

https://dspace.org/

The method of centres

Jordi, Kendal Clive

2015-02-06
http://hdl.handle.net/10539/16912

Downloaded from a DSpace repository. https://dspace.org/

THE METHOD OF CENTRES.

Kendal Clive Jordi

THE METHOD OF CENTRES.

Kendal Clive Jordi

An essay submitted to the Faculty of Science in partial fulfilment of the requirements for the degree of Master of Science.

University of the Witwatersrand,
Johannesburg.

1975.

Abstract.

A very general Method of Centres algorithm with proven convergence is considered. The Theoretical Method of Centres of Huard, the Method of Centres with Upper Bounding Functions of Huard, the Linearized Method of Centres of Huard and a modification by Pironneau and Polak are particular cases of the general algorithm.

An extension due to Hoshino to the Q-function of Fiacco and McCormick is then considered. Using a function defined by Hoshino it is then shown that the Method of Centres algorithm can be applied to a certain class of problems with equality constraints.

INTRODUCTION.

In this essay we are concerned with the problem of minimizing some function $f^0(x)$: $R^n \to R$ where the variables $x \in R^n$ are constrained to belong to some non-empty set $A \subset R^n$.

The method that we are interested in here, to obtain the solution to this problem computationally, is referred to as the Method of Centres. The Method of Centres operates with moving truncations of the *constraint set*. Convergence to a minimum solution of the problem is controlled by a sequence $\{f^0(x_k)\}$ of truncation levels. Each of the points (centres) x_k is obtained by the unconstrained maximization of a distance function which characterizes some distance from the point x_k to the boundary of the truncated constraint set.

The Method of Centres was initially proposed in 1964 by Huard [11]. It was then explored theoretically and computationally by Faure and Huard [3],[4], Bui-Trong-Lieu and Huard [1], Huard [12],[13] and Tremolieres [22]. Mifflin [16], Denel and Huard [2], Pironneau and Polak [18] proposed modifications to improve the rate of convergence. Lootsma [15] pointed out that it is a variant of a barrier (penalty) function method without any arbitrary parameters. One of the distance functions provides the moving truncations counterpart of the Logarithmic Barrier Function technique originally proposed by Frisch [8]. Huard [14] treated the convergence of this method using approximate centres, and he related this method to certain other well known methods. Polak [19] described various possible versions of this method and included some numerical examples. Fiacco and McCormick [5], [6] developed a distance function which they called the Q-function.

Hoshino [10] proposed an extension of the Q-function to include a parameter which is chosen to accelerate convergence. Polak, Mukai and Pironneau [20] applied the Method of Centres to solve the optimal control problem.

In this essay, after some preliminary assumptions and definitions, we state a general method of centres algorithm and we prove that it converges. Special cases of the general algorithm are then considered. The first particular case is the Theoretical Method of Centres of Huard [12], which is a generalization of the algorithm initially proposed in [11]. Pironneau and Polak [18] showed that this algorithm has linear convergence. This algorithm requires the maximization of the distance function which can be quite time consuming. We truncate the search for a point x' which maximizes the distance function by searching for a point which is an ε distance from x' by using approximate centres.

We then consider simpler upper bounding functions. Approximate centres are considered again. The variant obtained by considering upper bounding functions is still a fairly general algorithm, since by particularizing it still further Huard [14] gets at the Linearized Method of Centres of Huard [13] as well as other well-known methods. We then consider the modification to the Linearized Method of Centres proposed by Piron-eau and Polak [19]. This has the effect of increasing the rate of convergence.

The Q-function investigated by Fiacco and McCormick [5],[6] is then extended to define a new distance function proposed by Hoshino [10]. Hoshino asserts that the new distance function demonstrates a marked improvement in convergence over the distance function proposed in [12].

Pietrykowski [17] considered an exact potential method for constrained minima. This has a major disadvantage from a computational point of view. This was overcome by Hoshino [10] who considered a new function which enables us to use the Method of Centres algorithm to solve a certain class of problems which have inequality constraints. The result of Hoshino's broadens the scope and range of applicability of of the Method of Centres algorithm considerably.

1. DEFINITIONS, ASSUMPTIONS AND PRELIMINARY RESULTS.

STATEMENT OF PROBLEM

The mathematical programming problem is to determine a vector $x^* = (x_1^*, x_2^*, \ldots, x_n^*)^T$ that solves the problem

(P.1)

$$\text{minimize} \quad f^0(x) \tag{1.1}$$

$$\text{subject to} \quad f^i(x) \leq 0 \tag{1.2}$$

where restrictions on the functions $f^i: R^n \to R^1$, $i = 0,\ldots,m$ will be defined later.

Define the constraint set by

$$C \triangleq \{ x : f^i(x) \leq 0, \; i = 1,\ldots,m \} \tag{1.3}$$

ASSUMPTIONS

A1.) We assume that there is a point $x_0 \in C$ such that the set $C'(x_0)$ defined by

$$C'(x_0) \triangleq \{ x : f^0(x) - f^0(x_0) \leq 0, \; f^i(x) \leq 0, \; i = 1,\ldots,m \} \tag{1.4}$$

is compact and has an interior.

This assumption is essential, since without it the method of centres cannot be applied to problem (P.1). It also ensures that a solution to the problem exists.

A2.) We assume that

(i) the set C defined in (1.3) has an interior and that the closure of the interior of C is equal to C.

(ii) for every $x \in C^0$, the interior of C, $f^i(x) < 0$, $i = 1,\ldots,m$.

The reason for A2.(i) is that given a point $x_o \in C$, the Method of Centres picks as its successor a point x_1 in the interior of the set $C^1(x_o)$, and hence it can find an optimal point for (P,1) only if that point is in the closure of the interior of C. The reason for A2.(ii) will become clear later when we define an " F-distance".

AN F-DISTANCE

Let Σ be a set whose elements are the subsets of R^n and let $d: R^n \times \Sigma \to R^1$ be a real function.

Notation: The interior of a set $E \in \Sigma$ is denoted E^o, and the boundary is denoted by $Fr(E)$.

Definition 1.1: d is called an F-distance on $R^n \times \Sigma$ if it satisfies

(i) $d(x,E) = 0 \quad \forall E \in \Sigma, \forall x \in Fr(E)$,

(ii) $d(x,E) > 0 \quad \forall E \in \Sigma, \forall x \in E^o$,

(iii) $\forall E \in \Sigma, \forall E' \in \Sigma: E \subset E'$, then, $\forall x \in E$, there exists a scalar $\rho(x) > 0$ such that

$$d(x,E) \leq \rho(x).d(x,E').$$

Definition 1.2: An F-distance d is said to be regular if it satisfies in addition

(iv) $\forall$ sequences $\{ E_k \in \Sigma : k \in N \}$ and $\{ x_k \in R^n : k \in N \}$ such that

$$E_k \supset E_{k+1} \supset K \in E, \quad E \neq \phi$$

$$x_k \in E_k, \quad x_k \notin E^o_{k+1}$$

we have that $d(x_k,E_k) \to 0$ as $k \to \infty$

Examples of regular F-distances.

Let assumptions A1 and A2 hold, and let $f^i: R^n \to R^1$, $i = 0,..,m$ be continuous real functions. Define

$$E \triangleq \{ C'(x_0) : f^0(x_0) \in K \subset R \} \tag{1.5}$$

$$K \triangleq (f(x^*), \sup \{ f(x) : x \in C \} \tag{1.6}$$

where x^* is the solution to (P.1), which exists by assumption. Let

$$E_{k+1} \triangleq C_k \triangleq C'(x_k)$$

Then each of the following functions :

$$\text{(i)} \quad d(x,C'(x_0)) = \min\{ f^0(x_0) - f^0(x), \ -f^i(x), \ i = 1,...,m \}, \tag{1.7}$$

$$\text{(ii)} \quad d(x,C'(x_0)) = (f^0(x_0) - f^0(x)) \prod_{i=1}^{m} (-f^i(x)) \tag{1.8}$$

is a regular F-distance, defined on $R^n \times E$, where $x_0 \in C$.

Proof: The following proof holds for both the functions defined above. Let $x_0 \in C$.

(i) $d(x,C'(x_0)) = 0$ for all $x \in Fr(C'(x_0))$ by inspection.

(ii) Let $x \in (C'(x_0))^0$. Then, by A2.(ii),

$$f^0(x) - f^0(x_0) < 0, \text{ and } f^i(x) < 0, \ \forall\, i = 1,...,m$$
$$\Rightarrow f^0(x_0) - f^0(x) > 0, \text{ and } f^i(x) > 0, \ \forall\, i = 1,...,m$$
$$\Rightarrow d(x,C'(x_0)) > 0.$$

(iii) $$C'(x_0) \subset C'(x_0')$$
$$\Rightarrow f^0(x_0) \leq f^0(x_0')$$
$$\Rightarrow f^0(x_0) - f^0(x) \leq f^0(x_0') - f^0(x), \ \forall\, x \in C'(x_0)$$
$$\Rightarrow d(x,C'(x_0)) \leq d(x,C'(x_0')), \ \forall\, x \in C'(x_0).$$

(iv) Consider an infinite sequence $\{ x_k \in C : k \in N \}$ and the sequence of corresponding $C_k = C'(x_k)$ such that, $\forall\, k \in N$,

$$f^0(x_k) \geq f^0(x_{k+1}) \geq f^0(x^*)$$

where x^* minimizes $f^0(x)$, $\forall\, x \in C$. Then

$$x_{k+1} \in C_k \triangleq E_{k+1}, \; x_{k+1} \in C^0_{k+1} \triangleq E^0_{k+2}.$$

Now, because $f^0(x)$ has a lower bound,

$$f^0(x^*) > -\infty$$

Hence,

$$f^0(x_k) \to f^0(x) \geq f^0(x^*) \quad \text{as} \quad k \to \infty, \text{ and}$$
$$f^0(x_k) \geq f^0(x) \quad \forall\, k \in N,$$

Let

$$E \triangleq C'(x) \neq \phi \text{ , because } x^* \in C'(x).$$

Therefore,

$$E_k \supset E_{k+1} \supset E.$$

We now have

$$f^0(x_k) - f^0(x_{k+1}) \to 0 \quad \text{as} \quad k \to \infty, \; k \in N$$
$$\Rightarrow \; d(x_{k+1}, C'(x_k)) \to 0 \quad \text{as} \quad k \to \infty, \; k \in N$$
$$\Rightarrow \; d(x_{k+1}, E_{k+1}) \to 0 \quad \text{as} \quad k \to \infty, \; k \in N.$$

Assume that the functions $f^i(x)$, $i = 0,\ldots,m$ have continuous derivitives. The derivitive of the function defined in (1.7) is not continuous in general. Therefore this function is not convenient for numerical maximization, but the function defined in (1.8) gives a suitable distance function that has continuous derivitives.

If the $f^i(x)$, $i = 0,\ldots,m$ are convex functions, then the function defined in (1.7) is concave, while the function defined in (1.8) is not necessarily concave. The expression given in (1.8) is analogous, up to a logarithm, to the logarithmic potential of the well known Method of Frisch [8], but its behavior in the process of computation is very different, the derivitives being taken into consideration not being infinite.

2. A GENERAL METHOD OF CENTRES ALGORITHM.

The algorithm which is presented in this section will play a central role in the next few chapters, where certain particular cases will be discussed. The algorithm, which is stated below, is not as general as the algorithm described in Huard [14].

We will assume that A1 and A2 hold. Consider the following problem:

$$(P.2) \qquad \begin{array}{l} \text{minimize} \quad f^0(x) \\ \text{subject to} \quad x \in C \cap B \end{array}$$

where f^0 is a continuous function $\forall\, x \in C \cap B$, C is defined in (1.3) and $B \subset R^n$ is some set such that

(i) $C^0 \cap B \neq \phi$,

(ii) there exists an $x_0 \in C$ such that $C^1(x_0) \cap B$ is compact and has an interior.

The following hypothesis is made for C and B:

(H) $C^0 \cap B \cap A = \phi \Rightarrow C \cap B \cap A = \phi$, $\forall\, A \subset R^n$ open.

Let d be a regular F-distance.

Algorithm

0. Set $k = 0$. Select $x_0 \in C \cap B$ such that A1 is satisfied.

1. If $(C^1(x_k))^0 \cap B = \phi$, set $x^* = x_k$. Stop.
   Otherwise go to 2.

2. Determine an $x'_{k+1} \in C^1(x_k) \cap B$ such that
   $$d(x'_{k+1}, C^1(x_k)) \geq d(x, C^1(x_k)), \quad \forall\, x \in C^1(x_k) \cap B.$$

3. Determine $x_{k+1} \in C^1(x'_{k+1}) \cap B$.

4. Set $k = k+1$ and return to 1.

Theorem 2.1: Let the above assumptions hold. If the sequence $\{x_k\}$ constructed by the above algorithm is finite, then the last element is optimal for (P2). Otherwise, if the sequence is infinite it has accumulation points, all of which are optimal for (P2).

Proof: The algorithm stops if $(C'(x_k))^0 \cap B = \phi$,

$\Rightarrow C^0 \cap \{ x : f^0(x) - f^0(x_k) < 0 \} \cap B = \phi$,

$\Rightarrow C \cap \{ x : f^0(x) - f^0(x_k) < 0 \} \cap B = \phi$ by hypothesis

$\Rightarrow$ there does not exist an $x \in C \cap B$ such that

$$f^0(x) < f^0(x_k).$$

Hence x_k is an optimal solution.

By inspection we must have $C'(x_{k+1}) \subset C'(x_k)$, $k = 0,1,2,\ldots$. Since $C'(x_0)$ is compact and since $x_{k+1} \in C'(x_k) \subset C'(x_0)$, the infinite sequence $\{x_k\}$ must contain accumulation points. Suppose there exists some subsequence which converges to the point $\hat{x}$, i.e. $x_k \to \hat{x}$ for $k \in K \subset \{0,1,2,\ldots\}$. We show that $\hat{x}$ is a solution to problem (P.2).

We have from Steps 2 and 3 that

$$f^0(x_k) > f^0(x'_{k+1}) \geq f^0(x_{k+1}) \geq f^0(x^*)$$

where x^* is a solution to problem which exists by assumption A1. Therefore, by the continuity of f^0 we have that the sequence $f^0(x_k)$ tends to the limit

$$f^0(\hat{x}) \geq f^0(x^*)$$

as k increases, and

$$f^0(x_k) \geq f^0(\hat{x}), \ \forall\, k. \tag{2.1}$$

We now prove that

$$(C'(\hat{x}))^0 \cap B = \phi. \tag{2.2}$$

If not, then there exists some $\bar{x} \in (C'(\hat{x}))^0 \cap B$. (2.3)

We have

$$E_k \supset E_{k+1} \supset C'(\hat{x})$$

$$x_k^1 \in E_k \cap B, \quad x_k^1 \notin E_{k+1}^0 \cap B.$$

Suppose $x_k \in (C^1(\hat{x}))^0 \cap B$, $\forall\, k \in N$. We show this is not possible. Since d is a regular f-distance,

$$d(x_k^1, E_k) \to 0, \text{ when } k \to \infty. \tag{2.4}$$

We have from (2.3),

$$\begin{aligned} 0 &< d(\overset{\circ}{x}, C^1(\hat{x})) \\ &\leq \rho_o \cdot d(\overset{\circ}{x}, E_k) \quad \text{with } \rho_o < 0 \text{ since } C^1(x) \subset E_k \\ &\leq \rho_o \cdot d(x_k^1, E_k) \quad \text{since } x_k^1 \text{ maximizes } d \text{ on } E_k \cap B. \end{aligned}$$

Therefore,

$$0 < d(\overset{\circ}{x}, C^1(\hat{x})) \leq \rho_o \cdot d(x_k^1, E_k)$$

which contradicts (2.4).

So there exists some finite $k^* \in N$ such that

$$x_k \in (C^1(\hat{x}))^0 \cap B, \quad \forall\, k \geq k^*.$$

This implies that there exists some finite $k^* \in N$ such that

$$f^0(x_k) \leq f^0(\hat{x}), \quad x_k \in (C^1(\hat{x}))^0 \cap B, \quad \forall\, k \geq k^*.$$

This contradicts (2.1). Therefore,

$$\begin{aligned} & (C^1(\hat{x}))^0 \cap B = \phi, \\ \Rightarrow\ & C^0 \cap \{\, x : f^0(x) - f^0(\hat{x}) < 0 \,\} \cap B = \phi, \\ \Rightarrow\ & C \cap \{\, x : f^0(x) - f^0(\hat{x}) < 0 \,\} \cap B = \phi, \text{ by (H)}, \\ \Rightarrow\ & f^0(x^*) \geq f^0(\hat{x}) \quad \text{since } C \cap B = \phi, \\ \Rightarrow\ & f^0(x^*) = f^0(\hat{x}). \end{aligned}$$

3. A THEORETICAL METHOD OF CENTRES: Huard [12].

The Method of Centres to be studied in this section, which is a generalization of the algorithm of Huard given in [11], is a conceptual algorithm, since, if we use the distance function defined in (1.7) the maximization of the distance function cannot be expected to be accomplished in a finite number of digital computer operations. However, it leads naturally to several implementable algorithms.

Consider the problem (P.1), where $f^i: R^n \to R$, $i = 0,\ldots m$ are continuous functions satisfying assumptions A1,A2. Let $B \subset R^n$ be compact, large enough to contain C.

Then we need the following hypothesis to be satisfied:

(H) $\quad C^o \cap A = \phi \implies C \cap A = \phi, \quad \forall A \subset R^n$, open.

In the following algorithm Huard uses the f-distance defined in equation (1.7). Let the point $x_k \in C$. Intuitively, for point $x \in C'(x_k)$ the function $d(x,C'(x_k))$ is the "distance that the point is from the closest boundary of the set $C'(x_k)$". To find the "middle" or centre of the set $C'(x_k)$ we need to maximize the distance function i.e. we need to find an x' such that

$$d(x',C'(x_k)) \geq d(x,C'(x_k)) \quad \forall\, x \in C'(x_k).$$

This process is equivalent to finding a solution $(d(x_{k+1},C'(x_k)),x_{k+1})$ of the problem

$$\max_{(d(x,C'(x_k)),x)} \{\, d(x,C'(x_k)) \;:\; f^o(x_k) - f^o(x) \geq d,\ -f^i(x) \geq d,\ i = 1,\ldots,m_1 \} \tag{3.1}$$

For a geometric interpretation see Fig. 3.1.

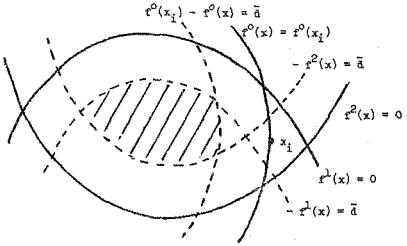


Figure 3.1. Suppose we have some distance $\bar{d}$. The shaded area represents the points which fall into the region

$$D = \{ x : f^0(x_i) - f^0(x) \geq \bar{d}, - f^i(x) \geq \bar{d}, i = 1,\ldots,m \}$$

Clearly $\bar{d}$ can still be increased, but if it is increased indefinitely, at some stage, $D = \phi$. The idea is to maximize $\bar{d}$ such that $D \neq \phi$.

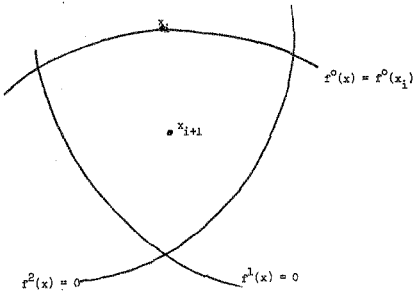


Figure 3.2.

In the following algorithm Steps 2 and 3 of the general algorithm are combined into one step i.e. x'_{k+1} and x_{k+1} are the same points.

Remark: If $d(x,C'(x_k)) = 0$, $\forall\, x \in C'(x_k))$ then $(C'(x_k))^o = \phi$ by inspection.

Algorithm

0. Set $k = 0$. Select $x_o \in C$ satisfying A1, A2.

1. Compute a solution $(d(x_{k+1},C'(x_k)),x_{k+1})$ to the problem defined in (3.1)

2. If $d(x_{k+1},C'(x_k)) = 0$, set $x^* = x_k$. Stop.
Otherwise, set $k = k+1$ and go to Step 1.

Theorem 3.1: Let the above assumptions hold. If the sequence $\{x_k\}$ constructed by the above algorithm is infinite, then it has accumulation points, all of which are optimal for (P.1). If the sequence $\{x_k\}$ is finite, then the last element is optimal for (P.1).

The above theorem follows as a result of theorem 2.1. Convergence can also be proved using a proof of Polak [19].

In addition to the previous assumptions in this section, assume that the functions $f^i(x)$, $i = 0,\ldots,m$ are all convex.

Remark: The convexity of the $f^i(x)$, $i = 0,\ldots,m$ implies that C is convex. This means that $\forall\, x_o$ such that $C'(x_o) \neq \phi$, $C'(x_o)$ is a convex set which is compact and has an interior by A1.

Theorem 3.2: Assume that A1, A2 are satisfied. Let the functions $f^i: R^n \to R$, $i = 0,\ldots,m$ be continuous and concave. Then

$$(C'(\hat{x}))^0 = \phi \Rightarrow C \cap \{ x : f^0(x) - f^0(\hat{x}) < 0 \} = \phi.$$

Proof: We prove the logical equivalent i.e.

$$C \cap \{ x : f^0(x) - f^0(\hat{x}) < 0 \} \neq \phi \Rightarrow (C'(\hat{x}))^0 \neq \phi. \tag{3.2}$$

Let $\overset{0}{x} \in C$ such that $f^i(\overset{0}{x}) < 0$, $i = 1,\ldots,m$ (3.3)
where $\overset{0}{x}$ exists by A1, A2. Let $\overset{0}{y} \in C \cap \{ x : f^0(x) - f^0(\overset{0}{x}) < 0 \}$
This exists by (3.2). Then,

$$f^0(\overset{0}{y}) - f^0(\overset{0}{x}) < 0 \tag{3.4}$$

$$f^i(\overset{0}{y}) \leq 0, \quad i = 1,\ldots,m \tag{3.5}$$

We assume (otherwise the theorem would be trivially verified),

$$\overset{0}{x} \neq \overset{0}{y}.$$

Let θ be any scalar, $0 \leq \theta \leq 1$, and $x(\theta)$ be a point defined by

$$x(\theta) = \overset{0}{x} + (1 - \theta)\overset{0}{y}.$$

The convexity of $f^i(x)$, $i = 1,\ldots,m$ implies that

$$f^i(x(\theta)) \leq \theta f^i(\overset{0}{x}) + (1 - \theta) f^i(\overset{0}{y}) \quad \forall\, i = 1,\ldots,m.$$

(3.3),(3.5) imply

$$f^i(x(\theta)) < 0. \tag{3.6}$$

The convexity of $f^0(x)$ implies that

$$f^0(x(\theta)) \leq \theta f^0(\overset{0}{x}) + (1 - \theta) f^0(\overset{0}{y})$$

$$= \theta (f^0(\overset{0}{x}) - f^0(\overset{0}{y})) + f^0(\overset{0}{y})$$

and by (3.4)

$$f^0(x(\theta)) - f^0(x) < 0 \quad \forall\, \theta : 0 \leq \theta \leq \bar{\theta} \tag{3.7}$$

where

$$\bar{\theta} = (f^0(\hat{x}) - f^0(\hat{y}))/(f^0(\hat{z}) - f^0(\hat{y})) > 0.$$

(3.6),(3.7) imply that

$$(C^1(\hat{x}_k))^0 \neq \phi.$$

The above result means that the hypothesis (H), used in the proof for convergence in Theorem 3.1, is automatically satisfied for convex functions.

Pironneau and Polak [18] have studied the rate of convergence of this algorithm with further restrictions. Their results are stated here in the form of two theorems, for completeness. The first theorem concludes that this Method of Centres algorithm converges at least linearly, the second, that it converges at most linearly under the additional assumption:

(A) there exists a compact convex set $C^u(x_o)$ containing $C^1(x_o)$ in its interior such that the function f^0 is strictly convex in $C^u(x_o)$.

This assumption implies that the solution to the problem is unique, where the solution exists by A1.

Theorem 3.3: Let $\{x_k\}$ be an infinite sequence generated by the last algorithm in solving problem (P.1), and suppose that the above assumptions are satisfied. Then, given any $\alpha \in (0,1)$, there exists an integer $k_o(\alpha)$ such that

$$f^0(x_{k+1}) - f^0(x^*) \leq (1 - \bar{\lambda}^0(1-\alpha))(f^0(x_k) - f^0(x^*)) \quad \forall\, k \geq k_o(\alpha)$$

where x^* is the unique solution of (P.1).

$$\bar{\lambda}^0 \triangleq \min \{ \langle \lambda, e \rangle : \sum_{i=0}^{m} \lambda^i \nabla f^i(x^*) = 0, \sum_{i=1}^{m} \lambda^i f^i(x^*) = 0$$

$$\sum_{i=0}^{m} \lambda^i = 1, \lambda \geq 0, \lambda \in R^{m+1} \} \qquad (3.8)$$

$$e \triangleq (1,0,0,\ldots,0) \in R^{m+1}$$

Note that the λ's are optimal multipliers.

