcher (ed.) Optimization. Academic Press, London.

10. Fletcher, R., and Powell, M. J. D. 1963. A Rapidly Convergent Descent Method for Minimization. Computer J. 6:1963-1968.

11. Fletcher, R., and Reeves, C. M. 1964. Function Minimization by Conjugate Gradients. Computer J. 7:1949-1954.

12. Friedman, B. 1956. Principles and Techniques of Applied Mathematics. John Wiley and Sons, Inc., New York.

13. Goldfarb, D. 1969. Sufficient Conditions for the Convergence of a Variable Metric Algorithm. pp. 273-283. In R. Fletcher (ed.) Optimization. Academic Press, London.

14. Greenstadt, J. 1970. Variations on Variable-Metric Methods. Math. of Computation 24:1-23.

15. Hanafy, L. M. 1970. The Linear Time Optimal Control Problem from a Calculus of Variations Point of View. NASA Contractor Report, NASA CR-1612.

16. Hermes, H., and LaSalle, J. P. 1969. Functional Analysis and Time-Optimal Control. Academic Press, New York.

17. Hestenes, M. R. 1966. Calculus of Variations and Optimal Control Theory. John Wiley and Sons, Inc., New York.

18. Hestenes, M. R. 1969. Multiplier and Gradient Methods. Journal of Optimization Theory and Applications 4(5):303-320.

19. Hestenes, M. R., and Steifel, E. 1952. Methods of Conjugate Gradients for Solving Linear Systems. Journ. of Res. NBS. 49 (6):409-436.

20. Horwitz, L. B., and Sarachik, P. E. 1968. Davidon's Method in Hilbert Space. SIAM J. Appl. Math. 16 (4):676-695.

21. Horwitz, L. B., and Sarachik, P. E. 1969. A Survey of Two Recent Iterative Techniques for Computing Optimal Control Signals. Proceedings of 10th Joint Automatic Control Conference. Boulder, Colorado:50-51.

22. Kantorovich, L. V., and Akilov, G. P. 1964. Functional Analysis in Normed Spaces. MacMillan, New York.

23. Lasdon, L. S., Mitter, S. K., and Waren, A. D. 1967. The Conjugate Gradient Method for Optimal Control Problems. IEEE Trans. Aut. Cont. AC-12 (2):132-138.

24. Leondes, C. T., and Niemann, R. A. 1969. On the Optimal Control Problem With Unrestricted Final Time. Journ. of Basic Engineering 91 (2):155-160.

25. Luenberger, D. G. 1969. Optimization by Vector Space Methods. John Wiley and Sons, Inc., New York.

26. Lusternik, L. A., and Sobelev, V. J. 1961. Elements of Functional Analysis. Frederick Unger Publishing Company, New York.

27. Myers, G. E. 1968. Properties of the Conjugate-Gradient and Davidon Methods. J. Opt. Theory Appl. 2 (2):209-219.

28. Park, S. K. 1970. On the Equivalence of the Optimal Control Problems and the Transformation of Optimal Control Problems With Compact Control Regions Into Lagrange Problems. Unpublished Ph.D. Thesis Department of Mathematics, North Carolina State University at Raleigh. University Microfilms, Ann Arbor, Michigan.

29. Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., and Mishchenko, E. F. 1962. The Mathematical Theory of Optimal Processes. Interscience Publishers, New York.

30. Powell, M. J. D. 1970. A Survey of Numerical Methods for Unconstrained Optimization. SIAM Review 12 (1):79-97.

31. Nashed, M. Z. 1970. Steepest Descent for Singular Linear Operator Equations. SIAM J. Numer. Anal. 7 (3):358-363.

32. Rosen, J. B. 1961. The Gradient Project Method for Nonlinear Programing II, Nonlinear Constraints. J. SIAM Appl. Math. 9:514-532.

33. Sagan, H. 1969. Introduction to the Calculus of Variations. McGraw-Hill Book Co., New York.

34. Sinnott, J. F., Jr., and Luenberger, D. G. 1967. Solution of Optimal Control Problems by the Method of Conjugate Gradients. Preprints of 1967 Joint Automatic Control Conference: 566-573.

35. Struble, R. A. 1962. Nonlinear Differential Equations. McGraw-Hill Book Co., New York.

36. Tokumaru, H., Adachi, N., and Goto, K. 1970. Davidon's Method for Minimization Problems in Hilbert Space With an Application to Control Problems. SIAM J. Control 8 (2):163-179.