Theorem 3.4: Let $\{x_k\}$ be an infinite sequence generated by the last algorithm in solving problem (P.1). Let x^* be the solution to the problem and let $\bar{\lambda}^0$ be defined as in (3.8). If the above assumptions are satisfied, then either $\bar{\lambda}^0 = 1$ and the sequence $\{f^0(x_k)\}$ converges superlinearly to $f^0(x^*)$, or $\bar{\lambda}^0 < 1$ and there exists an integer k_1 such that

$$f^0(x_{k+1}) - f^0(x^*) \geq (1 - \bar{\lambda}^0)(f^0(x_k) - f^0(x^*)) \quad \forall k \geq k_1.$$

THE USE OF ε - CENTRES.

In practice, approximate versions of the previous algorithm have been found to be rather inefficient, because even an approximate calculation of a x_{k+1} defined in (3.1) is quite time consuming. This has led to the development of methods which truncate the search for a point x satisfying (3.1) when one finds a point x' which is within a distance ε from the point x, where ε is made progressively smaller.

Definition 3.1: Given an f-distance d, defined on $R^n \times \Sigma$, a set $E \in \Sigma$ and a number ε such that $0 \leq \varepsilon \leq \sup \{ d(x,E) : x \in E \}$ we call the ε-centre of E (w.r.t. d) every point $x' \in E$ such that

$$d(x',E) \leq \sup \{ d(x,E) : x \in E \}$$

If $\varepsilon = 0$, *such a point is called a centre of* E.

Algorithm.

0. Set $k = 0$. Select $x_0 \in C$ satisfying A1, A2.

1. Compute a $x_{k+1} \in C'(x_k)$ such that

$$d(x_{k+1},C'(x_k)) \geq \max_{(d(x,C'(x_k)),x)} \{ d(x,C'(x_k)) : f^0(x_k) - f^0(x) \geq d, \; - f^i(x) \geq d, \; i = 1,\ldots,m \} - \varepsilon_k$$

where the ε_k form a sequence of strictly positive values tending to zero as $i \to \infty$, and such that $d(x_{k+1},C'(x_k)) - \varepsilon_k > 0$.

2. If $d(x_{k+1},C'(x_k)) = 0$, $\varepsilon = 0$, set $x^* = x_{k+1}$. Stop.
Otherwise set $k = k+1$ and return to Step 1.

The convergence of this algorithm can be proved in much the same way as convergence was proved for the general algorithm in Theorem 2.1. We need to reconstruct the reasoning after (2.4) as follows:

We have from (2.3),

$$\begin{aligned} 0 < d(\hat{x},C'(\hat{x})) &\leq \rho_0 . d(\hat{x},E_k) \quad \text{with } \rho_0 > 0 \text{ since } C'(\hat{x}) \subset E_k \\ &\leq \rho_0 . (d(x_k,E_k) + \varepsilon_{k-1}) \quad \text{since } x_k \text{ maximizes } d \text{ to within } \varepsilon_{k-1} \text{ on } E_k. \end{aligned}$$

Therefore,

$$0 < d(\hat{x},C'(\hat{x})) \leq \rho_0 . (d(x_k,E_k) + \varepsilon_{k-1}). \qquad (3.9)$$

However,

$$d(x_k,E_k) \to 0, \; \varepsilon_k \to 0$$

contradicts (3.9). The rest of the proof holds for $\varepsilon_k \to 0$ as $k \to \infty$.

4. THE METHOD OF CENTRES BY UPPER-BOUNDING FUNCTIONS: Huard [14].

In this section a very general procedure is developed using an upper bound of the f-distance. It will be shown in the next section how this can be particularized to the Linearized Method of Centres described by Huard in [13].

Assume the functions $f^i(x)\colon R^n \to R$, $i = 0,\ldots,m$ are continuous and convex, and that $f^o(x)$ attains its minimum on the set $C \cap B$ at x^*.

Suppose that $d\colon R^n \times \Sigma \to R$ is a continuous regular f-distance on $R^n \times \Sigma$ satisfying

$$(v)\quad d(x,E) < 0 \quad \forall\, x \notin E,\ \forall\, E \in \Sigma. \tag{4.1}$$

For simplification we shall consider C to be a subset of R^n instead of Σ i.e. we shall consider a function $d\colon R^n \times R^n \to R$ instead of the function $d\colon R^n \times R^n \to R$. B is some compact subset of R^n.

Definition 4.1: Define $d'\colon R^n \times R^n \times R^n \to R$ to be a continuous function satisfying:

(i) $d'(x,y,C'(x_o)) \geq d(x,C'(x_o)) \quad \forall\, y \in B,\ \forall\, x \in C'(x_o) \cap B$

$\forall\, x_o \in C$ such that $C'(x_o) \neq \phi$,

(ii) $d'(x,x,C'(x)) = 0 \quad \forall\, x \in C$,

(iii) $\forall\, a \in C,\ \forall\, b \in B$ fixed:

$$d(x,C'(a)) \leq 0,\ \forall\, x \in [a,b]$$

$$\Rightarrow\ d'(x,a,C'(a)) \leq 0\ \ \forall\, x \in [a,b].$$

Remark:

$$d(x,y,C'(x_o)) = 0\ \ \forall\, y \in B,\ \forall\, x \notin C'(x_o) \cap B$$

$$\Rightarrow\ d(x,C'(x_o)) = 0\ \ \forall\, x \notin C'(x_o) \cap B, \text{ by (i)},$$

$$\Rightarrow\ (C'(x_o))^o = \phi \Rightarrow (C'(x_o))^o \cap B = \phi.$$

Algorithm.

0. Set $k = 0$. Choose a point $x_0 \in C \cap B$. Choose a constant $\alpha \geq 1$.

1. Determine $y_k \in B$ such that

$$\alpha . d'(y_k, x_k, C'(x_k)) \geq d'(y, x_k, C'(x_k)), \quad \forall\, y \in B.$$

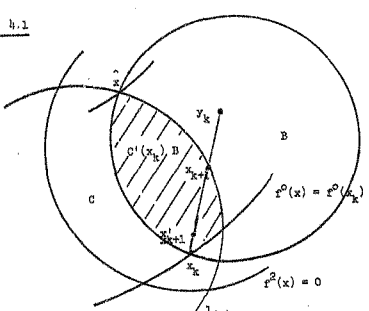
(For instance maximizing $d'(y, x_k, C'(x_k))$ on $C \cap B$ or on B).
If $d'(y_k, x_k, C'(x_k)) = 0$, set $x^* = x_k$. Stop.
Otherwise go to Step 2.

2. Determine $x'_{k+1} \in [x_k, y_k]$ such that

$$d(x'_{k+1}, C'(x_k)) \geq d(x, C'(x_k)) \quad \forall\, x \in [x_k, y_k].$$

3. Determine $x_{k+1} \in C'(x'_{k+1}) \cap B$.
(For instance minimizing $f^0(x)$ on $[x'_{k+1}, y_k] \cap C'(x'_{k+1})$.
Set $k = k+1$ and return to Step 1.

Figure 4.1



Remark:

(i) B convex $\Rightarrow [x_k, y_k] \subset B$

(ii) $d(x'_{k+1}, C'(x_k)) \geq d(x_k, C'(x_k)) = 0$

$\Rightarrow$ by (4.1) that $x'_{k+1} \in C'(x_k) \cap B \subset C \cap B$

i.e. $f^0(x'_{k+1}) \geq f^0(x^*)$

and so $C'(x'_{k+1}) \cap B \neq \phi$.

This algorithm falls within the ambit of the General Method of Centres described in Section 2 for which convergence has been proved. However, the algorithm is conceptual because of Step 2, which can be simplified, computationally, using ε-centres. This gives:

Algorithm.

0. Set $k = 0$. Select a point $x_0 \in C \cap B$. Choose a constant $\alpha \geq 1$.

1. Determine $y_k \in B$ such that

$$\alpha . d'(y_k, x_k, C'(x_k)) \geq d'(y, x_k, C'(x_k)) \quad \forall\, y \in B.$$

If $d'(y_k, x_k, C'(x_k)) = 0$, set $x^* = x_k$. Stop.

Otherwise go to Step 2.

2. Determine $x'_{k+1} \in [x_k, y_k]$ such that

$$d'(x'_{k+1}, C'(x_k)) \geq \max \{ d(x, C'(x_k)) : f^0(x_k) - f^0(x) \geq d, \; (d(x, C'(x_k)), x) \qquad - f^i(x) \geq d,\ i = 1, .., m \} - \varepsilon_k$$

with $\varepsilon_k \to 0$ as $k \to \infty$.

3. Determine $x_{k+1} \in C'(x'_{k+1}) \cap B$.

Set $k = k+1$ and return to Step 1.

Theorem 4.1: Let the functions $f^i(x)$, $i = 0,\ldots,m$ be convex and continuously differentiable. Suppose our assumptions are satisfied. Then the sequence $\{x_k\}$, if it is infinite, generated by the previous algorithm has accumulation points, all of which are optimal. If the sequence is finite, then the last element is optimal.

Proof: The proof of convergence follows from Theorem 2.1 and remarks concerning convergence made after the Algorithm using ε-centres at the end of the last section.

We need to show that in the infinite case, the ε_k defined by

$$\varepsilon_k = \max \{ d(x,C'(x_k)) : f^0(x_k) - f^0(x) \geq d, \\ - f^i(x) \geq d,\ i = 1,\ldots,m \} - d(x'_{k+1},C'(x_k))$$

tends to zero when $k \to \infty$.

1. Firstly, we prove that

$$d(x'_{k+1},C'(x_k)) \to 0.$$

$$x'_{k+1} \in C'(x_k) \Rightarrow f^0(x'_{k+1}) \leq f^0(x_k)$$

$$x_{k+1} \in C'(x'_{k+1}) \Rightarrow f^0(x_{k+1}) \leq f^0(x'_{k+1})$$

$$x_{k+1} \in C \cap B \Rightarrow f^0(x^*) \leq f^0(x_{k+1})$$

$$\Rightarrow \quad f^0(x_k) \geq f^0(x'_{k+1}) \geq f^0(x_{k+1}) \geq f^0(x^*)$$

$$\Rightarrow \quad f^0(x_{k+1}) \to f^0(\hat{x}) \geq f^0(x^*)$$

$$f^0(x'_{k+1}) \to f^0(\hat{x}) \qquad \text{as } k \to \infty,\ k \in N,$$

$$\text{i.e. } f^0(x_{k+1}) - f^0(\hat{x}) \to 0$$

$$f^0(x'_{k+1}) - f^0(\hat{x}) \to 0 \qquad \text{as } k \to \infty,\ k \in N.$$

Now, $\forall\ k \in N$

$$C'(x_k) \supset C'(x'_{k+1}) \supset C'(x_{k+1}) \supset C'(\hat{x}),$$

$$x_{k+1} \in Fr(\ C'(x_{k+1})\ )$$

and d is a regular f-distance. This implies that

$$d(x_{k+1}, C'(x_k)) \to 0 \qquad \text{as } k \to \infty,\ k \in N.$$

Similarly

$$d(x'_{k+1}, C'(x_k)) \to 0 \qquad \text{as } k \to \infty,\ k \in N. \tag{4.3}$$

2. Secondly, we show that there is a subsequence for which

$$\lim\ d'(y_k, x_k, C'(x_k)) \le 0, \text{ when } k \to \infty,\ k \in S \subset N.$$

The point (x_k, y_k) belongs to the compact set $A \cap B \times B$. Hence, there exists $S \subset N$ such that:

$$x_k \to \hat{x} \in A \cap B$$

$$y_k \to \hat{y} \in B \qquad \text{when } k \to \infty,\ k \in S \subset N.$$

On the other hand, $\forall\ \theta \in [0,1]$ fixed, $\forall\ k \in N$, by definition of x'_{k+1}

$$d(x_k + \theta(y_k - x_k), C'(x_k)) \le d(x'_{k+1}, C'(x_k))$$

which gives, taking the limit $k \to \infty$, $k \in S$, with fixed θ, d being continuous, that

$$d(\hat{x} + \theta(\hat{y} - \hat{x}), C'(\hat{x})) \le 0, \quad \forall\ \theta \in [0,1].$$

This gives, by Definition 4.1 (iii),

$$d'(\hat{x} + \theta(\hat{y} - \hat{x}), \hat{x}, C'(\hat{x})) \le 0, \quad \forall\ \theta \in [0,1].$$

Setting $\theta = 1$, we have, since d' is continuous,

$$\lim\ d'(y_k, x_k, C'(x_k)) \leq 0, \text{ when } k \to \infty,\ k \in S. \qquad (4.4)$$

3. Finally, we show that $\varepsilon_k \to 0$.

Let c_k be a point of $C'(x_k)$ which maximizes $d(x, C'(x_k))$ on $C'(x_k) \cap B$. Such a point exists because d is continuous, B is compact, $C'(x_k) \cap B \neq \phi$, and by (4.1) $d(x, C'(x_k)) < 0 \ \forall\ x \notin C'(x_k)$.

We have from (4.3), that $d(x'_{k+1}, C'(x_k)) \to 0$ as $k \to \infty$, $k \in N$. Also, since $c_k \in C'(x_k) \cap B$,

$$\begin{aligned} 0 &\leq d(c_k, C'(x_k)) \\ &\leq d'(c_k, x_k, C'(x_k)) \\ &\leq \alpha . d'(y_k, x_k, C'(x_k)) \quad \text{by definition of } y_k. \end{aligned} \qquad (4.5)$$

Taking $k \to \infty$, $k \in S$, because $\alpha \geq 1$ we have from (4.4), (4.5)

$$d(c_k, C'(x_k)) \to 0 \quad \text{as} \quad k \to \infty.$$

Thus

$$\varepsilon_k = d(c_k, C'(x_k)) - d(x'_{k+1}, C'(x_k)) \geq 0$$

tends to zero as $k \to \infty$, $k \in S$.

The algorithms outlined in this section are fairly general, since by particularizing them still further Huard [14] manages to get at the well-known methods of Zoutendijk [23], Frank and Wolf [7], Rosen [21], and, in particular, the Linearized Method of Centres, Huard [13], which is covered in the next section.

5. THE LINEARIZED METHOD OF CENTRES: Huard [13].

Let A1, A2 be satisfied, B be a convex set which is large enough to contain C in its interior. Assume that the functions $f^i(x)$: $R^n \to R$, $i = 0,\ldots,m$ are continuously differentiable and convex. Let x^* be a point minimizing $f^0(x)$ on C, $x_0 \in C$. Define

$$d'(x,y,C'(x_0)) \triangleq \min \{ f^0(x_0) - f^0(x,y); -f^i(x,y),\ i = 1,..m \} \tag{5.1}$$

where

$$f^0(x,y) \triangleq f^0(y) + \nabla f^0(y)(x-y) \tag{5.2}$$

$$f^i(x,y) \triangleq f^i(y) + \nabla f^i(y)(x-y) \qquad \forall\ i = 1,\ldots,m. \tag{5.3}$$

We now show that the conditions of Definition 4.1 are satisfied where, because the functions are convex,

$$f^i(x,y) \leq f^i(x) \qquad \forall\ i = 0,\ldots,m,$$

(i) $d'(x,y,C'(x_0)) = \min \{ f^0(x_0) - f^0(x,y); -f^i(x,y),\ i = 1,..m \}$

$$\geq \min \{ f^0(x_0) - f^0(x), -f^i(x),\ i = 1,\ldots,m \}$$

$$= d(x,C'(x_0)).$$

(ii) $d'(x,x,C'(x)) = \min \{ 0; -f^i(x),\ i = 1,\ldots,m \}$

$$= 0 \qquad \forall\ x \in C.$$

(iii) Consider $a \in C$, $b \in B$ such that

$$d(x,C'(a)) \leq 0 \qquad \forall\ x \in [a,b]$$

It is easily verified that

$$\nabla f^0(a).(b-a) \geq 0$$

and/or there exists at least one $i \in \{1,\ldots,m\}$ such that

$$\nabla f^i(a).(b-a) \geq 0.$$

Hence, from (5.1) - (5.3), $d'(x,a,C'(a))$ is decreasing at a point x on $[a,b]$. Since $d'(a,a,C'(a)) = 0$ we have that

$$d'(x,a,C'(a)) \leq 0 \quad \forall\, x \in [a,b].$$

Also, it is easy to show that $d'(x,y,C'(x_0))$ is a continuous function, because the functions $\nabla f^i(x)$, $i = 0,\ldots,m$ are continuous by assumption.

Set $x = x_k$. Therefore, from the above, we can replace Step 1 in the Algorithm for the Method of Centres by Upper Bounding functions with:

Choose $\alpha = 1$ and solve

$$\max_{h \in S} \{ \min \{ < -\nabla f^0(x), h(x) >, -f^i(x) - < \nabla f^i(x), h(x) >, \quad i = 1,\ldots,m \} \} \quad (5.4)$$

where S is any subset of R^n containing the origin in its interior.

Let $h^0(x): R^n \to R$ be defined as the solution to (5.4), where

$$S \triangleq \{ h(x) \in R^n : |h^i| \leq 1,\ i = 1,\ldots,n \}. \quad (5.5)$$

Theorem 5.1: Suppose that $h(x) \in S$ is such that

$$h^o(x) = \min \{ -\langle \nabla f^o(x), h(x) \rangle, -f^i(x) - \langle \nabla f^i(x), h(x) \rangle, \ i=1,..m \} \quad (5.6)$$

Then (i) $(h^o(x), h(x))$ is optimal for the linear programming problem

$$\text{maximize } h^o(x) \quad (5.7)$$

subject to

$$-h^o(x) - \langle \nabla f^o(x), h(x) \rangle \geq 0 \quad (5.8)$$

$$-h^o(x) - f^i(x) - \langle \nabla f^i(x), h(x) \rangle \geq 0, \ i = 1,...m \quad (5.9)$$

$$|h^i(x)| \leq 1 \qquad i = 1,...,m \quad (5.10)$$

(ii) If $(h^o(x), h(x))$ is optimal for (5.7) - (5.10), it satisfies (5.6).

Proof: (i) h satisfies (5.10), by definition of S. Since h^o satisfies (5.6) we have

$$0 = \min \{ -\langle \nabla f^o(x), h(x) \rangle - h^o(x), -f^i(x) - \langle \nabla f^i(x), h(x) \rangle - h^o(x), \ i = 1,...,m \}$$

which implies that $(h^o(x), h(x))$ satisfies constraints (5.8) and (5.9).

Assume that $h^o(x)$ does not maximize (5.7) such that (5.8) - (5.9) are satisfied. This implies that there exists $\delta > 0$ s.t.

$$-\langle \nabla f^o(x), h \rangle \geq h^o(x) + \delta$$

$$-f^i(x) - \langle \nabla f^i(x), h \rangle \geq h^o(x) + \delta, \quad \forall \ i = 1,...,m.$$

Hence equality doesn't hold in (5.6). Contradiction.

(ii) Suppose that $(h^0(x),h(x))$ is optimal for (5.7) - (5.10). (5.10) implies that $h(x) \in S$. (5.8),(5.9) imply that

$$- \langle \nabla f^0(x), h \rangle \geq h^0(x)$$

$$- f^i(x) - \langle \nabla f^i(x), h \rangle \geq h^0(x), \quad i = 1,\ldots,m.$$

Hence (5.6) holds as $h^0(x)$ maximizes the RHS, so that equality holds.

Let $\bar{h}^0(x)$ be the maximum value of $h^0(x)$ such that constraints (5.8) - (5.10) are satisfied for x fixed.

Theorem 5.2: $\bar{h}^0(x^*) = 0$ iff x^* is optimal.

Proof: From the linear programming problem we have

$$\bar{h}^0(x) = \max_{(h^0,h)} \{ h^0(x) : \langle \nabla f^0(x), h(x) \rangle \leq - h^0(x), \; f^i(x) + \langle \nabla f^i(x), h(x) \rangle \leq - h^0(x), \; i = 1,\ldots,m \},$$

This is equivalent to finding

$$- \bar{h}^0(x) = \min_{(h^0,h)} \{ h^0(x) : \langle \nabla f^0(x), h(x) \rangle \leq h^0(x), \; f^i(x) + \langle \nabla f^i(x), h(x) \rangle \leq h^0(x), \; i = 1,\ldots,m \}.$$

Then, because of our assumptions at the beginning of this chapter it is straightforward to verify that the condition of the Strong Duality Theorem (Appendix A) are satisfied by the above problem. Therefore, from (A.4),

$$- \bar{h}^0(x) = \max_{\mu \geq 0} \{ \inf_{(h^0,h)} \{ h^0(x) + \mu^0 (\langle \nabla f^0(x), h(x) \rangle - h^0(x)) + \sum_{i=1}^{m} \mu^i (f^i(x) + \langle \nabla f^i(x), h(x) \rangle - h^0(x)) \} \}$$

$$= \max \{ \inf_{(h^0,h)} \{ \Big(1 - \sum_{i=0}^{m} \mu^i \Big) h^0(x) + \sum_{i=0}^{m} \mu^i f^i(x) + \sum_{i=0}^{m} \mu^i \langle \nabla f^i(x), h(x) \rangle \}.$$

Now, by (3.A), problem (2.A) has at least one solution, and by (11.A)

$$-\bar{h}^0(x) = \max_{\mu \geq 0} \{ \Big(1 - \sum_{i=0}^{m} \mu^i \Big) h^0(x) + \sum_{i=1}^{m} \mu^i f^i(x) + \sum_{i=0}^{m} \mu^i \langle \nabla f^i(x), h(x) \rangle \mid \sum_{i=0}^{m} \mu^i \geq 1, \sum_{i=0}^{m} \mu^i \nabla f^i(x) = 0 \}$$

$$= \max_{\mu \geq 0} \{ \sum_{i=1}^{m} \mu^i f^i(x) \mid \sum_{i=0}^{m} \mu^i = 1, \sum_{i=0}^{m} \mu^i \nabla f^i(x) = 0 \}. \qquad (5.11)$$

Now it can be seen from (5.11) that

$$\bar{h}^0(x) \geq 0 \quad \forall\, x \in C. \qquad (5.12)$$

Suppose now that $x^* \in C$ and $\bar{h}^0(x^*) = 0$. Then from (5.11) there exists $\bar{\mu} \geq 0$ such that

$$\sum_{i=1}^{m} \bar{\mu}^i f^i(x^*) = 0, \sum_{i=0}^{m} \bar{\mu}^i = 1, \sum_{i=0}^{m} \bar{\mu}^i \nabla f^i(x^*) = 0. \qquad (5.13)$$

From the convexity of the functions f^i, $i = 0,\ldots,m$

$$f^i(x) \geq f^i(x^*) + \langle \nabla f^i(x^*), x - x^* \rangle \quad \forall\, x \in R^n;\ i=0,..,m,$$

which, because of (5.13), becomes

$$\sum_{i=1}^{m} \bar{\mu}^i f^i(x) \geq \bar{\mu}^0 (f^0(x^*) - f(x)) \quad \forall\, x \in R^n. \qquad (5.14)$$

If $\bar{\mu}^0 = 0$, then A.2(i) is contradicted. Therefore $\bar{\mu}^0 > 0$. Now from the fact that $\bar{\mu} \geq 0$ we have from (5.14)

$$f^0(x^*) - f^0(x) \leq 0 \quad \forall\, x \text{ such that } f^i(x) \leq 0,\ i=0,..,m,$$

i.e. $f^0(x^*) \leq f^0(x) \quad \forall\, x \in C.$

So we have that x^* is optimal for (P.1).

Let x^* be optimal for problem (P.1). Then, by the well known Kuhn-Tucker conditions there exists an optimal multiplier $\lambda \in R^{m+1}$ such that

$$\sum_{i=1}^{m} \lambda^i f^i(x^*) = 0, \quad \sum_{i=0}^{m} \lambda^i \nabla f^i(x^*) = 0, \quad \sum_{i=0}^{m} \lambda^i = 1, \ \lambda \geq 0 \qquad (5.15)$$

and (5.11) and (5.15) imply that $\bar{h}^0(x^*) = 0$.

Remark: From (5.12) we have that $\bar{h}^0(x) \geq 0 \quad \forall x \in C$.

We replace Steps 2,3 in the Algorithm for the Method of Centres by Upper Bounding Functions with the Step Size determination problem

$$\max_{(d_h,\mu)} \{ d_h \mid f^0(x_k) - f^0(x_k + \mu h(x_k)) \geq d_h, \ -f^i(x_k + \mu h(x_k)) \geq d_h, \ i = 1,\dots,m \} \qquad (5.16)$$

and x'_{k+1} and x_{k+1} co-incide.

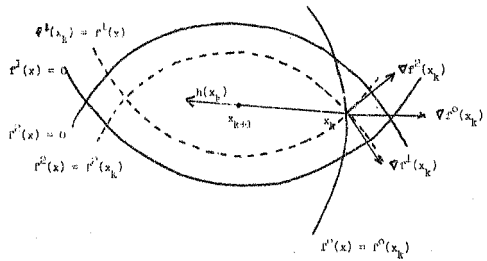


Figure 5.1

The purpose of the subproblem (5,7) - (5,10) at iteration k+1 is to find a vector which points into the region $C'(x_k)$, and (5.16) gives us a point x_{k+1} in $C'(x_k)$. See Figure 5.1. It can be seen that the Linearized Method of Centres is closely related to Zoutendijk's Method of Feasible Directions.

The previous discussion gives us:

Algorithm.

0. Set $k = 0$. Select $x_0 \in C$.

1. Set $x_k = x$. Solve (5.7) - (5.10) to obtain $(h^0(x_k), h(x))$.

2. If $h^0(x_k) = 0$, set $x^* = x_k$. Stop.
   Otherwise go to 3.

3. Compute solution (d_k, μ) to (5.16).

4. Set $x_{k+1} = x_k + \mu h(x_k)$.
   Set $k = k+1$ and return to Step 1.

That this algorithm converges for the infinite case and that its points of accumulation are optimal can be seen from Theorem 5.2 and Theorem 2.1.

Piyavskii and Polak [10] note that they have been able to show that the above algorithm converges at least as fast as $1 / \sqrt{k}$ with the above assumptions strengthened slightly to make $f^0(x)$ strictly convex in the neighbourhood of the optimum. They also show that the *algorithm did not converge linearly* under these assumptions, for a specific case. They then proposed a modified Method of Centres which exhibits improved convergence. This is outlined in the next section.

6. A MODIFIED METHOD OF CENTRES: Pironneau and Polak [18].

Consider the problem (P.1) where the functions $f^i(x):R^n \to R$, $i = 0,\ldots,m$ are continuously differentiable and convex, satisfying A1, A2. Define $k^0(x)$ by

$$k^0(x) \triangleq \max_{(h^0,h)} \{ h^0(x) - \tfrac{1}{2}||h||^2 : < \nabla f^0(x), h(x) > \leq - h^0(x), f^i(x) + < \nabla f^i(x), h(x) > \leq - h^0(x), i = 1,..,m \} \quad (6.1)$$

Algorithm.

0. Set $k = 0$. Select $x_0 \in C$.

1. Compute a solution $(h^0(x_k),h(x))$ to

$$\max_{(h^0,h)} \{ h^0 - \tfrac{1}{2}||h||^2 : - < \nabla f^0(x_k), h > \geq h^0, - f^i(x_k) - < \nabla f^i(x_k), h > \geq h^0, i = 1,\ldots,m \}. \quad (6.2)$$

2. If $k^0(x_k) = 0$, set $x^* = x_k$. Stop.
Otherwise go to 3.

3. Compute a solution $(d(x_{k+1},C'(x_k)), \mu(x_k))$ to

$$\max_{(d(x,C'(x_k)),\mu)} \{ d(x,C'(x_k)) : f^0(x_k) - f^0(x_k + \mu h(x_k)) \geq d, - f^i(x_k + \mu h(x_k)) \geq d, i = 1,\ldots,m \}.$$

4. Set $x_{k+1} = x_k + h(x_k)$.
Set $k = k+1$ and return to Step 1.

Note that (5.7) - (5.10) is a linear programming problem with $m+n+1$ constraints, while (6.2) is a quadratic problem with $m+1$ constraints.

Theorem 6.1: $k^0(x^*) = 0$ iff x^* is optimal.

The above result can be proved by amending the proof of Theorem 5.2 slightly. Otherwise the reader should consult [18]. Also

$$k^0(x) \geq 0 \qquad \forall\, x \in C.$$

(II) Assume that there exists $\varepsilon' > 0$ and $m^0 \in (0,1)$ such that

$$\langle y, \frac{\partial^2 f^0(x)}{\partial x^2} y \rangle \geq m^0 ||y||^2 \qquad \forall\, y \in R^n,\ \forall\, x \in B(x^*, \varepsilon') \cap C,$$

where x^* is the unique solution to (P.1) and C is as before.

The above assumption guarantees strict convexity in the neighbourhood of x^*, a solution to (P.1). Hence x^* is unique. Pironneau and Polak [18] show that this algorithm has linear convergence. The result is given here for completeness. The proof can be found in [18].

Theorem 6.2: Let $\{x_k\}$ be a sequence generated by the above algorithm in solving problem (P.1). Suppose that the assumptions made at the beginning of the chapter and (II) are satisfied. Then $x_k \to x^*$ linearly as $k \to \infty$, $f^0(x_k) \to f^0(x^*)$ linearly as $k \to \infty$, in accordance with the following bounds: Given any $\gamma \in (0,1)$, there exists $k_0(\gamma)$, $C_0(\gamma) > 0$ such that

$$||x_k - x^*|| \leq \left\{ 1 - \bar{\lambda}^{0^2} \frac{m^0}{L} (1-\gamma)^2 \right\}^{\frac{1}{2}} C_0(\gamma) \qquad \forall\, k \geq k_0(\gamma)$$

$$f^0(x_{k+1}) - f^0(x^*) \leq \left\{ 1 - \bar{\lambda}^{0^2} \frac{m^0}{L} (1-\gamma)^2 \right\} \left\{ f^0(x_k) - f^0(x^*) \right\} \quad \forall\, k \geq k_0(\gamma),$$

where m^0 is defined in (II),

$$\bar{\lambda}^0 \triangleq \min \{ <\lambda,e> : \sum_{i=1}^{m} \lambda^i f^i(x^*) = 0, \quad \sum_{i=0}^{m} \lambda^i \nabla f^i(x^*) = 0,$$

$$\sum_{i=0}^{m} \lambda^i = 1, \quad \lambda \geq 0, \quad \lambda \in R^{m+1} \},$$

$$e \triangleq (1,0,0,\ldots,0) \in R^{m+1},$$

$$L \triangleq \max \{ 1, M^i, i = 0,\ldots,m \}, \text{ where}$$

$$M^i = \max \{ \| \frac{\partial^2}{\partial x^2} f^i(x) \| : x \in B(x^*,\varepsilon') \}.$$

The Modified Algorithm is not implementable computationally because of the exact minimization required in Step 3. Pironneau and Polak [18] consider the application of a Golden Section Search to the function $d_h : R \to R$ defined by

$$d_h = \min_{\mu} \{ f^0(x_k) - f^0(x_k + \mu h(x_k)), - f^i(x_k + \mu h(x_k)), i = 1,.m \}$$

to find two points $\mu(x_k)$, $\mu'(x_k)$ with $\mu'(x_k) > \mu(x_k) > 0$ such that $[\mu(x_k),\mu'(x_k)]$ contains the maximizer of d_h, and such that

$$d_h(\mu(x_k)) \geq \beta(\mu'(x_k) - \mu(x_k))<\nabla f^0(x_k),h(x_k)>,$$

where $\beta > 0$. d_h is just the distance function defined in (1.7) along the vector $h(x_k)$. It is proved in [18] that the algorithm still converges to the unique solution to the problem, and that it does so linearly.

7. A METHOD OF CENTRES ALGORITHM WITH ARBITRARY PARAMETERS:

Hoshino [10].

Fiacco and McCormick [5],[6] proposed a distance function in $C'(x_k)$ which is defined by

$$Q_k(x) = \left(f^0(x_k) - f^0(x) \right)^{-1} - \sum_{i=1}^{m} \left(f^i(x) \right)^{-1} \qquad (7.1)$$

and they used this Q-function for their Sequential Unconstrained Minimization Technique (SUMT). It can be seen that if the functions $f^i(x)$, $i = 0,\ldots,m$ are convex then $Q_k(x)$ is convex.

If we were to consider the inverse of the Q-function then we would have a concave function in $C'(x_k)$ which would be identically zero for all points $x \in Fr(C'(x_k))$. Also, it would take on positive values for all $x \in (C'(x_k))^0$. We therefore consider the new distance function defined by

$$\alpha_k (d_k(x))^{-1} = \alpha_k \left(f^0(x_k) - f^0(x) \right)^{-1} - \sum_{i=1}^{m} \left(f^i(x) \right)^{-1} \quad \forall x \in (C'(x_k))^0$$

$$= 0 \qquad \forall x \in Fr(C'(x_k)) \qquad (7.2)$$

where α_k is a positive arbitrary parameter, used to accelerate the rate of convergence. Outside $C'(x_k)$ it is convenient to let $d_k(x)$ be the function

$$d_k(x) = \min \{ f^0(x_k) - f^0(x), -\alpha_k f^i(x), i = 1,\ldots,m \}. \qquad (7.3)$$

Remarks.

(i) Let $f^i(x)$, $i = 0,\ldots,m$ be convex functions. Then (7.2) implies that $d_k(x)$ is a convex function in $C'(x_k)$. Therefore any maximum of $d_k(x)$ in $C'(x_k)$ is the global maximum. When x is in $C'(x_k)$, then $d_k(x)$ given by (7.3) is

not less than that given by (7.2), and the function which is defined to be the smallest of concave functions, is concave (i.e. no points of inflection). Thus the extension to $d_k(x)$ given by (7.3) preserves the concavity. This concavity property is helpful for the unconstrained maximization, for it permits points outside $C'(x_k)$ to be accessed in the course of a linear search, although the *centres are confined to* $C'(x_k)$.

(ii) $d_k(x) = f^0(x_k) - f^0(x)$ when $f^0(x_k) - f^0(x) \leq -\alpha_k f^i(x)$
$i = 1,\ldots,m$

$d_k(x) = -\alpha_k f^i(x)$ when $-\alpha_k f^i(x) \leq f^0(x_k) - f^0(x)$
and $-\alpha_k f^i(x) \leq -\alpha_k f^j(x)$
$j = 1,..,i-1,i+1,..,m.$

Therefore the magnitude of the distance function tends to vary less than (1.8) for which there are dangers of overflow and underflow. This is desirable for numerical treatment.

(iii) The derivitive of $\alpha_k(x)$ is continuous at the boundary points of $C'(x_k)$ where only one of the functions $f^0(x_k) - f^0(x)$, $f^i(x)$, $i = 1,\ldots,m$ is zero.

Consider the Barrier (interior penalty) function of the form

$$B_r(x) = f^0(x) - r^\lambda \sum_{i=1}^{m} \left(f^i(x) \right)^{-1} \qquad (7.4)$$

described by Fiacco and McCormick [6], where r denotes a positive controlling parameter. Let the $f^i(x)$, $i = 0,\ldots,m$ be convex functions.

$B_r(x)$ is a convex function with $B_r(x) = \infty$, for any $x \in Fr(C)$. Then there exists a point $x \in C$ minimizing (7.4) over C for any $r > 0$. This is due to the second term in the expression which presents itself as a barrier in order to prevent violation of the constraints.

In addition to convexity assume that A1, A2 are satisfied, and that the functions $f^i(x)$, $i = 0,\ldots,m$ have continuous first order partial derivitives. Then the gradient of $d_k(x)$ vanishes at the centre x_{k+1}, whence

$$\nabla f^0(x_{k+1}) + \alpha_k^{-1}(f^0(x_k) - f^0(x_{k+1}))\sum_{i=1}^{m} \nabla f^i(x_{k+1})/(f^i(x_{k+1}))^2 = 0 \tag{7.5}$$

Obviously the point x_{k+1} solves the equation

$$\nabla f^0(x_{k+1}) - r^2 \sum_{i=1}^{m} \nabla f^i(x_{k+1})/(f^i(x_{k+1}))^2 = 0 \tag{7.6}$$

for r given by

$$r_k^2 = (f^0(x_k) - f^0(x_{k+1}))^2/\alpha_k. \tag{7.7}$$

Then it can be shown,[15], that the sequence $\{r_k\}$ generated in this manner is a monotonic, nonincreasing, null sequence.

Hoshino considers α_k given by

$$\alpha_k = r_k(f^0(x_{k+1}) - f^0(x_k))\sum_{i=1}^{m} (f^i(x_{k+1}))^{-1} \tag{7.8}$$

where r_k forms a null sequence. Then

$$r_k^2 = (f^0(x_{k+1}) - f^0(x_k))/(r_{k_1 \neq 1}\sum^{m} (f^i(x_{k+1}))^{-1}). \tag{7.9}$$

Algorithm.

0. Set $k = 0$. Select $x_0 \in C$.

1. Compute a solution $x_{k+1} \in C'(x_k)$ such that

$$d_k(x_{k+1}) \geq d_k(x) \quad \forall\, x$$

where $d_k(x)$ is given by (7.2) and (7.3), and α_k by (7.8). (From remark (i), the solution which maximizes $d_k(x)$ will lie in $C'(x_k)$).

2. If $d_k(x_{k+1}) = 0$, set $x^* = x_{k+1}$. Stop.
Otherwise set $k = k+1$ and return to Step 1.

The convergence of this algorithm to an optimal solution follows from the proof of Theorem 2.1, or, alternatively, can be deduced from the discussion before the algorithm and Theorem 8 in Fiacco and McCormick [6].

General results about the convergence rate of the Method of Centres using the distance function (7.2) are not known. However, under the additional assumptions:

(i) $f^i(x)$, $i = 0,\ldots,m$ are linear functions,

(ii) only one vector x solves the problem,

(iii) the effect of inactive constraints on $d_k(x)$ is ignored. In fact this effect tends to zero as the solution is approached.

Then Hoshino [9] asserts that:

(i) when α_k is given by (7.8), then the geometric distance

from x_k to x_{k+1} is greater than or equal to r_k times the minimum geometrical distance from x_{k+1} to a constraint.

(ii) All the centres x_k lie on a fixed straight line. This can be used to accelerate convergence.

(iii) If r_k is kept constant for all the maximizations, and α_k is given by (7.8), then the geometrical distance from the centre x_{k+1} to the solution x^* decreases by the factor $1/(1+r_k)$ on each iteration.

Implementation of the algorithm.

It can be seen that a major drawback to implementing the above algorithm is that at each iteration we should like to set

$$\alpha_k = r_g\,(f^0(x_{k+1}) - f^0(x_k) \,) \sum_{i=1}^{m} 1/f^i(x_{k+i}) \qquad (7.10)$$

where r_g is constant. However, this expression contains $f^i(x_{k+1})$, $i = 0,\ldots,m$. But x_{k+1} is as yet unknown and this would result in many trials for different α_k's to determine the value (7.10). Therefore, in place of α_k we use the quantity

$$\hat{\alpha}_k = r_g\,(f^0(x_k) - f^0(x_{k-1}) \,) \sum_{i=1}^{m} 1/f^i(x_k).$$

Then the additional assumptions (i),(ii),(iii) give,[9],

$$\hat{\alpha}_k/\hat{\alpha}_{k-1} = (\alpha_k/\hat{\alpha}_{k-1})^{\frac{1}{2}}.$$

If $\alpha_k > \hat{\alpha}_{k-1}$ then $\alpha_k > \hat{\alpha}_k > \hat{\alpha}_{k-1}$, and if $\alpha_k < \hat{\alpha}_{k-1}$ then $\alpha_k < \hat{\alpha}_k < \hat{\alpha}_{k-1}$. So $\hat{\alpha}_k$ changes slower than α_k. This is an approximate relation between α_k and $\hat{\alpha}_k$.

To accelerate convergence, compute a point y_k where the boundary of the feasible region is cut by the line through x_k and x_{k-1}. See Step 3 of the General Method of Centres algorithm.

For a more detailed description of the above results and some actual programming results, the reader should consult [10]. The test results in [10] show that the Method of Centres algorithm using the distance function defined in (7.2), (7.3) has a marked improvement in the rate of convergence over an algorithm using the distance function defined in (1.8).

To accelerate convergence, compute a point y_k where the boundary of the feasible region is cut by the line through x_k and x_{k-1}. See Step 3 of the General Method of Centres algorithm.

For a more detailed description of the above results and some actual programming results, the reader should consult [10]. The test results in [10] show that the Method of Centres algorithm using the distance function defined in (7.2), (7.3) has a marked improvement in the rate of convergence over an algorithm using the distance function defined in (1.8).

8. EQUALITY CONSTRAINTS AND THE METHOD OF CENTRES: Hoshino [10].

Consider the problem described by

$$\text{minimize} \quad f^0(x) \tag{8.1}$$

(P.4) $$\text{subject to} \quad f^i(x) \le 0, \quad i = 1,\ldots,m_1 \tag{8.2}$$

$$f^i(x) = 0, \quad i = m_1+1,..,m. \tag{8.3}$$

Clearly the interior of the feasible region defined by the constraints is empty, and so (P.4) cannot be handled by our Method of Centres in its present form.

In order to solve (P.4), Pietrykowski [17] minimized the potential function

$$p(x,\mu) = \mu f^0(x) + \sum_{i=1}^{m_1} \text{pos}(\, f^i(x)\,) + \sum_{i=m_1+1}^{m} |f^i(x)| \tag{8.4}$$

where μ is a parameter and where

$$\text{pos}(\tau) = \begin{cases} \tau, & \tau > 0 \\ 0, & \tau \le 0. \end{cases} \tag{8.5}$$

One disadvantage that (9.4) has from the computational point of view is that $p(x,\mu)$ is non-differentiable at some points. To overcome this difficulty, Hoshino [10] considers the additional constraint

$$\begin{aligned} f^i(x) &\ge 0 \qquad i = m+1,\ldots,m+m_1, \\ f^i(x) &= -f^{i-m}(x) \quad i = m+m_1+1,\ldots,2m. \end{aligned} \tag{8.6}$$

Consider solving (P.4) with the additional constraints by minimizing

$$\bar{p}(x,\mu) = \max_{i=1,,,2m} (\mu f^0(x) + f^i(x))$$

$$= \mu f^0(x) + \max_i (f^i(x)), \qquad (8.7)$$

Theorem 8.1: Let the functions $f^i(x)$, $i = 0,\ldots,m$ be continuously differentiable and let the gradient vectors $\nabla f^i(x)$, $i = i,\ldots,m$ be linearly independent. Then there is a real number $\bar{\mu} > 0$ such that for $0 < \mu < \bar{\mu}$, the minimum point $x(\mu)$ of $\bar{p}(x,\mu)$ coincides with the solution x^* of (P.4).

Proof: Firstly, suppose $x(\mu)$ lies in the interior of the infeasible region. Then there is an $\varepsilon > 0$ such that every $x \in N_\varepsilon(x(\mu))$ belongs to the infeasible region. From the definition of $x(\mu)$

$$\mu f^0(x) + \max_i (f^i(x)) \geq \mu f^0(x(\mu)) + \max_i (f^i(x(\mu))). \qquad (8.8)$$

Let x' be such that

$$H(x) = \{ \max_i (f^i(x(\mu))) - \max_i (f^i(x)) \} / ||x(\mu) - x||$$

takes its maximum value at x', $\forall x \in N_\varepsilon(x(\mu))$, which is denoted by M.

Now from (8.6) and the independence of the vectors $\nabla f^i(x)$, $i = 1,\ldots,m$,

$$M \geq \hat{M} > 0 \quad \text{when} \quad \max_i (f^i(x(\mu))) > 0.$$

Therefore, from (8.8) we have that

$$\mu \{ f^0(x) - f^0(x(\mu)) \} \geq \max_i (f^i(x(\mu))) - \max_i (f^i(x))$$

i.e. $\mu \{ f^0(x) - f^0(x(\mu)) \} / ||x(\mu) - x|| \geq \hat{M} > 0.$ (8.9)

As $f^0(x)$ is differentiable, there exists constant $\bar{M}$ such that

$$\{\,|f^0(x) - f^0(x(\mu))|\,\}\,/||x(\mu) - x|| \le \bar{M} \quad \text{for} \quad x \in N_\varepsilon(x(\mu)).$$

If we choose μ such that $\mu < \hat{M}/\bar{M} \ge \bar{\mu}$, then $\mu\bar{M} < \hat{M}$ which contradicts (8.9). Therefore, $x(\mu)$ is not interior to the feasible region for $\mu < \bar{\mu}$.

It follows that when $0 < \mu < \bar{\mu}$ then the vector $x(\mu)$ minimizes $\bar{p}(x,\mu)$ and also satisfies the condition $\max_i\,(\,f^i(x(\mu))\,) = 0$.

i.e. $x(\mu)$ minimizes $f^0(x)$ subject to the constraints.

We have shown above that if we minimize $\bar{p}(x,\mu)$, the solution will also be a solution to (P.4).

Minimizing $\bar{p}(x,\mu)$ given by (8.7) is equivalent to calculating an x and scalar $\hat{0}$ to

$$\text{minimize} \quad \hat{0}$$

$$\text{subject to} \quad \hat{0} \ge \mu f^0(x) + f^i(x), \quad i = 1,\ldots,2m.$$

Define

$$0 = \hat{0} - \mu f^0(x).$$

Then the problem becomes

$$\text{minimize} \quad f^0(x,0) = \mu f^0(x) + 0 \tag{8.10}$$

(P.5)

$$\text{subject to} \quad f^i(x) - 0 \ge 0, \quad i = 1,\ldots,2m. \tag{8.11}$$

to which the Method of Centres can be applied directly.

Corresponding to the distance function given in (7.2), each iteration calculates (x_{k+1},θ_{k+1}) from (x_k,θ_k) by minimizing the distance function $d_k(x,\theta)$ defined by

$$\alpha_k(d_k(x,\theta))^{-1} = \alpha_k(f^o(x_k,\theta_k) - f^o(x,\theta))^{-1} + \sum_{i=1}^{2m} (\theta - F^i(x))^{-1} \quad (8.12)$$

The above results will enable the Method of Centres to be applied to solve a far wider range of problems than has been realized previously.

CONCLUSION.

We have seen that the method of centres is a technique which computes the least value of $f(x)$ subject to (1.2) iteratively. Each iteration seeks a value of x that maximizes a distance function. It is found, in general, that the convergence by the method of centres is rather slow. Nevertheless, it has been found to give accurate solutions. Another attraction is clearly, except in the case of equality constraints discussed in the last chapter, the fact that at each iteration we have a feasible solution to the problem.

Clearly the rate of convergence is dependent on the choice of distance function. This is an area of considerable interest. For the theoretical method of centres, Huard [11], [12] proposed the use of the distance functions defined in equations (1.7) and (1.8). Pironneau and Polak [18] have investigated the rate of convergence of this algorithm. Mifflin [16] and Donel and Huard [2] have considered the advantages of introducing parameters into equations (1.9), (1.10) in an effort to improve convergence. Huard [14] considered the method of centres algorithm using simpler upper bounding functions for the

distance function. He then particularized this algorithm to obtain the linearized method of centres of [13], and also related the method of centres to the well-known methods of Zoutendijk [23], Rosen [21] and Frank and Wolf [7]. Pironneau and Polak [18] showed that for a certain example, the algorithm of 13 did not converge linearly. They proposed a modification to the algorithm which had the effect of improving the convergence.

Hoshino [10] extended the Q-function of Fiacco and McCormick [5], [6] by introducing a parameter α_k. The rate of convergence of the algorithm has been investigated in the linear case by Hoshino [9]. The extension to the nonlinear convex case and more general functions is an area for further research. Different choices for α_k may also be studied from the viewpoint of investigating the rate of convergence, and possibly obtaining a better $\hat{\alpha}_k$ for computational purposes.

Polak [19] has considered searching for 'desirable' points. This *enables certain hypotheses which we made* to be relaxed or discarded. He has shown that the algorithms of [11], [12], [13] converge to a desirable point even when a 'golden section search' is used.

Bui-Trong-Lieu and Huard [1] treated the method of centres in abstract spaces. Polak, Mukai, and Pironneau [20] have considered the method of centres in solving optimal control problems.

Finally, it has been seen that a certain class of problems with *equality* constraints can be transformed into equivalent problems with inequality constraints. This not only broadens the field of applicability of the method of centres considerably, but also means that interior penalty function methods can be applied directly to solve certain problems with equality constraints.

APPENDIX : The Strong Duality Theorem (Polak and Pironneau [18]).

Consider the problem

$$\text{(P.1)} \qquad \begin{array}{l} \text{minimize } f(x) \\ \text{subject to } f^i(x) \le 0, \quad i = 1,\ldots,m \end{array} \tag{1.A}$$

Assume that the $f^i : R^n \to R$ are convex and continuously differentiable for all i. Then the dual to (P.1) is given by

$$\max_{u \ge 0} \{ \inf_x \{ f^0(x) + \sum_{i=1}^m u^i f^i(x) \} \} \tag{2.A}$$

where $u \in R^m$.

Now suppose that (P.1) has at least one solution and that A2 is satisfied. Then

a) (i) problem (2.A) has at least one solution, (3.A)

(ii) $$\max_{u \ge 0} \{ \inf_x \{ f^0(x) + \sum_{i=1}^m u^i f^i(x) \} \} = \min \{ f^0(x) : f^i(x) \le 0, \quad i = 1,\ldots,m \}, \tag{4.A}$$

(iii) for any solution $\bar{u}$ to (2.A) and solution x^* to (P.1),

$$f^0(x^*) = \min_x \{ f^0(x) + \sum_{i=1}^m \bar{u}^i f^i(x) \}, \tag{5.A}$$

b) A vector $\bar{u} \ge 0$ in R^m is a solution of (2.A) iff there exists $x^* \in C$ such that

(i) $$f^0(x^*) + \sum_{i=1}^m \bar{u}^i f^i(x^*) = \min_x \{ f^0(x) + \sum_{i=1}^m \bar{u}^i f^i(x) \}, \tag{6.A}$$

(ii) $$\bar{u}^i f^i(x^*) = 0, \quad i = 1,\ldots,m. \tag{7.A}$$

Note that if $\hat{x}$ and $\tilde{u} \geq 0$ satisfy

$$\nabla f^0(\hat{x}) + \sum_{i=1}^{m} \tilde{u}^i \nabla f^i(\hat{x}) = 0, \tag{8.A}$$

then $\hat{x}$ is a solution to

$$\min_x \{ f^0(x) + \sum_{i=1}^{m} \tilde{u}^i f^i(x) \}. \tag{9.A}$$

Thus, if there exists a solution $\hat{x}$ to

$$\min_x \{ f^0(x) + \sum_{i=1}^{m} \tilde{u}^i f^i(x) \}, \tag{10.A}$$

then for any x^* a solution to problem (P.1),

$$f^0(x^*) = \max_{u \geq 0} \min_x \{ f^0(x) + \sum_{i=1}^{m} u^i f^i(x) \mid \nabla f^0(x) + \sum_{i=1}^{m} u^i \nabla f^i(x) = 0 \}. \tag{11.A}$$

This theorem is a particular case of the strong duality theorem stated in 'Geoffrion, A., Duality in Nonlinear Programming, Working Paper No. 150, Univ. of Calif., Los Angeles, Western Management Science Institute, August 1969'.

REFERENCES.

1. Bui-Trong-Lieu and P. Huard, Le méthode des centres dans un espace topologique, Numer. Math., 8, 56 - 67, 1966.

2. Donel, J., and P.Huard, Programmation non linéaire et linearisation, NATO Conference, Elsinore, Denmark, 1971.

3. Faure, P., and P. Huard, Résolution des programmes mathématiques à fonction nonlinéaire par la méthode du gradient réduit, Rev. Fr. Recherche Opér., 9, 167 - 205, 1965.

4. Fsure, P., and P. Huard, Résultats nouveaux relatifs à la méthode des centres, Quatrième Conférence de Recherche Opér., Cambridge, Mass., 1966.

5. Fiacco, A.V., and G.P. McCormick, The sequential unconstrained minimization technique without parameters, Oper. Res., 15, 820 - 827, 1967.

6. Fiacco, A.V., and G.P. McCormick, Nonlinear Programming, Sequential Unconstrained Minimisation Techniques, Wiley, New York, 1968.

7. Frank, M., and P. Wolf, An algorithm for quadratic programming, Nav. Res. Logist. Quart., 3, 95 - 120, 1956.

8. Frisch, R., The logarithmic potential method for solving linear programming problems, Memor. of Univ. of Inst. of Econ., Oslo, Denmark, 1955.

9. Hoshino, S., Data Processing Center Report No. 10, p 12, Kyoto Univ., Japan, 1972.

10. Hoshino, S., A method of centres algorithm with arbitrary parameter and its application, J. Inst. Maths. Applic., 12, 319 - 328, 1973.

11. Huard, P., Résolution de programmes mathématiques à constraintes non-linéaires par la méthode des centres, Note E.D.F., HR 5690/3/317, 1964.

12. Huard, P., Resolution of mathematical programming with nonlinear constraints by the method of centres, in 'Nonlinear Programming' ed. J. Abadie, 206 - 219, North Holland Pub. Co., Amster., 1967.

13. Huard, P., Programmation mathematique convexe, R.I.R.O. No. 7, 43 - 59, 1968.

14. Huard, P., A method of centers by upper bounding functions with applications, in 'Nonlinear Programming', ed. J.B. Rosen et al, Academic Press, 1970.

15. Loorsma, F.A., Boundary properties of penalty functions for constrained minima, Phillips Res. Reports, 23, 1 - 102, 1968.

16. Mifflin, R., Convergence rates for the method of centres algorithm, O.R. Centre Report ORC 71-10, Coll. of Eng., Univ. of Calif., Berkeley, 1971.

17. Pietrykowski, T., An exact potential method for constrained maxima, SIAM J. Num. Anal., 6, No. 2, 299 - 304, 1969.

18. Pironneau, O., and E. Polak, On the rate of covergence of certain methods of centres, Mem. No. ERL-M296, Elec. Res. Lab., Coll. of Eng., Univ. of Calif., Berkeley, 1971.

19. Polak, E., Computational Methods in Optimization, Academic Press, New York, 1971.

20. Polak, E., Mukai and O. Pironneau, The method of centres and method of feasible directions for the solution of optimal control problems, Proc. 1971 IEEE Conf. on Decision and Control, Miami, Florida, 1971.

21. Rosen, J.B., The gradient projection method for nonlinear programming, part 1, linear constraints, SIAM J., 8, 181 - 217, 1960.

13. Huard, P., Programmation mathematique convexe, R.I.R.O. No. 7, 43 - 59, 1968.

14. Huard, P., A method of centers by upper bounding functions with applications, in 'Nonlinear Programming', ed. J.B. Rosen et al, Academic Press, 1970.

15. Lootsma, F.A., Boundary properties of penalty functions for constrained minima, Phillips Res. Reports, 23, 1 - 102, 1968.

16. Mifflin, R., Convergence rates for the method of centres algorithm, O.R. Centre Report ORC 71-10, Coll. of Eng., Univ. of Calif., Berkeley, 1971.

17. Pietrykowski, T., An exact potential method for constrained maxima, SIAM J. Num. Anal., 6, No. 2, 299 - 304, 1969.

18. Pironneau, O., and E. Polak, On the rate of convergence of certain methods of centres, Mem. No. ERL-M296, Elec. Res. Lab., Coll. of Eng., Univ. of Calif., Berkeley, 1971.

19. Polak, E., Computational Methods in Optimization, Academic Press, New York, 1971.

20. Polak, E., Mukai and O. Pironneau, The method of centres and method of feasible directions for the solution of optimal control problems, Proc. 1971 IEEE Conf. on Decision and Control, Miami, Florida, 1971.

21. Rosen, J.B., The gradient projection method for nonlinear programming, part I, linear constraints, SIAM J., 8, 181 - 217, 1960.

22. Tremolierès, R., La méthode des centres à troncature variable, These, Paris, 1968.

23. Zoutendijk, G., Method of Feasible Direction, Elsevier Pub. Co., Amsterdam, 1960.

DUALITY THEORY IN CONVEX PROGRAMMING : A FUNCTIONAL ANALYSIS APPROACH.

Kendal Clive Jordi

DUALITY THEORY IN CONVEX PROGRAMMING : A FUNCTIONAL ANALYSIS APPROACH.

Kendal Clive Jordi

An essay submitted to the Faculty of Science in partial fulfilment of the requirements for the degree of Master of Science

University of the Witwatersrand,
Johannesburg.

1975.

Abstract.

The theory required for the proof of the Hahn-Banach Theorem is developed. This theorem in its geometric form leads to the seperation theorems for convex sets.

The theory of conjugate functions and sets is then developed. Fenchel's Duality Theorem is derived, making use of the Seperating Hyperplane Theorem. Lagrange Multipliers are discussed, and the Seperating Hyperplane Theorem is then used to develop a global duality theory for constrained optimization problems.

Gateaux and Frechet differentials are next defined. Fenchel's Theorem is then employed to derive the Kuhn-Tucker conditions for constrained optimization problems. Finally, a Dual Theorem and its converse are presented for a class of functions satisfying certain constraint qualifications.

1. INTRODUCTION.

In mathematical programming, a duality theorem gives a relationship between a constrained minimization problem and a constrained maximization problem. The two programming problems, one of which is called the primal and the other the dual, are related in such a way that the existence of a solution to one of these problems ensures the existence of a solution to the other. Furthermore, if a solution exists, the optimal function values for the two problems are equal.

It is our intention to develop a duality theory for convex nonlinear programming, using a functional analysis approach. Functional analysis is the study of vector spaces resulting from a merging of geometry, linear algebra, and analysis. Its appeal as a unifying discipline stems primarily from its geometric character. Most of the principal results in functional analysis are expressed as abstractions of intuitive geometric properties of ordinary 3-D space.

The following 4 chapters follow the presentation given by Luenberger [18] fairly closely. In the next chapter definitions of vector spaces, convex sets, transformations, Banach spaces and Hilbert spaces are given. The third chapter is concerned with the development of dual spaces, and the central mathematical result of this essay, the Hahn-Banach Theorem, is proved. This result is used to prove the seperating theorems for convex sets.

A discussion of convex and concave functionals, conjugate convex and conjugate concave functionals, and conjugate sets, follows. The Seperating Hyperplane Theorem is then used to prove Fenchel's Duality Theorem. In the fifth chapter Lagrange multipliers are covered briefly,

and the Seperating Hyperplane Theorem is used to prove the Lagrange Duality theorem for global constrained optimization.

Constrained optimization for differentiable functions is covered in the last chapter. Gateaux and Frechet differentials are defined and their relation to subgradients is discussed. Further properties of conjugate functionals are derived which are used in a derivation of a generalized Kuhn-Tucker theorem for convex functions satisfying Slater's constraint qualification. The Kuhn-Tucker conditions are then used to establish a duality theorem. The converse theorem is also derived, providing that the dual problem satisfies certain constraint qualifications. Finally, it is shown that if the primal problem has a quadratic objective function and linear constraints, its dual problem can be transformed into a maximization problem in a finite-dimensional Euclidean space.

2. LINEAR SPACES.

2.1 The Definition of a Vector Space.

A vector space X is a set of elements called vectors, together with two operations. The first operation is addition, which associates with any two vectors $x, y \in X$ a vector $x+y \in X$. The second operation is scalar multiplication, which associates with any vector $x \in X$ and any scalar α a vector αx. The set X and operations of addition and scalar multiplication satisfy the following axioms:

(i) $x + y = y + x$ commutative law,

(ii) $(x + y) + z = x + (y + z)$ associative law,

(iii) there exists $0 \in X$ such that $x + 0 = x$, $\forall\, x \in X$,

(iv) $\alpha(x + y) = \alpha x + \alpha y$,

(v) $(\alpha + \beta)x = \alpha x + \beta x$ distributive laws,

(vi) $(\alpha \times \beta)x = \alpha(\beta \times x)$,

(vii) $0 \times x = 0$, $1 \times x = x$, $\forall\, x \in X$.

2.2 Subspaces, Linear Combinations and Linear Manifolds.

2.2.1 Definition: A nonempty subset M of a vector space X is said to be a subspace of X if, for $\alpha, \beta \in \mathbb{C}$, $x, y \in M \Rightarrow \alpha x + \beta y \in M$.

Note: $0 \in M$.

2.2.2 Lemma: Let M and N be subspaces of a vector space X. Then $M \cap N$ is a subspace of X.

Proof: $M \cap N \neq \phi$, since 0 is contained in subspaces M, N.

$x, y \in M \cap N \Rightarrow x, y \in M$ and $x, y \in N$

$\Rightarrow \alpha x + \beta y \in M$ and $\alpha x + \beta y \in N$, $\alpha, \beta \in \mathbb{C}$

$\therefore \alpha x + \beta y \in M \cap N$.

2.2.3 Definition: The sum of two subsets S and T in a vector space, denoted S+T, consists of all vectors of the form s+t, where $s \in S$, $t \in T$.

2.2.4 Lemma: Let M and N be subspaces of a vector space X. Then M+N is a subspace of X.

Proof: Clearly $0 \in M+N$.

Let $x, y \in M+N$. Then there exists $m_1, m_2 \in M$, $n_1, n_2 \in N$ such that $x = m_1 + n_1$, $y = m_2 + n_2$. Therefore,

$$\begin{aligned} \alpha x + \beta y &= \alpha(m_1 + n_1) + \beta(m_2 + n_2) \\ &= (\alpha m_1 + \beta m_2) + (\alpha n_1 + \beta n_2) \\ &= m_o + n_o, \text{ where } m_o \in M, \ n_o \in N, \\ &= m + n, \text{ where } m \in M, \ n \in N. \end{aligned}$$

2.2.5 Definition: A linear combination of the vectors $x_1, x_2, \ldots, x_n$ in a vector space is a sum of the form $\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n$.

2.2.6 Definition: A necessary and sufficient condition for the set of vectors $x_1, x_2, \ldots, x_n$ to be linearly independent, is that the expression $\sum_{k=1}^{n} \alpha_k x_k = 0 \Rightarrow \alpha_k = 0 \ \forall \ k = 1, \ldots, n$.

2.2.7 Definition: Let S be a subset of the vector space X. Then the set [S], the subspace generated by S, consists of all vectors in X which are linear combinations of vectors in S.

Note: (i) That $[S]$ is a subspace follows from the fact that a linear combination of linear combinations is a linear combination.

(ii) $[S]$ is the smallest subspace containing S, in the sense that if M is a subspace containing S (i.e. it must contain all linear combinations from S), then M contains $[S]$.

2.2.8 Definition: A vector space X is the direct sum of two subspaces M and N, if every vector $x \in X$ has unique representation of the form $x = m + n$, $m \in M$, $n \in N$. Notation: $X = M \oplus N$.

2.2.9 Definition: The translation of a subspace is said to be a linear manifold. A linear manifold can be written as

$$V = x_0 + M \triangleq \{ x : x = x_0 + y : y \in M \}.$$

A linear manifold is not a linear subspace, as 0 is not necessarily a member of it. For the basic concepts of a vector space, an established standard is Halmos [7].

2.3 Convexity and Cones.

2.3.1 Definition: A set K in a linear vector space is convex if, given $x_1, x_2 \in K \Rightarrow \alpha x_1 + (1-\alpha)x_2 \in K$, $0 \leq \alpha \leq 1$.

Note that subspaces and linear manifolds are convex. The empty set is considered to be convex. The following result is elementary:

2.3.2 Lemma: Let G and K be convex sets in a vector space. Then

(i) $\alpha K = \{ x : x = \alpha k, \ k \in K \}$ is convex for any scalar α.

(ii) $K + G$ is convex.

2.3.3 Lemma: Let C be a collection of convex sets. Then $\bigcap_{K \in C} K$ is convex.

Proof: Let $G = \bigcap_{K \in C} K$. If $G = \phi$, then we have the result trivially.

Let $g_1, g_2 \in G$, and $\alpha : 0 \leq \alpha \leq 1$.

Then $g_1, g_2 \in K$ for all $K \in C$. Since K is convex,

$\alpha g_1 + (1-\alpha) g_2 \in K$ for all $K \in C$.

Thus $\alpha g_1 + (1-\alpha) g_2 \in C$.

2.3.4 Definition: A set C in a linear vector space is said to be a cone with vertex at the origin if $x \in C \Rightarrow \alpha x \in C \quad \forall\, \alpha \geq 0$.

A cone with vertex p is defined as a translation $p + C$ of a cone C with vertex at the origin.

2.4 Normed Linear Spaces.

2.4.1 Definition: A normed linear vector space $(X, ||\cdot||)$ is a real or complex linear space X and a function $||\cdot|| : X \to R$ satisfying

(i) $||x|| \geq 0 \quad \forall\, x \in X$ and $||x|| = 0$ iff $x = 0$,

(ii) $||\alpha x|| = |\alpha| \cdot ||x|| \quad \forall\, x \in X$, $\alpha \in R$ or $\mathbb{C}$,

(iii) $||x + y|| \leq ||x|| + ||y|| \quad \forall\, x, y \in X$.

2.4.2 Lemma: In a normed linear space $||x|| - ||y|| \leq ||x - y||$ for any two vectors x, y.

Proof: $||x|| - ||y|| = ||x - y + y|| - ||y||$

$\leq ||x - y|| + ||y|| - ||y|| = ||x - y||$.

2.5 Open and Closed Sets.

2.5.1 Definition: Let P be a subset of a normed space X. $p \in P$ is an interior point of P if there is an $\varepsilon > 0$ such that all vectors $x \in N(p,\varepsilon) \triangleq \{ x : ||p - x|| < \varepsilon \}$ are also members of P.

$$\overset{o}{P} \triangleq \{ p : p \text{ is an interior point of } P \}.$$

2.5.2 Definition: A set P is open if $P = \overset{o}{P}$.

2.5.3 Definition: A point $x \in X$ is said to be a closure point of a set P if given $\varepsilon > 0$, there is a point $p \in P$ satisfying $||x - p|| < \varepsilon$.

$$\bar{P} \triangleq \{ x : x \text{ is a closure point of } P \}.$$

2.5.4 Definition: A set P is closed if $P = \bar{P}$.

Suppose P is a set contained in a linear manifold V. $p \in P$ is an interior point of P relative to V if there is an $\varepsilon > 0$ such that all vectors $x \in V$ satisfying $||x - p|| < \varepsilon$ are also members of P. The set P is said to be open relative to V if every point in P is an interior point of P relative to V.

If V is the closed linear manifold generated by P, i.e. the intersection of all closed linear manifolds containing P, then x is a relative interior point of P if it is an interior point of P relative to manifold V. Similar meaning is given to relatively closed, etc.

2.6 Transformations.

2.6.1 Definition: Let X and Y be linear vector spaces and let D be a subset of X. A rule which associates with every element $x \in D$ an element $y \in Y$ is said to be a transformation from X to Y with domain D, i.e. $T : D \subset X \to Y$.
If y corresponds to x under T we write $y = Tx$.

2.6.2 Definition: If for every $y \in Y$ there is at most one $x \in D$ for which $T(x) = y$, the transformation T is said to be one-to-one.

2.6.3 Definition: If for every $y \in Y$ there is at least one $x \in D$ such that $T(x) = y$, T is said to be onto.

2.6.4 Definition: A transformation from a vector space X into the space of real (or complex) scalars is said to be a functional.

2.6.5 Definition: A transformation T mapping a vector space X into a vector space Y is said to be linear if for every $x_1, x_2 \in X$, and all scalars α_1, α_2 we have

$$T(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 T(x_1) + \alpha_2 T(x_2).$$

2.6.6 Definition: A transformation T mapping a normed space X into a normed space Y is continuous at $x_0 \in X$ if for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$||x - x_0|| < \delta \Rightarrow ||T(x) - T(x_0)|| < \varepsilon.$$

2.6.7 Lemma: Let $(X,||\cdot||)$ and $(Y,|||\cdot|||)$ be normed spaces and $T : X \to Y$ be a linear operator. Then T is continuous $\Rightarrow$ $\sup \{ |||Tx||| : x \in X, ||x|| \leq 1 \} < \infty$.

Proof: T is continuous. Therefore it is continuous at 0. Let $\varepsilon > 0$. Then there exists $\delta > 0$ such that

$$||x - 0|| < \delta \Rightarrow |||Tx - T0||| < \varepsilon,$$

so $$||x|| < \delta \Rightarrow |||Tx||| < \varepsilon.$$

Let $x \in X$ with $||x|| \leq 1$. Then

$$||(\delta/2)x|| \leq \delta/2 < \delta,$$

so $$|||T((\delta/2)x)||| < \varepsilon.$$

Hence $$|||Tx||| < 2\varepsilon/\delta.$$

Hence $$\sup \{ |||Tx||| : x \in X, ||x|| \leq 1 \} \leq 2\varepsilon/\delta.$$

2.6.8 Corollary: If $T : X \to Y$ is a linear operator, and if T is continuous, define

$$|||T||| \triangleq \sup \{ |||Tx||| : x \in X, ||x|| \leq 1 \}$$
$$\triangleq \inf \{ M : |||Tx||| \leq M||x||, \ \forall x \in X \}.$$

Then $$|||Tx||| \leq |||T|||\cdot||x|| \quad \forall x \in X.$$

Proof: If $x \in X$, $x \neq 0$, $||x/(||x||)|| \leq 1$. Then

$$|||Tx||| \leq ||x|| \sup \{ |||Tz||| : z \in X : ||z|| \leq 1 \}.$$

i.e. $\inf \{ M : |||Tx||| \leq M||x||, \ \forall x \in X \}$

$$\leq \sup \{ |||Tx||| : x \in X : ||x|| \leq 1 \}.$$

If $x \in X$, $x \neq 0$, $||x|| \leq 1$. Then $|||Tx||| \leq M$.

i.e. $\sup \{ |||Tx||| : x \in X : ||x|| \leq 1 \}$

$$\leq \inf \{ M : |||Tx||| \leq M||x|| \quad \forall x \in X \}.$$

2.7 Banach Spaces.

2.7.1 Definition: A sequence $\{x_n\}$ in a normed space is said to be a Cauchy sequence if $||x_n - x_m|| \to 0$ as $n, m \to \infty$. i.e. given $\varepsilon > 0$, there exists N such that $||x_n - x_m|| < \varepsilon \ \forall\, n, m > N$.

Note: In a normed space every convergent sequence is a Cauchy sequence, however, a Cauchy sequence may not be convergent.

2.7.2 Definition: A space in which every Cauchy sequence has a limit (and is therefore convergent) is said to be complete.

2.7.3 Definition: A set K in a normed space X is said to be compact if, given an arbitrary sequence $\{x_n\}$ in K, there is a subsequence $\{x_{n_i}\}$ converging to an element $x \in K$.

Note: In finite dimensions, compactness is to being closed and bounded, but this is not necessarily true in a general normed space.

2.7.4 Definition: A complete normed linear vector space is called a Banach space.

2.8 Hilbert Spaces.

An Hilbert space is a special form of normed space having an inner product defined which is analogous to the dot product of two vectors in analytic geometry.

2.8.1 Definition: An inner product on a linear space X is a function $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{C}$ (or R), satisfying

(i) $\langle x,x\rangle \geq 0$ and $\langle x,x\rangle = 0$ iff $x = \theta$,

(ii) $\langle \alpha x_1 + \beta x_2, y\rangle = \alpha\langle x_1, y\rangle + \beta\langle x_2, y\rangle$,

(iii) $\langle x,y\rangle = \overline{\langle y,x\rangle}$.

Then $(X,\langle\cdot,\cdot\rangle)$ is called an inner product space.

Note: (ii) shows that for fixed y, the function $\langle x,y\rangle$ is linear in x.

2.8.2 Lemma (The Cauchy Schwarz Inequality): Let $(X,\langle\cdot,\cdot\rangle)$ be an inner product space. Then

$$|\langle x,y\rangle| \leq ||x||\cdot||y|| \quad \forall\, x, y \in X \text{ where } ||x|| = \sqrt{\langle x,x\rangle}$$

and $$|\langle x,y\rangle| = ||x||\cdot||y|| \text{ iff } y = \beta x,\ \beta \in \mathbb{C},\ x \neq \theta.$$

Proof: There is an $\alpha \in \mathbb{C}$ such that $|\alpha| = 1$ and $|\langle x,y\rangle| = \alpha\langle x,y\rangle$.

Then $\langle t\alpha x + y, t\alpha x + y\rangle \geq 0 \quad \forall$ real t,

$$= t^2||x||^2 + \langle t\alpha x, y\rangle + \langle y, t\alpha x\rangle + ||y||^2$$

$$= t^2||x||^2 + 2t|\langle x,y\rangle| + ||y||^2.$$

Hence the discriminant ≤ 0, so

$$|\langle x,y\rangle|^2 \leq ||x||^2\cdot||y||^2,$$

and the discriminant $= 0$ iff there is an unique real root t_0.

Therefore $|\langle x,y\rangle| = ||x||\cdot||y||$ iff

$$\langle t_0\alpha x + y, t_0\alpha x + y\rangle = 0 \text{ iff}$$

$$t_0\alpha x + y = 0, \quad \text{i.e. } y = \beta x.$$

2.8.3 Definition: A complete inner product space is called an Hilbert space H.

2.8.4 Definition: If $x, y \in H$, and if $\langle x,y\rangle = 0$, then x is orthogonal to y and we write $x \perp y$.

If $M^{\perp}$ is a nonempty subset of H, write

$$M^{\perp} \triangleq \{ x \in H : x \perp y \ \forall\, y \in M \}.$$

2.8.5 Lemma: If M is a nonempty subset of H, then $M^{\perp}$ is a closed linear subspace of H, satisfying $M^{\perp\perp} \supset M^{\perp}$ and $M \cap M^{\perp} \subset \{0\}$.

Proof: $M^{\perp}$ is a linear space:

$x_1, x_2 \in M^{\perp}$, $\alpha, \beta \in \mathbb{C}$

$\Rightarrow \langle \alpha x_1 + \beta x_2, y\rangle = \alpha\langle x_1, y\rangle + \beta\langle x_2, y\rangle = 0$ if $y \in M$.

Therefore $\alpha x_1 + \beta x_2 \in M^{\perp}$.

$M^{\perp}$ is closed:

Let $(x_n) \subset M^{\perp}$ and let $\|x_n - x\| \to 0$, $x \in H$. If $y \in M$, then

$$|\langle x,y\rangle| = |\langle x,y\rangle - 0| = |\langle x - x_n, y\rangle|$$

$$\leq \|x - x_n\| \cdot \|y\| \quad \text{Cauchy's inequalty}$$

$$\to 0.$$

So $x \in M^{\perp} \Rightarrow M^{\perp}$ closed.

$M^{\perp\perp} \supset M^{\perp}$:

Let $x \in M$.

Then $y \in M^{\perp} \Rightarrow \langle x,y\rangle = 0$, so $\langle y,x\rangle = 0$.

So $x \in M^{\perp\perp}$.

$M \cap M^{\perp} \subset \{0\}$.

If $z \in M \cap M^{\perp}$, then $\langle z,z\rangle = 0$.

So $z = 0$.

2.8.6 Lemma: Let H be a Hilbert space, and let E be a closed linear subspace. Then $H = E \oplus E^{\perp}$.

Proof: $E \cap E^{\perp} = \{0\}$.

Let $x \in H$ and let $z \in E$ satisfying

$$||x - z|| = \inf\{\, ||x - y|| : y \in E \,\}.$$

Show $x - y \in E^{\perp}$. If $y \in E$,

$$||x - z||^2 \leq ||x - z - \alpha y||^2 \quad \text{as } E \text{ is a closed linear}$$

as E is a closed linear subspace. Hence

$$||x - z||^2 \leq ||x - z||^2 - 2\mathrm{Re}\,\bar{\alpha}\langle x - z, y\rangle + |\alpha|^2||y||^2.$$

Let $\alpha = t\langle x - z, y\rangle$, $t \in R$. Then we have

$$0 \leq (\, t^2||y||^2 - 2t \,)\, |\langle x - z, y\rangle|^2$$

and because t is an arbitrary real number,

$$\langle x - z, y\rangle = 0.$$

Therefore, $x - z \in E^{\perp}$.

For a general discussion of functional analysis, the reader should refer to the excellent introduction by Simmons [25]. Other important references include Douglas [3], Riesz and Sz-Nagy [21], Dunford and Schwartz [4], Kelley and Namioka [16], and Hille and Phillips [9].

3. DUAL SPACES.

The modern theory of optimization in normed linear spaces is largely centred about the interrelation between a space and its corresponding dual - the space consisting of all continuous linear functionals on the original space. In this chapter we consider the general construction of dual spaces and develop the central theorem of this essay, the Hahn-Banach Theorem.

3.1 Linear Functionals.

If X and Y are normed spaces, we let $L(X,Y)$ or $B(X,Y)$ denote the set of continuous linear operators from X into Y.

3.1.1 Definition: If X is a normed space, let

$$X^* \triangleq L(X,\mathbb{C}) \quad \text{if } X \text{ is over } \mathbb{C}$$
$$= L(X,\mathbb{R}) \quad \text{if } X \text{ is over } \mathbb{R}.$$

X^* is called the dual of X, and its elements are continuous linear functionals.

The value of the linear functional $x^* \in X^*$ at the point $x \in X$ is denoted by $x^*(x)$ or by $\langle x, x^* \rangle$.

3.1.2 Definition: An isometry $T : X \to X$ is a function such that

$$||Tx - Ty|| = ||x - y|| \quad \forall\, x, y \in X.$$

We now show that in an Hilbert space, bounded linear functionals are generated by elements of the space itself, and that all bounded

linear functionals on an Hilbert space are of this form.

3.1.3 Theorem (Riesz): Let H be an Hilbert space, and for each y in H, define f_y by $f_y(x) = \langle x,y\rangle$. Then the map $y \to f_y : H \to H^*$ is an isometric conjugate linear operator (from H onto H).

Proof: For each y, f_y is a linear functional, and

$$|f_y(x)| = |\langle x,y\rangle| \le ||x||\cdot||y|| \quad \text{(Cauchy Schwarz)},$$

so $$||f_y|| \le ||y||.$$

$$|f_y(y)| = \langle y,y\rangle = ||y||^2. \quad \text{So} \quad ||f_y|| \ge ||y||.$$

Therefore $||f_y|| = ||y||$ and so the map is isometric.

$$f_{\alpha y+\beta z}(x) = \langle x,\alpha y + \beta z\rangle = \bar\alpha\langle x,y\rangle + \bar\beta\langle x,z\rangle$$
$$= \bar\alpha f_y + \bar\beta f_z, \quad \text{so the map is conjugate linear.}$$

Let $f \in H^*$, and suppose $f \neq 0$.

Let $E = \ker f \ (= \{ x \in H : f(x) = 0 \})$. Then $H = E \oplus E^\perp$ by Theorem 2.8.6, and $E^\perp \neq \{0\}$ because $E \neq H$.

Let $z \ (\neq 0) \in E^\perp$. We shall show that $f = f_{\alpha z}$ for some $\alpha \in \mathbb{C}$.

If this holds, then

$$f(z) = f_{\alpha z}(z) = \langle z,\alpha z\rangle = \bar\alpha||z||^2.$$

Let $x \in H$. Then there is a $w \in E$, $\beta \in \mathbb{C}$, such that $x = w + \beta z$.

Then
$$f(x) = \beta f(z) \quad \text{because } f(w) = 0$$
$$= \beta\bar\alpha||z||^2 \quad \text{by choice of } \alpha$$
$$= \beta\langle z,\alpha z\rangle$$
$$= \langle w + \beta z,\alpha z\rangle \quad \text{because } w \perp z$$
$$= \langle x,\alpha z\rangle = f_{\alpha z}(x).$$

3.2 The Hahn-Banach Theorem.

Most of the theory of dual spaces rests on the Hahn-Banach Theorem, which asserts that any functional defined on a linear subspace of a normed linear space can be extended linearly and continuously to the whole space without increasing its norm.

3.2.1 Definition: Let f be a linear functional defined on a subspace M of a vector space X. A linear functional F is said to be an extension of f from M to N if

(i) F is defined on subspace $N \supset M$,

(ii) $Fx = fx$, $\forall\, x \in M$.

3.2.2 Definition: (a) A relation R on a nonempty set Ω is said to be a partial order if

(i) $x \,R\, x \quad \forall\, x \in \Omega$

(ii) $x \,R\, y,\ y \,R\, z \Rightarrow x \,R\, z$.

(b) A partially ordered set Ω is said to be totally ordered if $x, y \in \Omega \Rightarrow x \,R\, y$ or $y \,R\, x$.

(c) An element $x \in \Omega$ is said to be an upper bound for a subset F of Ω if $x \,R\, y$. $\forall\, y \in F$.

(d) An element $x \in \Omega$ is said to be maximal if $y \subset \Omega$, $y \,R\, x \Rightarrow x \,R\, y$,

3.2.3 Zorn's Lemma: If each totally ordered subset of a partially ordered set has an upper bound, then there is a maximal element of Ω.

3.2.4 The Hahn-Banach Theorem: Let X be a real linear normed space and Y be a subspace of X, and let $p : X \to R$ be a function satisfying

(i) $p(x + y) \leq p(x) + p(y) \quad \forall\, x, y \in X$

(ii) $p(\alpha x) = \alpha p(x) \quad \forall\, \alpha \geq 0.$

If f is a linear functional on Y satisfying $f(y) \leq p(y)$, $\forall\, y \in Y$, then there is an extension F of f to X that satisfies $F(x) \leq p(x)$ $\forall\, x$ in X.

3.2.5 Lemma: The theorem is true if $X = Y + Rx_o$.

Proof: We want an $F_o : X \to R$ linear that satisfies

$$F_o(y + \alpha x_o) \leq p(y + \alpha x_o) \quad \forall\, \alpha \in R, \ \forall\, y \text{ in } Y. \tag{3.1}$$

Dividing by $|\alpha|$, $\alpha \neq 0$, we obtain the equivalent condition

$$F_o(-y + x_o) \leq p(-y + x_o) \qquad (\alpha > 0)$$

$$F_o(y - x_o) \leq p(y - x_o) \qquad (\alpha < 0), \ \forall\, y \in Y$$

(renaming $-y = y/|\alpha|$, $y = y/|\alpha|$). This is equivalent to

$$F_o(x_o) \leq f(y) + p(x_o - y)$$

and $$f(y) - p(y - x_o) \leq F_o(x_o) \tag{3.2}$$

because F_o is to be an extension of f. Thus we can choose F_o to satisfy (3.1) iff there is a real number $F_o(x_o)$ that satisfies (3.2).

If there is a real number $\beta = F_o(x_o)$ satisfying (3.2), we define

$$F_o : X = Y + Rx_o \to R : y + \alpha x_o \to f(y) + \alpha F_o(x_o)$$

and this is linear and satisfies (3.1) by (3.2).

The inequalities (3.2) hold for some real $F_o(x_o)$ iff

$$\sup \{ f(y) - p(y - x_o) : y \in Y \} \leq \inf \{ f(z) + p(x_o - z) : z \in Y \}.$$

So we must show

$$f(y) - p(y - x_o) \leq f(z) + p(x_o - z) \quad \forall\ y,\ z \in Y.$$

Consider $f(y) - f(z) = f(y - z)$

$$\leq p(y - z)$$

$$\leq p(y - x_o) + p(x_o - y).$$

Proof of theorem: Let Ω = set of all extensions g of f to linear subspaces Dom $g \supset Y$ satisfying $g(z) \leq p(z)$, $\forall$ z in Dom g. Define $g_1 \geq g_2$ iff Dom $g_1 \supset$ Dom g_2, and g_1 is an extension (*) of g_2. Then Ω is a partially ordered non-empty set that satisfies the conditions of Zorn's Lemma.

Let Ψ be a totally ordered subset of Ω. Let

$$Z = \cup \{ \text{Dom } g : g \in \Psi \}.$$

Then Z is a linear subspace of X because Ψ is totally ordered. Define $g_o : Z \to R$ by

$$g_o(z) = g(z) \quad \text{if} \quad z \in \text{Dom } g \quad \text{for} \quad g \in \Psi.$$

g_o is well defined, by (*). Then g_o is a linear functional and

$$g_o(z) \leq p(z) \quad \forall\ z \text{ in } Z = \text{Dom } g_o.$$

Further, g_o is in Ω and is an upper bound for Ψ.

Let F be a maximal element of Ω which exists by Zorn's Lemma. If $\text{Dom } F \neq X$, then we can contradict the maximality of F by applying the lemma with $f = F$ and $Y = \text{Dom } F$.

3.2.6 Corollary: Let f be a bounded linear functional defined on a subspace M of a real normed vector space X. Then there is a bounded linear functional F defined on X which is an extension of f and which has norm equal to the norm of f on M.

Proof: Take $p(x) = ||f||_M||x||$ in the Hahn-Banach Theorem.

3.2.7 Corollary: If X is a normed linear space and x_o is a non-zero vector in X, then there exists a functional f_o in X^* such that $f_o(x_o) = ||x_o||$ and $||f_o|| = 1$.

Proof: Let $Y = \{\alpha x_o\}$ be the linear subspace of X spanned by x_o and define f on Y by $f(\alpha x_o) = \alpha||x_o||$. Clearly, f is a functional on Y such that $f(x_o) = ||x_o||$ and $||f|| = 1$. By the Hahn-Banach Theorem, f can be extended to a functional f_o in X^* with the required properties.

Among other things, this result shows that X^* seperates the vectors in X; for if x and y are any two distinct vectors so that $x - y \neq \theta$, then there exists a functional f in X^* such that $f(x - y) \neq 0$, or equivalently, $f(x) \neq f(y)$.

3.3 The Second Dual Space.

Let $x^* \in X^*$. $\langle x,x^* \rangle$ denotes the value of the functional x^* at the point $x \in X$. Now, given $x \in X$, $f(x^*) = \langle x,x^* \rangle$ defines a functional on the space X^*. The functional f defined on X^* is linear, since

$$f(\alpha x_1^* + \beta x_2^*) = \langle x, \alpha x_1^* + \beta x_2^* \rangle = \alpha \langle x, x_1^* \rangle + \beta \langle x, x_2^* \rangle$$
$$= \alpha f(x_1^*) + \beta f(x_2^*).$$

Furthermore, since

$$|f(x^*)| = |\langle x,x^* \rangle| \leq ||x|| \cdot ||x^*||,$$

it follows $||f|| \leq ||x||$. By Corollary 3.2.7, there is a non zero $x^* \in X^*$ such that

$$\langle x,x^* \rangle = ||x|| \cdot ||x^*||,$$

so $||f|| = ||x||$.

Thus, depending on whether we consider x or x^* fixed in $\langle x,x^* \rangle$, both X and X^* define bounded linear functionals on each other, which motivates the symmetric notation $\langle x,x^* \rangle$.

Since the conjugate space X^* of a normed linear space X is itself a normed linear vector space, it is possible to form a dual space $(X^*)^*$ of X^*. We denote this space by X^{**}, and call it the second dual space of X. Each point in X gives rise to a functional x^{**} in X^{**}. If $x^* \in X^*$, then $\langle x^*,x^{**} \rangle$ is called the natural mapping of X into X^{**}, i.e. $\phi : X \to X^{**} : \phi(x) = x^{**}$ maps members of X into the functionals they generate on X^* through the geometric notation. This mapping is linear, norm preserving, but, generally, not onto.

3.3.1 Definition: A normed linear space is reflexive if $X = X^{**}$.

3.3.2 Example: Any Hilbert space is reflexive. $H = H^* = H^{**}$.

3.4 The Seperation Theorems.

There is a major conceptual difference between the approach taken in the remainder of this chapter and that taken in the preceding section. Linear functionals, rather than being visualised as elements of a dual space, are visualised as hyperplanes generated in the primal space. Luenberger points out that this difference combines the relevant aspects of both the primal and the dual into a single geometric image, and thereby frees our intuition of the burden of visualising two distinct spaces.

3.4.1 Definition: A hyperplane in a normed space X is the set $\{ x \in X : \langle x, x^* \rangle = \alpha \}$ for some $\alpha \in \mathbb{C}$, $x^* \in X^*$.

Luenberger: A hyperplane H in the linear vector space X is a maximal proper linear manifold. i.e. let H be a linear manifold such that $H \neq X$. Then if V is any linear manifold containing H, either $V = X$ or $V = H$.

The geometric form of the Hahn-Banach Theorem, in its simplest form, says that given a convex set K containing an interior point, and given a point x_0 not in $\overset{o}{K}$, there is a closed hyperplane containing x_0, but disjoint from $\overset{o}{K}$.

We introduce a sublinear functional in the space that depends on the shape of K.

3.4.2 Definition: Let K be a convex set in a normed linear vector space X, and suppose $\theta \notin \overset{o}{K}$. Then the Minkowski functional p of K is defined on X by

$$p(x) \triangleq \inf \{ r : x/r \in K, \ r > 0 \}.$$

If K were the unit sphere, x_o a point such that $||x_o|| = 1$, then for every $x \in K$,

$$\langle x, x_o^* \rangle \leq ||x_o^*|| \cdot ||x|| < ||x_o^*|| \cdot ||x_o|| = \langle x_o, x_o^* \rangle$$

by the Hahn-Banach Theorem. i.e. $\{ x : \langle x, x_o^* \rangle = \langle x_o, x_o^* \rangle \}$ is a hyperplane disjoint from the interior of the unit sphere.

In the general case, $p(x)$ defines a kind of distance from the origin to x measured w.r.t. K; it is the factor by which K must be expanded so as to include x.

3.4.3 Lemma: Let K be a convex set containing θ as an interior point. Then the Minkowski functional p of K satisfies

(i) $\infty > p(x) \geq 0$, $\forall\, x \in X$,

(ii) $p(\alpha x) = \alpha p(x)$ for $\alpha > 0$,

(iii) $p(x_1 + x_2) \leq p(x_1) + p(x_2)$,

(iv) p is continuous,

(v) $\bar{K} = \{ x : p(x) \leq 1 \}$, $\overset{o}{K} = \{ x : p(x) < 1 \}$.

Proof: (i) Since K contains a sphere about θ, given x, there is an $r > 0$ such that $x/r \in K$. Therefore $p(x) < \infty$ $\forall\, x$.

Obviously $p(x) \geq 0$.

(ii) For $\alpha > 0$.

$$\begin{aligned} p(\alpha x) &= \inf \{ r : \alpha x / r \in K,\ r > 0 \} \\ &= \inf \{ \alpha r' : x/r' \in K,\ r' > 0 \} \\ &= \alpha \inf \{ r' : x/r' \in K,\ r' > 0 \} = \alpha p(x). \end{aligned}$$

(iii) Given x_1, x_2, $\epsilon > 0$, choose r_1, r_2 such that

$$p(x_i) < r_i < p(x_i) + \epsilon, \quad i = 1,2.$$

By (ii), $p(x_i/r_j) < 1$, so $x_i/r_i \in K$. Let $r = r_1 + r_2$. By the convexity of K,

$$(r_1/r)(x_1/r_1) + (r_2/r)(x_2/r_2) = (x_1 + x_2)/r \in K.$$

Thus $p(x_1 + x_2/r) \leq 1$. By (ii),

$$p(x_1 + x_2) \leq r < p(x_1) + p(x_2) + 2\varepsilon,$$

giving the result, as ε is arbitrary.

(iv) Let ε be the radius of a closed sphere centred at θ and contained in K. Then for any $x \in X$, $\varepsilon x/||x|| \in K$, thus $p(\varepsilon x/||x||) \leq 1$. By (ii), $p(x) \leq (1/\varepsilon)||x||$. i.e p is continuous at θ. From (iii)

$$p(x) = p(x - y + y) \leq p(x - y) + p(y)$$

$$p(y) = p(y - x + x) \leq p(y - x) + p(x),$$

or $$-p(y - x) \leq p(x) - p(y) \leq p(x - y)$$

from which continuity on X follows from the continuity at θ.

(v) Follows from (iv).

3.4.4 Mazur's Theorem: Let K be a convex set having a nonempty interior in a real normed linear vector space X. Suppose V is a linear manifold in X containing no interior points of K. Then there is a closed hyperplane in X containing V but containing no interior points of K. i.e. there is an element $x^* \in X^*$, and a constant c such that $\langle v, x^* \rangle = c \ \forall \ v \in V$ and $\langle k, x^* \rangle < c \ \forall \ k \in K$.

Proof: By appropriate translation, assume Θ is an interior point of K. Let M be the subspace of X generated by V. Then V is a hyperplane in M and does not contain Θ, i.e. V is a translation of a subspace N in M, say $V = x_0 + N$.

Now $x_0 \notin N$, because $\Theta \notin V$, so $[x_0 + N] = M$. For $x \in M$, $x = \alpha x_0 + n$, $n \in N$, define linear functional f on M by $f(x) = \alpha$. Then $V = \{ x : f(x) = 1 \}$.

Let p be the Minkowski functional of K. Since V contains no interior points of K, $f(x) = 1 \leq p(x)$, $\forall\, x \in V$.

For $\alpha > 0$, $f(\alpha x) = \alpha \leq p(\alpha x)$ by homogeneity

$\alpha < 0$, $f(\alpha x) \leq 0 \leq p(\alpha x)$ $\forall\, x \in V$.

Thus, $f(x) \leq p(x)$ $\forall\, x \in M$.

By the Hahn-Banach Theorem there is an extension F of f from M to X with $f(x) \leq p(x)$. Let $H = \{ x : F(x) = 1 \}$. Since $F(x) \leq p(x)$ on X and since by Lemma 3.4.3 p is continuous, F is continuous, $F(x) < 1$ for $x \in \overset{\circ}{K}$. Therefore H is the desired closed hyperplane.

3.4.5 Definition: A closed hyperplane H in a normed space X is said to be a support (or supporting hyperplane) for the convex set K if K is contained in one of the closed half spaces determined by H and H contains a point of K.

3.4.6 Theorem (Support Theorem): If x is not an interior point of a convex set K which contains interior points, there is a closed hyperplane H containing x such that K lies on one side of H.

3.4.7 Theorem (Eidelheit Seperation Theorem): Let K_1, K_2 be convex sets in X such that K_1 has interior points and K_2 contains no interior points of K_1. Then there is a closed hyperplane H seperating K_1 and K_2. i.e. there is an $x^* \in X^*$ such that

$$\sup_{x \in K_1} \langle x, x^* \rangle \leq \inf_{x \in K_2} \langle x, x^* \rangle.$$

In other words, K_1 and K_2 lie in opposite half spaces determined by H.

Proof: Let $K = K_1 - K_2$, then K contains an interior point, and θ is not one of them. By the Support Theorem, there is an $x^* \in X^*$, $x^* \neq 0$ such that $\langle x, x^* \rangle \leq 0$ for $x \in K$.

Thus for $x_1 \in K_1$, $x_2 \in K_2$, $\langle x, x^* \rangle \leq \langle x_2, x^* \rangle$.

i.e. there is a real number c such that

$$\sup_{x \in K_1} \langle x, x^* \rangle \leq c \leq \inf_{x \in K_2} \langle x, x^* \rangle.$$

The desired hyperplane is $H = \{ x : \langle x, x^* \rangle = c \}$.

3.4.8 Theorem: If K is a closed convex set in a normed space, then K is equal to the intersection of all the closed half-spaces that contain it.

The above theorem is often regarded as the geometric foundation of duality theory for convex sets. By associating closed hyperplanes (or half-spaces) with elements of X^*, the theorem expresses a convex set in X as a collection of elements in X^*.

4. CONJUGATE FUNCTIONALS AND DUALITY.

In this section we consider convex and concave functionals, from which we obtain a global theory of optimization. This development is based on the geometric representation of a nonlinear functional in terms of its graph, and builds on the theory of convex sets. The interesting theory of conjugate functionals produces a duality theory for a class of optimization problems.

4.1 Convex and Concave Functionals.

4.1.1 Definition:

(i) A real valued functional f defined on a convex subset C of a linear space is convex if

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2), \quad \forall\, x_1, x_2 \in C \text{ and } \alpha \in (0,1).$$

(ii) If strict inequality holds whenever $x_1 \neq x_2$, then f is strictly convex.

(iii) A real valued functional g defined on a convex set is said to be (strictly) concave if $-g$ is (strictly) convex.

4.1.2 Definition: A proper convex functional f on a vector space X is an everywhere defined convex functional with values $(-\infty, \infty]$ not all identically $+\infty$, i.e. $f(x) < +\infty$ for at least one $x \in X$, and $f(x) > -\infty$ for every $x \in X$.

A finite-valued convex functional defined on a nonempty convex set C in X can always be extended to a proper convex functional on X by assigning the value $+\infty$ to it outside of C.

4.1.3 Definition: Let f be a functional on the vector space X.

(i) If there exists an $\bar{x} \in X$ such that

$$f(\bar{x}) = \min_{x \in X} f(x),$$

then $\bar{x}$ is a global minimum for f.

(ii) If there exists an $\bar{x} \in X$ such that

$$x \in N(\bar{x}, \delta) \subset X \Rightarrow f(x) \geq f(\bar{x}),$$

then $\bar{x}$ is a local minimum for f.

4.1.4 Lemma: Let f be a convex functional defined on a convex subset C of a normed space. Let $\mu = \inf_{x \in C} f(x)$. Then

(i) The subset Ω of C where $f(x) = \mu$, is convex.

(ii) If $\bar{x}$ is a local minimum of f, then $f(\bar{x}) = \mu$ and, hence $\bar{x}$ is a global minimum.

Proof:

(i) Let $x_1, x_2 \in \Omega$. Then for $0 \leq \alpha \leq 1$

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2) = \mu.$$

Also, $x_1, x_2 \in C \Rightarrow \alpha x_1 + (1-\alpha)x_2 \in C$

Therefore, $f(\alpha x_1 + (1-\alpha)x_2) \geq \mu$.

Hence $f(\alpha x_1 + (1-\alpha)x_2) = \mu$.

(ii) Suppose that N is a neighbourhood about $\bar{x}$ in which $\bar{x}$ minimizes f. For any $x_1 \in C$ there exists $x \in N$ such that $x = \alpha \bar{x} + (1-\alpha)x_1$ for some α, $0 \leq \alpha \leq 1$.

Then $f(\bar{x}) \leq f(x) \leq \alpha f(\bar{x}) + (1-\alpha) f(x_1)$

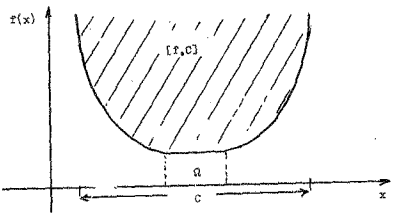
i.e. $f(\bar{x}) \leq f(x_1)$.

We now reduce the study of convex fuctionals to the study of convex sets by considering the region above the graph of the functional.

4.1.5 Definition: Let f be a convex functional defined on a convex set $C \subset X$, where X is a vector space. Define $[f,C]$ by

$$[f,C] \triangleq \{ (r,x) \in R \times X : x \in C,\ f(x) \leq r \}.$$

Example: Figure 4.1



4.1.6 Lemma: $[f,C]$ is a convex set.

Proof: Let $(r_1,x_1), (r_2,x_2) \in [f,C]$, $\alpha : 0 \leq \alpha \leq 1$.

$\alpha(r_1,x_1) + (1-\alpha)(r_2,x_2) = (\alpha r_1 + (1-\alpha)r_2, \alpha x_1 + (1-\alpha)x_2)$.

Now $\alpha x_1 + (1-\alpha)x_2 \in C$ because C is convex.

$f(x_1) \leq r_1$, $f(x_2) \leq r_2$. Therefore,

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2)$$
$$\leq \alpha r_1 + (1-\alpha)r_2.$$

i.e. $\alpha(r_1,x_1) + (1-\alpha)(r_2,x_2) \in [f,C]$.

We now discuss the properties of the set $[f,C]$.

4.1.7 Theorem: If f is a convex functional on the convex domain C in a normed space X and C has a nonempty relative interior $\overset{o}{C}$, then the convex set $[f,C]$ has a relative interior point (r_o,x_o) iff f is continuous at the point $x_o \in \overset{o}{C}$.

Proof: Assume that f is continuous at a point $x_o \in \overset{o}{C}$. Note that $v([f,C])$, the linear manifold generated by $[f,C]$ is equal to $R \times v(C)$. Given ε, $0 < \varepsilon < 1$, there is a $\delta > 0$ such that for $x \in N(x_o,\delta) \cap v(C)$ we have $x \in \overset{o}{C}$ and $|f(x) - f(x_o)| < \varepsilon$. Let $r_o = f(x_o) + 2$. Then the point $(r_o,x_o) \in [f,C]$ is a relative interior point of $[f,C]$, since $(r,x) \in [f,C]$ for $|r-r_o| < 1$ and $x \in N(x_o,\delta) \cap v(C)$.

Now suppose that (r_o,x_o) is an interior point of $[f,C]$. Then there is $\varepsilon_o > 0$, $\delta_o > 0$ such that for $x \in N(x_o,\delta_o) \cap v(C)$ and $|r-r_o| < \varepsilon_o$ we have $r \geq f(x)$. Thus f is bounded above by $f(x_o)+\varepsilon_o$ on the neighbourhood $N(x_o,\delta_o) \cap v(C)$. Without loss of generality, assume $x_o = \theta$, $f(x_o) = 0$. For any ε, $0 < \varepsilon < 1$, and for any $x \in N(x_o,\varepsilon\delta_o) \cap v(C)$

$$f(x) = f\{(1-\varepsilon)\theta + \varepsilon(1/\varepsilon)x\} \leq (1-\varepsilon)f(\theta) + \varepsilon f((1/\varepsilon)x) \leq \varepsilon\varepsilon_o$$

where ε_o is the bound on f in $N(x_o,\delta_o) \cap v(C)$. Further,

$$0 = f(\theta) = f\{(1/1+\varepsilon)x + (1 - 1/1+\varepsilon)(-1/\varepsilon)x\}$$
$$\leq (1/1+\varepsilon)f(x) + (1 - 1/1+\varepsilon)f((-1/\varepsilon)x).$$

So $f(x) \geq -\varepsilon f((-1/\varepsilon)x) \geq -\varepsilon\varepsilon_o$. Therefore for $x \in N(x_o,\varepsilon\delta_o) \cap v(C)$, we have $|f(x)| \leq \varepsilon\varepsilon_o$. Thus f is continuous at x_o.

4.1.8 Theorem: A convex functional f defined on a convex domain C and continuous at a single point in the relative interior $\overset{o}{C}$ of C is continuous throughout $\overset{o}{C}$.

Proof: Without loss of generality assume f is continuous at $\theta \in \overset{o}{C}$ $f(\theta) = 0$. By restricting our attention to $\nu(C)$, we may assume C has interior points rather than relative interior points. Let y be an arbitrary point in $\overset{o}{C}$. Since $\overset{o}{C}$ is (relatively) open, there is a $\beta > 1$ such that $\beta y \in \overset{o}{C}$. Given $\varepsilon > 0$, let $\delta > 0$ be such that $||x|| < \delta$ implies $|f(x)| < \varepsilon$. Then for $||z - y|| < (1 - 1/\beta)\delta$ we have

$$z = y + (1 - 1/\beta)x = 1/\beta(\beta y) + (1 - 1/\beta)x$$

for some $x \in \overset{o}{C}$ with $||x|| < \delta$. Thus $z \in C$ and

$$\begin{aligned} f(z) &\leq 1/\beta f(\beta y) + (1 - 1/\beta)f(x) \\ &< 1/\beta f(\beta y) + (1 - 1/\beta)\varepsilon. \end{aligned}$$

Thus f is bounded above in the sphere $||z - y|| < (1 - 1/\beta)\delta$. Therefore, for r sufficiently large, the point (r,y) is an interior point of $[f,C]$. Thus, by Theorem 4.1.7, f is continuous at y.

4.2 Conjugate Convex Functionals and Support Functionals.

We now investigate the dual representation of the set $[f,C]$ in terms of closed hyperplanes. From this we obtain a very general duality principle for optimization problems.

4.2.1 Definition: Let f be a convex functional defined on a convex set C in a normed space X. The conjugate set C^* is defined as

$$C^* \triangleq \{ x^* \in X^* : \sup_{x \in C} (\langle x,x^* \rangle - f(x)) < \infty \}.$$

and the functional f^* conjugate to f is defined on C^* by

$$f^*(x^*) \triangleq \sup_{x \in C} (\langle x,x^* \rangle - f(x)).$$

4.2.2 Definition:

$$[f^*,C^*] \triangleq \{ (s,x^*) \in R \times X^* : x^* \in X^*, f^*(x^*) \leq s \}.$$

4.2.3 Lemma:

(i) f^* is a convex functional and C^* is a convex set.

(ii) $[f^*,C^*]$ is a closed convex subset of $R \times X^*$.

Proof: (i) Let $x_1^*, x_2^* \in X^*$, and $\alpha: 0 < \alpha < 1$.

$$\sup_{x \in C} (\langle x,\alpha x_1^* + (1-\alpha)x_2^* \rangle - f(x))$$

$$= \sup_{x \in C} \{ \alpha(\langle x,x_1^* \rangle - f(x)) + (1-\alpha)(\langle x,x_2^* \rangle - f(x)) \}$$

$$\leq \alpha \sup_{x \in C} (\langle x,x_1^* \rangle - f(x)) + (1-\alpha) \sup_{x \in C} (\langle x,x_2^* \rangle - f(x)).$$

(ii) Let $\{(s_i,x_i^*)\}$ be a sequence from $[f^*,C^*]$ such that $(s_i,x_i^*) \to (s,x^*)$ as $i \to \infty$. Then $\forall\, i, \forall\, x \in C$, $s_i \geq f^*(x_i^*) \geq \langle x,x_i^* \rangle - f(x)$. Letting $i \to \infty$, $s \geq \langle x,x^* \rangle - f(x)$,

4.2 Conjugate Convex Functionals and Support Functionals.

We now investigate the dual representation of the set $[f,C]$ in terms of closed hyperplanes. From this we obtain a very general duality principle for optimization problems.

4.2.1 Definition: Let f be a convex functional defined on a convex set C in a normed space X. The conjugate set C^* is defined as

$$C^* \triangleq \{ x^* \in X^* : \sup_{x \in C} (\langle x,x^* \rangle - f(x)) < \infty \},$$

and the functional f^* conjugate to f is defined on C^* by

$$f^*(x^*) \triangleq \sup_{x \in C} (\langle x,x^* \rangle - f(x)).$$

4.2.2 Definition:

$$[f^*,C^*] \triangleq \{ (s,x^*) \in R \times X^* : x^* \in X^*, f^*(x^*) \leq s \}.$$

4.2.3 Lemma:

(i) f^* is a convex functional and C^* is a convex set.

(ii) $[f^*,C^*]$ is a closed convex subset of $R \times X^*$.

Proof: (i) Let $x_1^*, x_2^* \in X^*$, and α: $0 < \alpha < 1$.

$$\sup_{x \in C} (\langle x,\alpha x_1^* + (1-\alpha)x_2^* \rangle - f(x))$$

$$= \sup_{x \in C} \{ \alpha(\langle x,x_1^* \rangle - f(x)) + (1-\alpha)(\langle x,x_2^* \rangle - f(x)) \}$$

$$\leq \alpha \sup_{x \in C} (\langle x,x_1^* \rangle - f(x)) + (1-\alpha) \sup_{x \in C} (\langle x,x_2^* \rangle - f(x)).$$

(ii) Let $\{(s_i,x_i^*)\}$ be a sequence from $[f^*,C^*]$ such that $(s_i,x_i^*) \to (s,x^*)$ as $i \to \infty$. Then $\forall\, i, \forall\, x \in C$, $s_i \geq f^*(x_i^*) \geq \langle x,x_i^* \rangle - f(x)$. Letting $i \to \infty$, $s \geq \langle x,x^* \rangle - f(x)$,

$\forall\, x \in C$. Therefore $s \geq \sup_{x \in C} (\langle x, x^* \rangle - f(x))$. i.e. $x^* \in C^*$, $s \geq f^*(x^*)$.

4.2.3 Definition: Given any convex set $C \subset X$, define $f(x) \equiv 0$ on C. The dual set to $[f,C]$ is of the form $[h,C^*]$ where

$$h(x^*) = \sup_{x \in C} \langle x, x^* \rangle$$

and the domain C^* is the set of vectors for which

$$\sup_{x \in C} \langle x, x^* \rangle < \infty.$$

Call $h(x^*)$ the support functional of the set C.

We write $[f,C]^* = [f^*, C^*]$. We now establish a relationship between the conjugate functional and seperating hyperplanes. On $R \times X$, closed hyperplanes are represented by

$$sr + \langle x, x^* \rangle = k$$

where s, k and x^* determine the hyperplane. A hyperplane is nonvertical for the defining linear functional (s,x^*) if $s \neq 0$. We restrict our attention to nonvertical hyperplanes, and without any further loss of generality we consider only those linear functionals of the form $(-1,x^*)$. Any nonvertical closed hyperplane can then be obtained by appropriate choice of x^* and k.

Now as k varies, the solutions (r,x) of $\langle x, x^* \rangle - r = k$ describe parallel hyperplanes in $R \times X$ (See Fig 4.2). Now $f^*(x^*)$ is the supremum of the values of k for which the hyperplane intersects $[f,C]$. Thus $\langle x, x^* \rangle - r = f^*(x^*)$ is a support hyperplane of $[f,C]$.

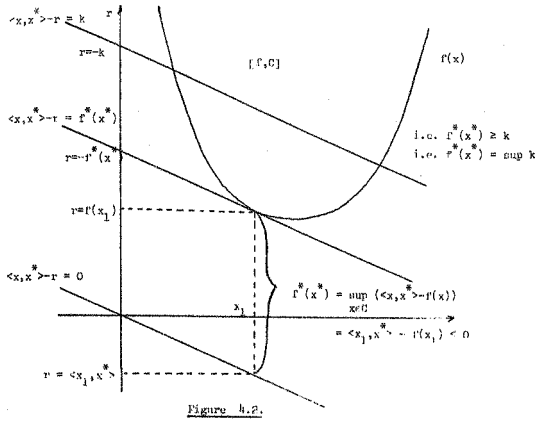


Figure 4.2.

The most important interpretation of the conjugate functional for application to optimization problems is that it measures the vertical distance from the origin to the support hyperplane, i.e. the hyperplane $\langle x,x^* \rangle - r = f^*(x^*)$ intersects the vertical axis ($x = 0$) at $(-f^*(x^*), 0)$.

Remark: Given the point $(s,x^*) \in [f^*,C^*]$, associate the half-space consisting of all $(r,x) \in R \times X$ satisfying

$$\langle x,x^* \rangle - r \leq s.$$

Then the set $[f^*,C^*]$ associates those (nonvertical) halfspaces that contain the set $[f,C]$. Hence $[f^*,C^*]$ is the dual representation of $[f,C]$.

4.3 Conjugate Concave Functionals.

4.3.1 Definition: Given a concave functional g defined on a convex subset D of a vector space, define

$$[g,D] \triangleq \{ (r,x) : x \in D, \quad r \leq g(x) \}.$$

4.3.2 Lemma: $[g,D]$ is a convex set.

Proof: Similar to that for $[f,C]$.

4.3.3 Definition: Let g be a concave functional on the convex set D. The conjugate set D^* is defined as

$$D^* \triangleq \{ x^* : \inf_{x \in D} (\langle x,x^* \rangle - g(x)) > -\infty \}$$

and the functional g^* conjugate to g is defined as

$$g^*(x^*) \triangleq \inf_{x \in D} (\langle x,x^* \rangle - g(x)).$$

It can be verified that D^* is convex and g^* is concave. Write $[g,D]^* = [g^*,D^*]$.

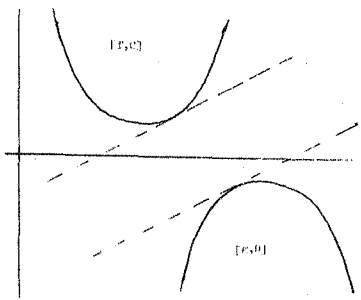
Note: Our notation does not distinguish completely between convex and concave functionals. It must be made clear which is being employed in any given context.

The interpretation of concave conjugate functionals is similar to that for convex conjugate functionals. The hyperplane $\langle x,x^* \rangle - r = g^*(x^*)$ supports the set $[g,D]$. Furthermore, $-g^*(x^*)$ is the intercept of that hyperplane with the vertical axis.

4.4 Dual Optimization Problems.

We now apply the theory of conjugate functionals to optimization. Let f be a convex functional over convex set C and g be a concave functional over convex set D. We seek $\inf_{C \cap D} (f(x) - g(x))$.

Figure 4.3



We wish to find the minimal vertical seperation between the sets (See Fig. 4.3). This is equivalent to finding the maximal vertical seperation of two parallel hyperplanes supporting $[f,C]$ and $[g,D]$. Therefore, $g^*(x^*) - f^*(x^*)$ is the vertical seperation of the two hyperplanes.

4.4.1 Theorem (Fenchel Duality Theorem): Let f, g be convex and concave functions respectively on the convex sets C and D in a normed space X. Let $C \cap D$ contain points in the relative interior of

C and D and/or [f,C] or [g,D] have nonempty interior. Let

$$\mu = \inf_{x \in C \cap D} (f(x) - g(x)) < \infty.$$

Then

$$\mu = \inf_{x \in C \cap D} (f(x) - g(x)) = \max_{x^* \in C^* \cap D^*} (g^*(x^*) - f^*(x^*))$$

where the maximum on the right is achieved by some $x_o^* \in C^* \cap D^*$. If the infinum on the left is achieved by some $x_o \in C \cap D$, then

$$\max_{x \in C} (\langle x, x_o^* \rangle - f(x)) = \langle x_o, x_o^* \rangle - f(x_o)$$

and

$$\min_{x \in D} (\langle x, x_o^* \rangle - g(x)) = \langle x_o, x_o^* \rangle - g(x_o).$$

Proof: By definition, $\forall\, x^* \in C^* \cap D^*$, $x \in C \cap D$

$$f^*(x^*) \geq \langle x, x^* \rangle - f(x)$$
$$g^*(x^*) \leq \langle x, x^* \rangle - g(x).$$

Thus

$$f(x) - g(x) \geq g^*(x^*) - f^*(x^*).$$

Hence

$$\inf_{x \in C \cap D} (f(x) - g(x)) \geq \sup_{x^* \in C^* \cap D^*} (g^*(x^*) - f^*(x^*))$$

Now, the convex set $[f-\mu, C]$ is a vertical displacement of $[f,C]$. By definition of μ, the sets $[f-\mu, C]$ and $[g, D]$ have disjoint relative interiors, however arbitrarily close. Since one of these sets has nonempty interior, there is a closed hyperplane in $R \times X$ seperating them. This hyperplane cannot be vertical otherwise its vertical projection on X would seperate C and D.

This nonvertical hyperplane can be represented as

$\{(t,x) \in R \times X : \langle x,x_0^* \rangle - r = c\}$, for some $x_0^* \in X^*$, $c \in R$.

Now since $[g,D]$ lies below, and maybe on, this hyperplane

$$c = \inf_{x \in D} (\langle x,x_0^* \rangle - g(x)) = g^*(x_0^*).$$

Likewise

$$c = \sup_{x \in C} (\langle x,x_0^* \rangle - f(x)+\mu) = f^*(x_0^*) + \mu.$$

Therefore,

$$\mu = g^*(x_0^*) - f^*(x_0^*).$$

i.e. there is an $x_0^* \in C^* \cap D^*$ such that

$$\inf_{C \cap D} (f(x) - g(x)) = g^*(x_0^*) - f^*(x_0^*).$$

Therefore the equality in the theorem is proved.

If the infimum μ is achieved by some $x_0 \in C \cap D$, then $(g(x_0),x_0) \in [f-\mu,C] \cap [g,D]$, and $(g(x_0),x_0)$ lies on the seperating hyperplane.

In applying this theorem to minimize a convex functional f on a convex domain D (the set D representing constraints), take $C = X$ and $g \equiv 0$ in this theorem.

Conjugate functionals on finite-dimensional spaces were introduced by Fenchel [5]. For an excellent presentation of this topic, consult Karlin [14]. Some extensions and related discussions are to be found in Rockafellar [23], Radstrom [20] and Whinston [28]. Karlin [13]-[15], Luenberger [18] demonstrate how the well-known Min-Max Theorem, of special importance in Game Theory, can be derived from Fenchel's Duality Theorem, for reflexive spaces.

We shall make use of the Fenchel Duality Theorem later on in this essay to obtain a duality theory for a class of optimization problems.

5. THE GLOBAL THEORY OF CONSTRAINED OPTIMIZATION AND DUALITY.

The general optimization problem treated in this essay is to minimize a functional f within a given subset of a vector space. Lagrange multipliers will dominate our attention, as they somehow always unscramble a difficult constrained problem.

Although we have encountered Lagrange multipliers at several points in the last chapter, they were treated as the result of certain duality calculations. It will be seen that the Lagrange Multiplier can be interpreted as a hyperplane. As a result, it may be suspected that the theory is simplest and most elegant for problems involving convex function. Indeed this is so. In this chapter, we therefore consider a global or convex theory based on the geometric interpretation in the constraint space where the Lagrange multiplier appears as a seperating hyperplane.

5.1 Positive Cones and Convex Mappings.

5.1.1 Definition: Let P be a convex cone in a vector space X. For $x, y \in X$, write $x \geq y$ (w.r.t P) if $x-y \in P$. The cone defining this relation is called the positive cone in X. The cone $N = -P$ is called the negative cone in X and we write $y \leq x$ for $y-x \in N$.

5.1.2 Definition: Given a normed space X together with a positive convex cone $P \subset X$. Define a corresponding cone $P^{\oplus}$ in the dual space X^* by

$$P^{\oplus} \triangleq \{ x^* \in X^* : \langle x, x^* \rangle \geq 0, \quad \forall x \in P \}.$$

5.1.3 Lemma: Let the positive cone P in the normed space X be closed. If $x \in X$ satisfies $\langle x, x^* \rangle \geq 0 \ \forall \ x^* \geq \theta$, then $x \geq \theta$.

Proof: Assume $x \notin P$. Then by the seperating hyperplane theorem, there is a $x^* \in X^*$ such that $\langle x, x^* \rangle < \langle p, x^* \rangle \ \forall \ p \in P$.

Since P is a cone, $\langle p, x^* \rangle$ can never be negative, otherwise $\langle x, x^* \rangle > \langle \alpha p, x^* \rangle$ for some $\alpha > 0$. Thus $x^* \in P^{\oplus}$. Also, $\inf_{p \in P} \langle p, x^* \rangle = 0$, so $\langle x, x^* \rangle < 0$.

5.1.4 Definition: Let X be a vector space and Z be a vector space having positive cone P. A mapping $G : X \to Z$ is said to be convex if the domain Ω of G is a convex set, and if for $\alpha : 0 < \alpha < 1$,

$$G(\alpha x_1 + (1-\alpha)x_2) \leq \alpha G(x_1) + (1-\alpha)G(x_2), \ \forall \ x_1, x_2 \in \Omega.$$

5.1.5 Lemma: Let G be a convex mapping as defined above. Then for every $z \in Z$ the set $\{ x : x \in \Omega \ ; \ G(x) \leq z \}$ is convex.

Proof: Let $x_1, x_2 \in \{ x : x \in \Omega : G(x) \leq z \}$.

Then $\alpha x_1 + (1-\alpha)x_2 \in \Omega$ because Ω is convex. For $\alpha : 0 < \alpha < 1$,

$$G(\alpha x_1 + (1-\alpha)x_2) \leq \alpha G(x_1) + (1-\alpha)G(x_2)$$

$$\leq \alpha z + (1-\alpha)z = z,$$

5.2 Lagrange Multipliers.

Consider the problem

$$\text{minimize} \quad f(x)$$

$$\text{subject to} \quad G(x) \leq \Theta, \quad x \in \Omega.$$

Ω is a convex subset of the vector space X, f is a real valued convex functional on Ω. G is a convex mapping from Ω into the normed space Z, having positive cone P.

Imbed the above problem in the family of problems

$$\text{minimize} \quad f(x)$$

$$\text{subject to} \quad G(x) \leq z, \quad x \in \Omega, \quad \text{where } z \in Z.$$

Define $\Gamma \subset Z$ as

$$\Gamma \triangleq \{ z : \text{There is an } x \in \Omega \text{ with } G(x) \leq z \},$$

i.e. the set of z such that the problem has at least one feasible solution.

5.2.1 Lemma: Γ is a convex set.

Proof: Let $z_1, z_2 \in \Gamma$. i.e. there exists $x_1, x_2 \in \Omega$ with $G(x_1) \leq z_1$ and $G(x_2) \leq z_2$. Then for $0 < \alpha < 1$,

$$G(\alpha x_1 + (1-\alpha)x_2) \leq \alpha z_1 + (1-\alpha)z_2.$$

5.2.2 Definition: On the set Γ, define the primal functional ω (which may or may not be finite) as

$$\omega(z) = \inf \{ f(x) : x \in \Omega, G(x) \leq z \}.$$

The original problem can then be regarded as determining the value $\omega(\theta)$.

5.2.3 Lemma: The functional ω is convex.

Proof: $\omega(\alpha z_1 + (1-\alpha)z_2)$

$$= \inf \{ f(x) : x \in \Omega : G(x) \leq \alpha z_1 + (1-\alpha)z_2 \}$$

$$\leq \inf \{ f(x) : x = \alpha x_1 + (1-\alpha)x_2,\ x_1 \in \Omega,\ x_2 \in \Omega : G(x_1) \leq z_1,\ G(x_2) \leq z_2 \}$$

$$\leq \alpha \inf \{ f(x_1) : x_1 \in \Omega,\ G(x_1) \leq z_1 \} + (1-\alpha)\inf \{ f(x_2) : x_2 \in \Omega : G(x_2) \leq z_2 \}$$

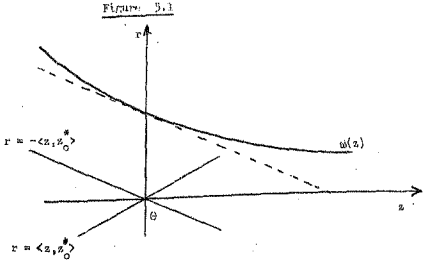
$$\leq \alpha\omega(z_1) + (1-\alpha)\omega(z_2).$$

5.2.4 Lemma: The functional ω is decreasing. i.e. if $z_1 \leq z_2$ then $\omega(z_1) \geq \omega(z_2)$.

Proof: $\omega(z_1) = \inf \{ f(x) : x \in \Omega : G(x) \leq z_1 \}$

$$\geq \inf \{ f(x) : x \in \Omega : G(x) \leq z_2 \} = \omega(z_2).$$

Figure 5.1



Since ω is convex it has a supporting hyperplane at $z = 0$, and lying below ω over all z. By adding an appropriate linear functional $\langle z, z_0^* \rangle$ to $\omega(z)$, the resulting combination $\omega(z) + \langle z, z_0^* \rangle$ is minimized at $z = 0$ (See Fig. 5.1). The functional z_0^* is the Lagrange multiplier for the problem; the tangent hyperplane illustrated in Fig. 5.1 corresponds to the element $(1, z_0^*) \in R \times Z^*$.

Lagrange multipliers for problems having inequality constraints were first treated explicitly by John [12] and Kuhn and Tucker [17].

5.2.5 Theorem: Let X be a linear vector space, Z a normed space, Ω a convex subset of X and P the positive cone in Z. Assume P contains an interior point. Let

$$f : \Omega \to R, \quad G : \Omega \to Z, \quad f, G \text{ convex.}$$

Assume there exists $x_1 \in \Omega$ for which $G(x_1) < 0$. (i.e. $G(x_1)$ is an interior point of $N = -P$). Let

$$\mu_0 = \inf f(x), \text{ subject to } x \in \Omega, G(x) \leq 0. \tag{5.1}$$

Assume $\mu_0 < \infty$. Then there is an element $z_0^* \geq 0$ in Z^* such that

$$\mu_0 = \inf_{x \in \Omega} \{ f(x) + \langle G(x), z_0^* \rangle \}. \tag{5.2}$$

Furthermore, if the infimum is achieved in (5.1) by an $x_0 \in \Omega$, $G(x_0) \leq 0$, it is achieved by x_0 in (5.2) and

$$\langle G(x_0), z_0^* \rangle = 0. \tag{5.3}$$

Proof: In $R \times Z$, define

$$A = \{ (r,z) : r \geq f(x), z \geq G(x) \text{ for some } x \in \Omega \},$$

$$B = \{ (r,z) : r \leq \mu_0, z \leq 0 \}.$$

A and B are convex, because f and G are convex.

By definition of μ_0, A contains no interior points of B. Now, N contains an interior point, therefore B contains an interior point. Applying the seperating hyperplane theorem, there exists $0 \neq (r_0, z_0^*) \in R \times Z^*$ such that

$$r_0 r_1 + \langle z_1, z_0^* \rangle \geq r_0 r_2 + \langle z_2, z_0^* \rangle$$

for $(r_1, z_1) \in A$, $(r_2, z_2) \in B$. From the nature of B it follows that $r_0 \geq 0$, $z_0^* \geq 0$. Show that $r_0 > 0$.

$(\mu_0, 0) \in B$, therefore

$$r_0 r + \langle z, z_0^* \rangle \geq r_0 \mu_0 \quad \forall (r, z) \in A.$$

Suppose that $r_0 = 0$. Then, in particular, $\langle G(x_1), z_0^* \rangle \geq 0$ and $z_0^* \neq 0$. Since $G(x_1)$ is an interior point of N, i.e. $G(x_1) < 0$, and $z_0^* \geq 0$, it follows that $\langle G(x_1), z_0^* \rangle < 0$ (if $\langle G(x_1), z_0^* \rangle \geq 0$ $\Rightarrow G(x_1) \geq 0$ by Lemma 5.1.3). Contradiction.

Therefore, $r_0 > 0$. Take $r_0 = 1$ (without loss of generality). Then

$$\mu_0 \leq \inf_{(r,z) \in A} (r + \langle z, z_0^* \rangle)$$

$$\leq \inf_{x \in \Omega} (f(x) + \langle G(x), z_0^* \rangle)$$

$$\leq \inf_{\substack{x \in \Omega \\ G(x) \leq 0}} f(x) = \mu_0.$$

i.e. $$\mu_0 = \inf_{x \in \Omega} (f(x) + \langle G(x), z_0^* \rangle).$$

If there exists an x_0 such that $G(x_0) \leq 0$, $\mu_0 = f(x_0)$, then

$$\mu_0 \leq f(x_0) + \langle G(x_0), z_0^* \rangle \leq f(x_0) = \mu_0$$

and hence $\langle G(x_0), z_0^* \rangle = 0$.

Note: The condition $G(x_1) < 0$, called a regularity condition, is typical of the assumptions that must be made in Lagrange multiplier theorems. It guarantees that the seperating hyperplane is nonvertical. i.e. $r_0 \neq 0$.

5.2.6 Corollary: Let the conditions of Theorem 5.2.5 be satisfied and assume that x_0 achieves the constrained miminum. Then there is a $z_0^* \geq 0$ such that the Lagrangian

$$L(x,z^*) = f(x) + \langle G(x), z^* \rangle$$

has a saddle point, i.e.

$$L(x_0,z^*) \leq L(x_0,z_0^*) \leq L(x,z_0^*) \quad \forall\, x \in \Omega,\ z^* \geq 0.$$

Proof: From (5.2), $L(x_0,z_0^*) \leq L(x,z_0^*)$, and

$$L(x_0,z^*) - L(x_0,z_0^*)$$

$$= \langle G(x_0), z^* \rangle - \langle G(x_0), z_0^* \rangle$$

$$= \langle G(x_0), z^* \rangle \leq 0, \quad \text{by (5.3)}.$$

An equality constraint of the form $H(x) = 0$, with $H(x) \equiv Ax - b$ where $A : X \to Y$ is equivalent to two convex inequalities $H(x) \leq 0$ and $-H(x) \leq 0$, and can be included in the constraints $G(x) \leq 0$. However, these cannot be included trivially in the above theorem as there never exists x_1 which satisfies $H(x_1) < 0$ and $-H(x_1) < 0$ (See Luenberger).

It should be noted that we are primarily concerned with the part played by Lagrange multipliers in Duality theory. Thus it is not our intention to give, or exploit, all the geometric properties of the Lagrange multipliers and of the problem, in this essay.

5.3 Duality.

Consider again the basic convex programming problem

$$\text{minimize} \quad f(x)$$

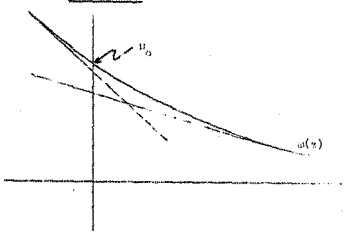
$$\text{subject to} \quad G(x) \leq \theta, \quad x \in \Omega.$$

f, G, Ω are convex. As previously, define

$$\omega(z) \triangleq \inf \{ f(x) : G(x) \leq z, \quad x \in \Omega \}, \tag{5.4}$$

and let $\mu_0 = \omega(\theta)$.

Figure 5.2.



μ_0 is equal to the maximum intercept with the vertical axis of all closed hyperplanes lying below ω (See Fig. 5.2); that hyperplane being determined by the Lagrange multiplier of the problem.

Define the dual functional ϕ corresponding to (5.4) to be defined in the positive cone in Z^* as

$$\phi(z^*) \triangleq \inf_{x \in \Omega} (f(x) + \langle G(x), z^* \rangle).$$

5.3.1 Lemma: The dual functional is concave and can be expressed as

$$\phi(z^*) = \inf_{z \in \Gamma} \; (\omega(z) + \langle z, z^* \rangle).$$

Proof: $\alpha : 0 < \alpha < 1$. $z_1^*, z_2^* \in Z^*$.

$$\phi(\alpha z_1^* + (1-\alpha) z_2^*)$$

$$= \inf_{x \in \Omega} \; (f(x) + \langle G(x), \alpha z_1^* + (1-\alpha) z_2^* \rangle)$$

$$\geq \inf_{x \in \Omega} \; (\alpha f(x) + \alpha \langle G(x), z_1^* \rangle) + \inf_{x \in \Omega} \; ((1-\alpha) f(x) + (1-\alpha) \langle G(x), z_2^* \rangle)$$

$$= \alpha \phi(z_1^*) + (1-\alpha) \phi(z_2^*).$$

For any $z^* \geq 0$, $z \in \Gamma$,

$$\phi(z^*) = \inf_{x \in \Omega} \; (f(x) + \langle G(x), z^* \rangle)$$

$$\leq \inf \; (f(x) + \langle z, z^* \rangle : G(x) \leq z, \; x \in \Omega)$$

$$= \omega(z) + \langle z, z^* \rangle.$$

Now, for any $x_1 \in \Omega$, with $z_1 = G(x_1)$

$$f(x_1) + \langle G(x_1), z^* \rangle$$

$$\geq \inf \; (f(x) + \langle z_1, z^* \rangle : G(x) \leq z_1, \; x \in \Omega)$$

$$= \omega(z_1) + \langle z_1, z^* \rangle.$$

Therefore

$$\phi(z^*) \geq \inf_{z \in \Gamma} \; (\omega(z) + \langle z, z^* \rangle)$$

Thus

$$\phi(z^*) = \inf_{z \in \Gamma} \; (\omega(z) + \langle z, z^* \rangle).$$

Therefore, ϕ is essentially the conjugate functional of ω. Although there is a sign discrepancy, the concepts are the same. $(1,z^*) \in R \times Z^*$ determines a family of hyperplanes in $R \times Z$, each hyperplane satisfying $r + \langle z, z^* \rangle = k$. For $k = \phi(z^*)$, this hyperplane supports the set $[\omega,\infty]$, the region above the graph of ω. Furthermore at $z = 0$, $r = \phi(z^*)$; hence $\phi(z^*)$ is equal to the intercept of this hyperplane with the vertical axis.

5.3.2 Theorem (Lagrange Duality Theorem): Let X be a vector space, Z be a normed space, Ω a convex subset of X, f, G convex, such that $f : \Omega \subset X \to R$, $G : X \to Z$. Suppose there exists x_1 such that $G(x_1) \leq 0$ and that

$$\mu_0 = \inf \{ f(x) : G(x) \leq 0, \quad x \in \Omega \}$$

is finite. Then

$$\inf_{\substack{x \in \Omega \\ G(x) \leq 0}} f(x) = \max_{z^* \geq 0} \phi(z^*) \tag{5.5}$$

and the maximum on the right is achieved by some $z_0^* \geq 0$. If the infimum on the left is achieved by some $x_0 \in \Omega$, then $\langle G(x_0), z_0^* \rangle = 0$ and x_0 minimizes

$$f(x) + \langle G(x), z_0^* \rangle, \quad x \in \Omega.$$

Proof: Let $z^* \geq 0$.

$$\inf_{x \in \Omega} \{ f(x) + \langle G(x), z^* \rangle \}$$

$$\leq \inf_{\substack{x \in \Omega \\ G(x) \leq 0}} \{ f(x) + \langle G(x), z^* \rangle \}$$

$$\leq \inf_{\substack{x \in \Omega \\ G(x) \leq 0}} f(x) = \mu_0.$$

i.e. $\max_{z^* \geq 0} \phi(z^*) \leq \mu_0$.

Conversely, Theorem 5.2.5 establishes the existence of z_0^* which gives equality. The remainder of the theorem is also proved in Theorem 5.2.5.

Since ω is decreasing, $\omega(0) \leq \omega(z)$ for $z \leq 0$. Hence, an alternative symmetric formulation of (5.5) is

$$\min_{z \leq 0} \omega(z) = \max_{z^* \geq 0} \phi(z^*).$$

5.3.3 Example: As a simple application of the duality theorem, we calculate the dual of the quadratic program:

$$\text{minimize } \tfrac{1}{2} x^T Q x - b^T x$$
$$\text{subject to } A x \leq c,$$

where $x \in R^n$, to be determined, $b \in R^n$, $c \in R^m$, A is an $m \times n$ matrix, and Q is an $n \times n$ positive definite symmetric matrix. Assuming there is an x satisfying $Ax \leq c$, the problem is equivalent to

$$\max_{\lambda \geq 0} \min_{x} \{ \tfrac{1}{2} x^T Q x - b^T x + \lambda^T (Ax - c) \}.$$

The minimization over x is unconstrained and attained by

$$x = Q^{-1}(b - A^T \lambda).$$

Substituting this above, the problem becomes

$$\max_{\lambda \geq 0} \{ -\tfrac{1}{2} \lambda^T P \lambda - \lambda^T d - \tfrac{1}{2} b^T Q^{-1} b \},$$

where $P = A Q^{-1} A^T$, $d = c - A Q^{-1} b$.

Thus the dual is also a quadratic programming problem. Note that the dual problem may be much easier to solve than the primal problem since the constraint set is much simpler and the dimension is smaller if $m < n$.

The approach presented in this chapter, which required no differentiability assumptions, was developed by Slater [26] and extended to infinite-dimensional spaces by Hurwicz [11]. In fact, our presentation closely follows Hurwitz. For some interesting extensions, see Arrow, Hurwitz and Uzawa [1]. In the absence of convexity constraints and an interior point, if the hyperplane does exist, the Lagrange technique for location of the optimum still applies. See Gould [6].

6. THE LOCAL THEORY OF CONSTRAINED OPTIMIZATION AND DUALITY.

Historically the local theory of Lagrange multipliers, stated in differential form, predates the global theory presented in the last chapter by almost a century. Its wider range of applicability, and its general convenience for most problems, continues to make the local theory the better known and most used of the two.

6.1 Gateaux and Frechet Differentials.

Let X be a vector space, Y a normed space, and T a transformation defined on a domain $D \subset X$, with range $R \subset Y$.

6.1.2 Definition: Let $x \in D$ and let h be arbitrary in X. If

$$\delta T(x,h) = \lim_{\alpha \to 0} \frac{1}{\alpha} \left(T(x + \alpha h) - T(x) \right)$$

exists, it is called the Gateaux differential of T at x with increment h. If the limit exists for each $h \in X$, T is said to be Gateaux differentiable at x.

Remarks:

(i) It makes sense to consider the above limit only if $x+\alpha h \in D$ for all α sufficiently small. The limit is taken in the usual sense of norm convergence in Y.

(ii) Fixed $x \in D$, h regarded as a variable, the Gateaux differential defines a transformation from X to Y. In the case of T linear, $\delta T(x,h) = T(h)$.

(iii) By far the most frequent application of this definition is in the case where Y is the real line, and hence the trans-

formation reduces to a functional on X. Thus if f is a functional on X, the Gateaux differential of f, if it exists, is

$$\delta f(x,h) = \frac{d}{d\alpha} f(x + \alpha h)\Big|_{\alpha=0}$$

and for each fixed $x \in X$, $\delta f(x,h)$ is a functional w.r.t. the variable $h \in X$.

The Gateaux differential generalizes the concept of the directional derivitive familiar in finite dimensions. The existence of the Gateaux differential is a rather weak requirement, however, since its definition requires no norm on X; hence properties of the Gateaux differential are not easily related to continuity. When X is normed a more satisfactory definition is given by the Frechet differential.

6.1.2 Definition: Let T be a transformation defined on an open domain D in a normed space X and having range in a normed space Y. If for fixed $x \in D$ and $h \in X$ there exists $\delta T(x;h) \in Y$ which is linear and continuous w.r.t. h such that

$$\lim_{\|h\| \to 0} \frac{\|T(x+h) - T(x) - \delta T(x;h)\|}{\|h\|} = 0,$$

Then T is said to be Frechet differentiable at x and $\delta T(x;h)$ is said to be the Frechet differential of T at x with increment h.

Note: We use the same symbol for Frechet and Gateaux differentials, since it is usually apparent from the context which is meant.

Suppose that the transformation T defined on an open domain $D \subset X$ is Frechet differentiable throughout D. At a fixed point $x \in D$

the Frechet differential $\delta T(x;h)$ is by definition of the form $\delta T(x;h) = \langle h, A_x \rangle$ where A_x is a bounded linear operator from X to Y. Thus, as x varies over D, the correspondence $x \to A_x$ defines a transformation D into the normed linear space $B(X,Y)$; this transformation is called the Frechet derivitive T' of T. Thus we have, by definition, $\delta T(x;h) = \langle h, T'(x) \rangle$.

6.1.3 Definition: In the special case where the original transformation is a functional f on X, we have $\delta f(x;h) = \langle h, f'(x) \rangle$ where $f'(x) \in X^*$ for each x. The element $f'(x)$ is called the gradient of f at x, and is denoted $\nabla f(x)$ rather than $f'(x)$.

6.1.4 Lemma: If the transformation T has a Frechet differential, then it is unique.

Proof: Suppose both $\delta T(x;h)$ and $\delta' T(x;h)$ satisfy the requirements to be Frechet differentials. Then

$$||\delta T(x;h) - \delta' T(x;h)|| \leq ||T(x;h) - T(x) - \delta T(x;h)|| + ||T(x;h) - T(x) - \delta' T(x;h)||.$$

or $$||\delta T(x;h) - \delta' T(x;h)|| = o(||h||).$$

Since $\delta T(x;h) - \delta' T(x;h)$ is bounded and linear in h it must be zero.

6.1.5 Definition: Let f be a convex functional defined on a normed space X. An element $x_0^* \in X^*$ is said to be a subgradient of f at x_0 if

$$f(x) - f(x_0) \geq \langle x - x_0, x_0^* \rangle \quad \forall\, x \in X.$$

Denote the set of subgradients of f at x by $\partial f(x)$.

6.1.6 Lemma: If f has a gradient $\nabla f(x_o)$ at x_o in the sense of Gateaux or Frechet, then the subgradient is equal to the gradient.

Proof: Let x_o^* be a subgradient at x_o. Then

$$f(x) - f(x_o) \geq \langle x-x_o, x_o^* \rangle \quad \forall\, x \in X.$$

Let $x = x_o + \lambda h$. Then

$$1/\lambda\,(\, f(x_o+\lambda h) - f(x)\,) \geq \langle y, x_o^* \rangle$$

for every y and $\lambda > 0$. Taking the limit as $\lambda \to 0$ through positive values

$$\delta f(x_o;y) \geq \langle y; x_o^* \rangle$$

and by definition

$$\langle y; \nabla f(x_o) \rangle \geq \langle y; x_o^* \rangle$$

for all y. For this to hold $\nabla f(x_o) = x_o^*$.

Remark: If f has a gradient $\nabla f(x)$ at x in the sense of Gateaux or Frechet, then the two previous lemmas give $\partial f(x) = \{\nabla f(x)\}$.

Now, from the definitions of f^* and $\partial f(x)$, $x^* \in \partial f(x)$ iff

$$f(x) + f^*(x^*) \geq \langle x, x^* \rangle. \tag{6.1}$$

Subgradients of the concave function g have analogous properties (with the defining inequality in (6.1) reversed).

6.1.h Lemma: $\lambda \partial f = \partial(\lambda f)$ for $\lambda > 0$.

Proof: Follows immediately from the definition of subgradient.

6.2 The Kuhn-Tucker Conditions.

To obtain the well known Kuhn-Tucker conditions for the constrained optimization problem that we have been considering, we shall follow the approach used by Rockefellar [24]. We shall make use of Fenchel's Duality Theorem.

6.2.1 Definition: A convex functional h is the indicator of a convex set C in X if

$$h(x) = 0 \quad x \in C$$
$$= \infty \quad x \notin C.$$

6.2.2 Theorem: Let f and g be a proper convex and proper concave functionals respectively. Assume that either f or g is continuous at some point where they are both finite. Then $\bar{x}$ is a point where $f - g$ achieves its minimum over X iff $\partial f(\bar{x})$ and $\partial g(\bar{x})$ have some $\bar{x}^*$ in common. Moreover, such vectors $\bar{x}^*$ are then precisely the points where $g^* - f^*$ achieves its minimum over X^*.

Proof: Let $A = \{ x : f(x) < \infty \}$

$B = \{ x : g(x) > -\infty \}$.

Now, by assumption there exists an $x_0 \in A \cap B$ where f and g are both finite and one is continuous; say g. Thus

$$\mu \leq f(x_0) - g(x_0) < \infty,$$

and the continuity of g means that $x_0 \in \overset{o}{B} \neq \phi$. Therefore, from Theorem 4.1.7 we have that $[g,B]$ has nonempty relative interior.

Now we have

$$f^*(x^*) = \sup_{x \in X} \{ \langle x, x^* \rangle - f(x) \} \geq \langle x, x^* \rangle - f(x) \quad \forall\, x \in X$$

$$g^*(x^*) = \inf_{x \in X} \{ \langle x, x^* \rangle - g(x) \} \leq \langle x, x^* \rangle - g(x) \quad \forall\, x \in X,$$

i.e. $\quad f(x) + f^*(x^*) \geq g(x) + g^*(x^*).$

Now, from (6.1) and its concave analogue, $\bar{x}^* \in \partial f(\bar{x})$, $\bar{x}^* \in \partial g(\bar{x})$

iff $\quad f(\bar{x}) + f^*(\bar{x}^*) \leq \langle \bar{x}, \bar{x}^* \rangle \leq g(\bar{x}) + g^*(\bar{x}^*).$

i.e. $\quad f(\bar{x}) + f^*(\bar{x}^*) = g(\bar{x}) + g^*(\bar{x}^*).$

The conclusion follows from Fenchel's Duality Theorem and the above discussion.

6.2.3 Theorem: Let f_1 and f_2 be proper convex functions on the vector space X. Assume that either f_1 or f_2 is continuous at some point where they are both finite. Then, $\forall\, x \in X$, $\forall\, x^* \in X^*$

(i) $(f_1 + f_2)^*(x^*) = \min \{ f_1^*(x^* - z^*) + f_2^*(z^*) : z^* \in X^* \}$,

(ii) $\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$.

Proof: (i) For any fixed $x^* \in X^*$, define

$$f(x) = f_2(x), \quad g(x) = \langle x, x^* \rangle - f_1(x).$$

Then by definition,

$$f^*(z^*) = f_2^*(z^*), \quad g^*(z^*) = -f_1^*(x^* - z^*).$$

Our hypothesis guarantees, via Fenchel's Theorem, that

$$\inf_{x \in X} \{ f_1(x) + f_2(x) - \langle x, x^* \rangle \} = \max_{z^* \in X^*} \{ -f_1^*(x^* - z^*) - f_2^*(z^*) \}.$$

But the LHS is $-(f_1 + f_2)^*(x^*)$ by definition.
Hence, we have proved the negative equivalent of (i).

(ii) From (i) and (6.1), $x^* \in \partial(f_1 + f_2)$ iff

$$f_1(x) + f_2(x) + f_1^*(x^* - z^*) + f_2^*(z^*) \leq \langle x, x^* \rangle = \langle x, x^* - z^* \rangle + \langle x, z^* \rangle$$

for some $z^* \in X^*$. By definition of f_1^* and f_2^* this is to

$$f_1(x) + f_1^*(x^* - z^*) \leq \langle x, x^* - z^* \rangle,\ f_2(x) + f_2^*(z^*) \leq \langle x, z^* \rangle,$$

and by (6.1) again, $x^* \in \partial(f_1 + f_2)(x)$ iff there exists some $z^* \in X^*$ such that $x^* - z^* \in \partial f_1(x)$ and $z^* \in \partial f_2(x)$.

Let f be a proper convex functional defined on the vector space X. Let $g_1, g_2, \ldots, g_m$ be convex functions defined on the vector space X which are everywhere finite, continuous and Gateaux differentiable. Define

$$D \triangleq \{ x : g_i(x) \leq 0,\ \forall\ i = 1, \ldots, m \}.$$

We assume that the problem satisfies the following constraint qualification which serves to exclude singularities which might otherwise occur on the boundary of the convex domain D.

6.2.4 Definition (Slater's Constraint Qualific tion): A point $x \in D$ satisfies *Slater's constraint qualification* if

$$g_i(x) < 0 \quad \forall\ i = 1, \ldots, m.$$

6.2.5 Theorem (The Kuhn-Tucker Theorem): Suppose that f is finite at some point x_0 satisfying Slater's constraint qualification. Then $\bar{x}$ is a point where f achieves a minimum on D iff there exists Lagrange multiplier $\lambda \in R^m$ which satisfies

$$\text{(i)} \quad -\sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}) \in \partial f(\bar{x}),$$

$$\text{(ii)} \quad \lambda_i g_i(\bar{x}) = 0 \quad \text{for all } i,$$

$$\text{(iii)} \quad \lambda_i \geq 0 \quad \text{for all } i.$$

Proof: Define the convex indicator functions $h_i(x)$ by

$$h_i(x) = 0 \quad \text{if } g_i(x) \leq 0,$$
$$h_i(x) = \infty \quad \text{if } g_i(x) > 0.$$

Define $g(x) = h_1(x) + \ldots + h_m(x)$.

Then minimizing f on D is the same as minimizing $f + g$ on X. i.e., minimize $f - (-g)$ on X, where $-g$ is a proper concave functional finite on D.

Now, at the point x_0 where f is finite and $g_i(x_0) < 0$, $i = 1,\ldots,m$, the functions $g, h_1, \ldots, h_m$ are all finite and continuous.

Thus by Theorem 6.2.2 and the concave analogue of Theorem 6.2.1, f achieves its minimum on D at $\bar{x}$ iff $f(\bar{x})$ contains an element of $-\partial h_1(\bar{x}) - \partial h_2(\bar{x}) \ldots - \partial h_m(\bar{x})$.

Now $x_i^* \in \partial h_i(\bar{x})$ iff

$$h_i(x) - h_i(\bar{x}) \geq \langle x - \bar{x}, x_i^* \rangle \quad \text{for all } x \in X.$$

Clearly, if $g_i(\bar{x}) > 0$, i.e. $\bar{x} \notin D$, then $\partial h_i(\bar{x}) = \phi$.

We have that $\langle \cdot, x_i^* \rangle$ achieves a maximum on $D_i = \{ x : g_i(x) \leq 0 \}$ at $\bar{x}$. Since $\sup g_i < 0$, we have from an elementary argument using the continuity and differentiability of g_i,

$$\partial h_i(\bar{x}) = \{0\} \quad \text{if} \quad g_i(\bar{x}) < 0,$$

$$\partial h_i(\bar{x}) = \{ \lambda_i \nabla g_i(\bar{x}) : \lambda_i \geq 0 \} \quad \text{if} \quad g_i(\bar{x}) = 0.$$

If X is a normed vector space, and f and g are continuous and differentiable in the sense of Frechet, then, providing the other assumptions of the theorem hold, we can write the conditions of the theorem as

$$\nabla f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}) = 0$$

$$\lambda_i g_i(\bar{x}) = 0$$

$$\lambda \geq 0.$$

These results were derived for the case $X = R^n$ by Kuhn and Tucker [17]. It must be emphasized that if the functions are Frechet differentiable, the constraint qualification that we have used in the above theorem is fairly restrictive. The Kuhn-Tucker necessary conditions for optimality can be derived for a more general class of functions (i.e., convexity is not required) satisfying a relaxed constraint qualification (i.e. satisfaction of *Slater's constraint* qualification implies the satisfaction of the relaxed constraint qualification). However, convexity is required for the conditions to be sufficient.

6.3 The Dual Problem.

Suppose that $f(x)$, $g_i(x)$, $i = 1,\ldots,m$ are convex functionals defined over the normed vector space X, which are differentiable in the sense of Frechet. The derivitives, which are elements of X^*, are denoted by $\nabla f(x)$ and $\nabla g_i(x)$, $i = 1,\ldots,m$.

The primal problem which we shall consider is to

$$\text{minimize} \quad f(x)$$
$$\text{subject to} \quad g_i(x) \leq 0, \quad i = 1,\ldots,m.$$

To develop the dual problem to the primal problem above, and to derive the duality theorems, we shall follow the approach taken by Ritter [22]

6.3.1 Theorem: Suppose $f(x)$, $g_i(x)$, $i = 1,\ldots,m$ are convex differentiable functions on X. If there exists an $x \in X$ satisfying Slater's constraint qualification and $\bar{x}$ is an optimal solution of the primal problem, then there exists $\bar{\lambda} \in R^m$ such that $(\bar{x},\bar{\lambda})$ is an optimal solution of the dual problem

$$\text{maximize} \quad f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \tag{6.2}$$

$$\text{subject to} \quad \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) = 0 \tag{6.3}$$

$$\lambda \geq 0. \tag{6.4}$$

Proof: Let $\bar{x}$ be an optimal solution to the primal problem. By the Kuhn-Tucker theorem, there exists $\bar{\lambda} \in R^m$ such that

$$\nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(\bar{x}) = 0$$

$$\bar{\lambda}_i g_i(\bar{x}) = 0, \quad \forall\, i,$$

$$\lambda \geq 0.$$

Thus $(\bar{x},\bar{\lambda})$ is a feasible point for the dual problem, and

$$f(\bar{x}) = f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i g_i(\bar{x}).$$

For any dual feasible point (x,λ) we have from the convexity and differentiability of f that

$$\begin{aligned} f(\bar{x}) - f(x) &\geq \langle \bar{x}-x, \nabla f(x) \rangle \\ &= -\langle \bar{x}-x, \sum_{i=1}^{m} \lambda_i \nabla g_i(x) \rangle \qquad \text{by (6.3)} \\ &\geq -\sum_{i=1}^{m} \lambda_i \langle \bar{x}-x, \nabla g_i(x) \rangle \\ &\geq \sum_{i=1}^{m} \lambda_i \left(g_i(x) - g_i(\bar{x}) \right) \quad \text{because } g_i \text{ are convex, } \forall\, i \\ &\geq \sum_{i=1}^{m} \lambda_i g_i(x) \quad \text{since } -\lambda_i g_i(\bar{x}) \geq 0, \forall\, i. \end{aligned}$$

So we have, for any dual feasible point (x,λ), that

$$f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \geq f(\bar{x}) = f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i g_i(\bar{x}).$$

The above theorem can still be derived even if we were to relax the constraint qualification.

In order to establish the statement that the existence of a solution to the dual problem implies the existence of a solution to the primal problem, the domain (6.3),(6.4) has to satisfy a certain kind of constraint qualification.

6.3.2 Constraint Qualification of the Dual Problem: Let $(\bar{x},\bar{\lambda})$ be an optimal solution of the dual problem. Then, it is supposed that that for each j, $1 \leq j \leq m$, *there exist* differentiable mappings

$h_j(\lambda_j)$, $r_{ij}(\lambda_j)$, $i = 1,\ldots,m$, $i \neq j$, of the interval

$$I = \{ \lambda_j : \bar{\lambda}_j - \epsilon_j \leq \lambda_j \leq \bar{\lambda}_j + \epsilon_j,\ \epsilon_j > 0 \} \cap \{ \lambda_j : \lambda_j \geq 0 \}$$

into X and R^1 respectively, such that

(i) $(h_j(\lambda_j), r_{1j}(\lambda_j),\ldots,r_{j-1,j}(\lambda_j), r_{j+1,j}(\lambda_j),\ldots,r_{mj}(\lambda_j))$ satisfies the conditions (6.3),(6.4) for $\lambda_j \in I$.

(ii) $h_j(\bar{\lambda}_j) = \bar{x}$; $r_{ij}(\bar{\lambda}_j) = \bar{\lambda}_i$, $i = 1,\ldots,m$, $i \neq j$.

(iii) $\sum_{\substack{i=1 \\ i\neq j}}^{m} E_i(\bar{x})\, \nabla_i r_{ij}(\bar{\lambda}_j) = \delta_j E_j(\bar{x})$ for some $\delta_j > -1$.

This constraint qualification is satisfied if there exists a differentiable mapping $x = \xi(\lambda_1,\ldots,\lambda_m)$ of some open neighbourhood N of the point $(\bar{\lambda}_1,\ldots,\bar{\lambda}_m)$ into X such that $\bar{x} = \xi(\bar{\lambda}_1,\ldots,\bar{\lambda}_m)$ *and* $\{\xi(\lambda_1,\ldots,\lambda_m),\lambda_1,\ldots,\lambda_m\}$ *is dual feasible for* $(\lambda_1,\ldots,\lambda_m) \in N \cap \{ \lambda_j : \lambda_j \geq 0,\ j = 1,\ldots,m \}$.

In this case we can choose

$$h_j(\lambda_j) = \xi(\bar{\lambda}_1,\ldots,\bar{\lambda}_{j-1},\lambda_j,\bar{\lambda}_{j+1},\ldots,\bar{\lambda}_m),$$

$$r_{ij}(\lambda_j) = \bar{\lambda}_i \quad \text{for } i = 1,\ldots,m,\ i \neq j.$$

It follows immediately that these mappings satisfy the conditions (i) to (iii).

6.3.3 *Theorem:* Suppose that $f(x)$, $g_i(x)$, $i = 1,\ldots,m$ are convex, differentiable functions on X. If $(\bar{x},\bar{\lambda})$ is a solution to the dual problem defined in equations (6.2) to (6.4), which satisfies the constraint qualification for the dual problem, then $\bar{x}$ *is an optimal solution to the primal* problem, and the extreme values are equal.

Proof: Let j, $1 \le j \le m$ be an arbitrary, but fixed index. Using the mappings of the constraint qualification of the dual problem we define

$$\psi(\lambda_j) = f(h_j(\lambda_j)) + \sum_{\substack{i=1 \\ i \neq j}}^{m} \tau_{ij}(\lambda_j) g_i(h_j(\lambda_j)) + \lambda_j g_j(h_j(\lambda_j)).$$

Then

$$\psi(\lambda_j) - \psi(\bar{\lambda}_j) = \nabla\psi(\bar{\lambda}_j)(\lambda_j - \bar{\lambda}_j) + o(|\lambda_j - \bar{\lambda}_j|)$$

$$= \Big(\sum_{\substack{i=1 \\ i \neq j}}^{m} \nabla\tau_{ij}(\bar{\lambda}_j) g_i(\bar{x}) + g_j(\bar{x}) \Big)(\lambda_j - \bar{\lambda}_j) + o(|\lambda_j - \bar{\lambda}_j|), \text{ by dual}$$

constraint qualification (ii), and because $(h_j(\bar{\lambda}_j), \tau_{ij}(\bar{\lambda}_j), i \neq j, \bar{\lambda}_j)$ is dual feasible,

$$= (\lambda_j + 1) g_j(\bar{x})(\lambda_j - \bar{\lambda}_j) + o(|\lambda_j - \bar{\lambda}_j|), \text{ by dual}$$

constraint qualification (ii).

Since $(\bar{x}, \bar{\lambda})$ is a maximum, it follows that

$$0 \ge \psi(\lambda_j) - \psi(\bar{\lambda}_j) = (\lambda_j + 1) g_j(\bar{x})(\lambda_j - \bar{\lambda}_j) + o(|\lambda_j - \bar{\lambda}_j|).$$

So, if $\bar{\lambda}_j > 0$, $g_j(\bar{x}) \ge 0$

$\bar{\lambda}_j > 0$, $g_j(\bar{x}) = 0$, for $1 \le j \le m$.

Therefore, $\bar{x}$ is a feasible point for the primal problem, and

$$f(\bar{x}) = f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i g_i(\bar{x}).$$

For any primal feasible point x, we have, from the convexity and differentiability of f, that

$$f(x) - f(\bar{x}) \geq \langle x-\bar{x}, \nabla f(\bar{x})\rangle$$

$$= -\langle x-\bar{x}, \sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(\bar{x})\rangle$$

$$= -\sum_{i=1}^{m} \bar{\lambda}_i \langle x-\bar{x}, \nabla g_i(\bar{x})\rangle$$

$$\geq \sum_{i=1}^{m} \bar{\lambda}_i (g_i(\bar{x}) - g_i(x)) \quad \text{by convexity}$$

$$\geq \sum_{i=1}^{m} \bar{\lambda}_i g_i(\bar{x}).$$

Hence, for any primal feasible point x, we have that

$$f(x) \geq f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i g_i(\bar{x}) = f(\bar{x}).$$

The duality theorem evolved from Dorn [2], Wolf [29], Hansen [8], Huard [10], and Mangasarian [19].

6.3.4 Example: Consider a primal problem which has the following special form (see Example 5.3.3):

$$\text{minimize} \quad \tfrac{1}{2}(Ax)x - f(x) \tag{6.5}$$

$$\text{subject to} \quad f_i(x) \leq \alpha_i, \quad i = 1,\ldots,m,$$

where f and f_i are continuous linear functionals, α_i are real constants and A is a bounded linear operator mapping X onto X^*. Suppose that A has the properties

(i) A^{-1} exists and is bounded,

(ii) $(Ax_1)x_2 = (Ax_2)x_1$ for any pair $x_1, x_2 \in X$,

(iii) $(Ax)x \geq 0$ for any $x \in X$.

The last property asserts that (6.5) is a convex function. This problem is of particular interest because it turns out that the dual problem is equivalent to an m-dimensional quadratic

maximization problem in a Euclidean space. Thus, the solution of the primal problem may be obtained by the computational solution of a finite-dimensional maximization problem.

The dual problem is as follows:

$$\text{maximize} \quad \tfrac{1}{2}(Ax)x - f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) - \sum_{i=1}^{m} \lambda_i \alpha_i \qquad (6.6)$$

$$\text{subject to} \quad Ax + \sum_{i=1}^{m} \lambda_i f_i = f, \quad \lambda_i \geq 0, \; i = 1,\ldots,m. \qquad (6.7)$$

The expression (6.7) follows from (6.3) since $f_i(x)$ and $f(x)$ are linear, and therefore $f_i(x) = f$, $f(x) = f$. Because of (6.7) the objective function (6.6) is equivalent to

$$-\tfrac{1}{2}(Ax)x - \sum_{i=1}^{m} \lambda_i \alpha_i. \qquad (6.8)$$

Since A^{-1} exists, (6.7) yields

$$x = A^{-1} - \sum_{i=1}^{m} \lambda_i A^{-1} f_i. \qquad (6.9)$$

Substituting this into (6.8), and using the fact that property (ii) of A implies that $f_i(A^{-1}f_j) = f_j(A^{-1}f_i)$, we have that the dual problem can be written as

$$\text{maximize} \quad -\sum_{i=1}^{m} c_i\lambda_i - \tfrac{1}{2}\sum_{i,j=1}^{m} c_{ij}\lambda_i\lambda_j - \tfrac{1}{2}f(A^{-1}f)$$

$$\text{subject to} \quad \lambda_i \geq 0, \quad i = 1,\ldots,m,$$

where

$$c_{ij} = c_{ji} = f_i(A^{-1}f_j),$$
$$c_i = \alpha_i - f(A^{-1}f_i).$$

Let $\bar{\lambda} \in R^m$ be the optimal solution. (6.9) defines a mapping whose existence was shown to be sufficient for the dual constraint

qualification of the dual problem to hold. Therefore, Theorem 6.3.3 *applies and*

$$\bar{x} = A^{-1}f - \sum_{i=1}^{m} \bar{\lambda}_i A^{-1}f_i$$

is an optimal solution of the primal problem.

7. CONCLUSION.

Duality theory is of considerable interest, and has many applications in the fields of constrained optimization and control theory. For certain problems, it is often a good deal easier to solve the dual problem than it is to solve the primal problem, and , providing certain assumptions are satisfied, the solutions are equal. In the last section it was shown that if the primal problem has a quadratic objective function and linear constraints, its dual problem can be transformed into a maximization problem in a finite dimensional Euclidean space. The duality theorem then gives directly, the solution to the original primal problem. This case is therefore of considerable interest because of its potential application to continuous programming problems and problems of optimal control theory.

Duality theorems supply the basis for a number of computational procedures. These procedures, like the duality theorems themselves, often can be *justified only for convex problems*, but in many cases the basic idea can be modified so as to be effective for other problems. The penalty function method emerges as a particularly nice implementation of the primal-dual philosophy.

In this essay we have attempted to provide a rigorous mathematical treatment of duality theory for convex constrained optimization. Much of the theory developed in this essay was obtained from Luenberger [18]. A great deal of the material that we have discussed is also covered in the book by Stoer and Witzgall [27]. In addition, it includes a very general extension of Fenchel's Duality Theorem, from which several other duality theorems, including Rockefellar's extension of Fenchel's Theorem, are deduced.

REFERENCES.

1. Arrow, K.J., L. Hurwitz, and H. Uzawa, Studies in Linear and Non-Linear Programming, Chapter 4, 'Programming in Linear Spaces", Stanford Univ. Press, Stanford, Calif., 38 - 102, 1964.

2. Born, W.S., A Duality Theorem for Convex Programs, IBM J. of Research and Development, 4, 407 - 413, 1960.

3. Douglas, R.G., Banach Algebra Techniques in Operator Theory, Academic Press, New York, 1972.

4. Dunford, N., and J.T. Schwartz, Linear Operators, Part I, Interscience, New York, 1958.

5. Fenchel, W., Convex Cones, Sets and Functions, Lecture Notes, Dept. of Math., Princeton Univ., 1953.

6. Gould, F.J., Extensions of Lagrange Multipliers in Nonlinear Programming, SIAM J. Appl. Math., Vol. 17, No. 6, 1969.

7. Halmos, P.R., Finite-Dimensional Vector Spaces, D.Van Nostrand Co., Princeton, N.J., 1958.

8. Hanson, M.A., A Duality Theorem in Nonlinear Programming with Nonlinear Constraints, Austral. J. Statistics, 3, 64 - 72, 1961.

9. Hille, E., and R.S. Phillips, Functional Analysis and Semi-Groups, Amer. Math. Soc. Colloquium Pub., 31, New York, 1957.

10. Huard, P., Dual Programs, IBM. J. Research and Dev., 6, 137 - 139, 1962.

11. Hurwicz, L., Programming in Linear Spaces, in K.J. Arrow, L. Hurwicz, and H. Uzawa, 'Studies in Linear and Nonlinear Programming', Stanford Univ. Press, 1958.

12. John, F., Extremum Problems with Inequalities as Subsidiary Conditions, Studies and Essays Presented to R. Courant on his 60th birthday, 187 - 204, Interscience, New York, 1948.

13. Karlin, S., Operator Treatment of Minimax Principle, in Contr. to the Theory of Games, ed. by H.W. Kuhn and A.W. Tucker, 133 - 154, Princeton Univ. Press, 1950.

14. Karlin, S., Mathematical Methods and Theory in Games, Programming and Economics, Vol I, Addison-Wesley, Reading Mass., 1959.

15. Karlin, S., Mathematical Methods and Theory in Games, Programming and Economics, Vol II, Addison-Wesley, Reading Mass., 1959.

16. Kelley, J.L., and I. Namioka, Linear Topological Spaces, Van Nostrand, Princeton, N.J., 1963.

17. Kuhn, H.W., and A.W. Tucker, Nonlinear Programming, Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability, 481 - 492, Univ. of Calif. Press, Berkeley, 1961.

8. Hanson, M.A., A Duality Theorem in Nonlinear Programming with Nonlinear Constraints, Austral. J. Statistics, 3, 64 - 72, 1961.

9. Hille, E., and R.S. Phillips, Functional Analysis and Semi-Groups, Amer. Math. Soc. Colloquium Pub., 31, New York, 1957.

10. Huard, P., Dual Programs, IBM. J. Research and Dev., 6, 137 - 139, 1962.

11. Hurwicz, L., Programming in Linear Spaces, in K.J. Arrow, L. Hurwicz, and H. Uzawa,'Studies in Linear and Nonlinear Programming', Stanford Univ. Press, 1958.

12. John, F., Extremum Problems with Inequalities as Subsidiary Conditions, Studies and Essays Presented to R. Courant on his 60th birthday, 187 - 204, Interscience, New York, 1948.

13. Karlin, S., Operator Treatment of Minimax Principle, in Contr. to the Theory of Games, ed. by H.W. Kuhn and A.W. Tucker, 133 - 154, Princeton Univ. Press, 1950.

14. Karlin, S., Mathematical Methods and Theory in Games, Programming and Economics, Vol I, Addison-Wesley, Reading Mass., 1959.

15. Karlin, S., Mathematical Methods and Theory in Games, Programming and Economics, Vol II, Addison-Wesley, Reading Mass., 1959.

16. Kelley, J.L., and I. Namioka, Linear Topological Spaces, Van Nostrand, Princeton, N.J., 1963.

17. Kuhn, H.W., and A.W. Tucker, Nonlinear Programming, Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability, 481 - 492, Univ. of Calif. Press, Berkeley, 1961.

18. Luenberger, D.G., Optimization by Vector Space Methods, John Wiley, 1969.

19. Mangasarian, O.L., Duality in Nonlinear Programming, Quart. Appl. Math., 20, 300 - 302, 1962.

20. Rådström, H., Polar Reciprocity, Seminar on Convex Sets, 27 - 29, Inst. for Advanced Study, Princeton, N.J., 1949 - 1950.

21. Riesz, F., and B. Sz-Nagy, Functional Analysis, Frederick Ungar, New York, 1955.

22. Ritter, K., Duality for Nonlinear Programming in a Banach Space, SIAM J. Appl. Math., 15(2), 294 - 302, 1967.

23. Rockafellar, R.T., Duality Theorems for Convex Functions, Amer. Math. Soc. Bull., 70, 189 - 192, 1964.

24. Rockafellar, R.T., Extensions of Fenchel's Duality Theorem for Convex Functions, Duke Math. J., 33, 81 - 90, 1966.

25. Simmons, G.F., Introduction to Topology and Modern Analysis, McGraw-Hill, New York, 1963.

26. Slater, M., Lagrange Multipliers Revisited: A Contribution to Nonlinear Programming, Cowles Commission Discussion Paper, Math. 403, Nov. 1950.

27. Stoer, J., and C. Witzgall, *Convexity and Optimization in Finite* Dimension, Springer-Verlag, New York, 1970.

28. Whinston, A., Some Applications of the Conjugate Function Theory to Duality, in 'Nonlinear Programming' ed. by J. Abadie, Interscience, New York, 1967.

29. Wolf, P., A Duality Theorem for Nonlinear Programming, Quart. Appl. Math., 19, 239 - 244, 1961.

THE MAXIMUM PRINCIPLE.

Kendal Clive Jordi

THE MAXIMUM PRINCIPLE.

Kendal Clive Jordi

An essay submitted to the Faculty of Science, in partial fulfilment of the requirements for the degree of Master of Science.

University of the Witwatersrand,
Johannesburg.
1975.

Abstract.

The basic optimal control problem for a system described by an ordinary differential equation is defined. An alternative formulation to the basic optimal control problem, which is required to establish Pontryagin's Maximum Principle, is developed. An historical development of the necessary conditions for the the optimality of a control system is given. The proofs of the Maximum Principle by Pontryagin, Boltyanskii, Gamkrelidze and Mischenko, Halkin, and Bryant and Mayne are then discussed in some detail.

1. INTRODUCTION.

The modern theory of control is not only of interest to mathematicians, but has also attracted attention in many other diverse fields of interest.

In control theory one is interested in a process, i.e., some action or motion which can be influenced by certain controls or policies. To analyse the process it is necessary to formulate a structure called the dynamics of the process, or a law which governs the change of state. This provides the means whereby one can determine the state $x(t)$ for $t > t_0$ provided the state is known for $t \leq t_0$. It is usually required that some goal be achieved by the process through a properly applied control policy, i.e., there is some objective. This is usually specified as the acquisition of some desired state target for the process. One question which arises naturally is whether or not means for influencing the process are sufficiently strong to allow the achievement of a specified objective. If such means exist, then we have a properly formulated control structure.

In control problems there are generally several ways in which the objective for a process may be accomplished. Within the set of possibilities, taking into account imposed constraints, it may be desirable to systematically choose the 'best' approach with respect to some performance criterion. If with respect to some performance criterion one seeks, in the set of all policies for achieving an objective, the one that is best, then the formulation is an optimal control problem. The control policy, if it

exists, which solves the optimal control problem is known as the optimal control (policy) for the problem.

The conditions which a control is required to satisfy in order to be optimal are of considerable interest. These conditions can be derived for a certain class of problems using a classical variational approach. However, as the problem becomes increasingly complex, this approach becomes increasingly difficult to apply, and in certain cases cannot be used at all.

In this essay we are interested in the necessary conditions formulated by Pontryagin and his associates, and called Pontryagin's Maximum Principle. These conditions are, in general, not sufficient, and are local in nature. An outline of the historical development of the various statements and proofs of the necessary conditions is given. In the rest of the essay we are concerned with contrasting the proof of the Maximum Principle by Pontryagin et al [52], which is based on the construction of the convex cones of McShane, with the proofs of Halkin, [25], [27], and Bryant and Mayne [13], which are based on the construction of convex sets, and are generalizations of the proof by La Salle [38] for the linear system.

2. THE CONTROL PROBLEM.

Consider the system defined by the ordinary differential equation

$$\dot{x}(t) = f(x(t),u(t),t) \qquad (E)$$

where $f: R^n \times \Omega \times [t_o,t_f] \to R^n$, and $\Omega \subset R^m$ denotes the control space.

U is said to be the class of admissible control functions if (i) each $u \in U$, $u : [t_o,t_f] \to \Omega$, is a bounded, measurable (in the sense of Lebesque) function, and

(ii) U is closed under juxtaposition and translation.

Assume that for each $u \in U$ and each $x_o \in R^n$ there exists one *absolutely continuous function*

$$x(t,u(\cdot)) = x(t;t_o,x_o,u(\cdot))$$

such that

$$\dot{x}(t,u(t)) = f(x(t,u(t)),u(t),t) \quad \text{a.e.} \quad t \in [t_o,t_f] \qquad (2.1)$$

and

$$x(t_o,u) = x_o.$$

If at least one solution $x(\cdot,u)$ of (2.1) exists on all of $[t_o,t_f]$ then we call $u(\cdot)$ a control and $x(\cdot,u)$ a response.

Define

$$S_* \triangleq \{ S \subset R^n : S \text{ is closed} \}.$$

Let $-\infty < \tau_o < \tau_1 < \infty$ and

$$G : [\tau_o,\tau_1] \to S_*.$$

Then a control problem consists of the following five items:

an ordinary differential equation (E),

a control region Ω,

an admissible control class U,

an initial point x_o,

and a target $G(\cdot)$ on $[\tau_o,\tau_1]$.

For mathematical reasons we consider measurable controls. However, measurable controls in general are difficult if not impossible to implement physically. Thus there may exist some measurable control steering some point to the required target, but not a physically 'reasonable' control. For this reason it is often of interest to consider restricting U (e.g. to piecewise continuous functions) or restricting Ω (e.g. to a finite number of points).

The problem of existence and uniqueness of solutions to (2.1) for a given bounded, measurable function $u(\cdot)$ is discussed in Coddington and Levinson [15]. If the function f satisfies the Caratheodary conditions on $[t_o,t_f]$ then for each given initial condition $x(t_o,u) = x_o$, there exists a solution $x(\cdot,u)$ of (2.1) on $[t_o,t_f]$. In the usual formulation of the control problem it is required that (E) has uniqueness. If it is assumed that f is either linear in x, or continuously differentiable in x, or locally Lipschitz, then each of these assumptions implies that, provided (E) has a solution, that solution is unique.

The Optimal Control Problem.

For each $u \in U$ and unique response $x(\cdot,u)$ to $u(\cdot)$, define

$$J = J[u(\cdot)] = \int_{t_o}^{t_f} f^o(x(t,u),u(t),t)dt$$

where $f^o: R^n \times \Omega \times [t_o,t_f] \to R$.

Then an optimal control problem consists of a control problem together with a cost functional J.

Given an optimal control problem, let $\Delta \triangleq \Delta((B),\Omega,U,x_o,G)$ be the set of all controls $u \in U$, defined on all subintervals $[t_o,t_f]$ of $[\tau_o,\tau_1]$ such that $u(\cdot)$ 'steers x_o to the target', i.e., such that

$$x(t_f;t_o,x_o,u(\cdot)) \in G(t_f).$$

Then a control $v \in \Delta$ is optimal (with respect to Δ) if

$$J[v(\cdot)] \leq J[u(\cdot)] \quad \forall\, u \in \Delta.$$

Two of the major problems of control theory can now be posed :

(i) Existence : When does Δ contain an optimal control?

(ii) Necessity : What properties must an optimal control have?

The existence of optimal controls occupies a special place in the literature. For a general discussion the reader should consult Lee and Markus [41], for example. As in the calculus of variations, the existence of a control $u \in \Delta$ which minimizes $J[u(\cdot)]$ implies the validity of certain equations - necessary conditions.

An Alternative Formulation of the Optimal Control Problem.

The formulation of the optimal control problem given above is not convenient for the derivation of the Maximum Principle that we shall present in later sections. We give an equivalent formulation.

Consider the non-autonomous (time dependent) system

$$\dot{x}(t,u) = f(x(t,u),u(t),t), \quad x(t_o) = x_o,$$

with cost functional

$$J[u] = \int_{t_o}^{t_f} f^o(x(t,u),u(t),t)dt.$$

Define an additional phase co-ordinate x^o by

$$x^o(t) = \int_{t_o}^{t} f^o(x(\tau,u),u(\tau),\tau)d\tau, \quad t\in[t_o,t_f].$$

We observe that

$$x^o(t_o,u) = 0, \quad x^o(t_f,u) = J[u(t_f)],$$

$$\dot{x}^o(t,u) = f^o(x(t,u),u(t),t), \quad t\in[t_o,t_f].$$

Consider the vectors $\bar{x}$, $\bar{f}$ where

$$\bar{x} = (x^o,x), \quad \bar{f} = (f^o,f).$$

Then the above equations can be combined to form the system

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

The initial condition for the system is given by

$$\bar{x}(t_o,u) = \bar{x}_o = (0,x_o)$$

while the final state is given by

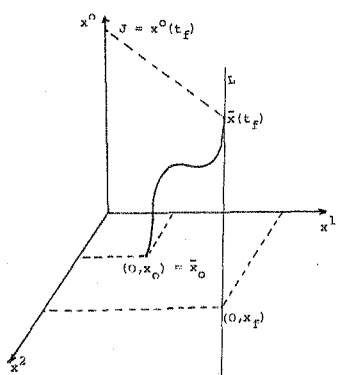
$$\bar{x}(t_f,u) = (J,x_f).$$

We are now required to find the admissible control $u \in \Lambda$, if it exists, which transfers the given initial point $\bar{x}_o$ to the point $\bar{x}_f$ so that the intersection of the trajectory $\bar{x}(\cdot)$ with the line L defined by

$$L \triangleq \{ (\xi,x_f) : \xi \in R \}$$

has the smallest x^o co-ordinate. See Fig. 2.1. It is clear that the autonomous optimal control problem can be formulated in the same way.

Figure 2.1



The Reachable Set.

We shall often consider the control problem with either t_o or t_f fixed in advance. For a fixed t_o, fixed initial condition x_o and for each $t > t_o$, we define the reachable set at time t by

$$W(t) \triangleq \{ x(t;x_o,t_o,u(\cdot)) : u \in U, \ t_o \text{ fixed} \}.$$

The reachable cone, the set of all possible points lying on the graph of all responses to controls in U for fixed t_o, is defined by

$$R.C. \triangleq \{ (t,W(t)) : t \geq t_o \}.$$

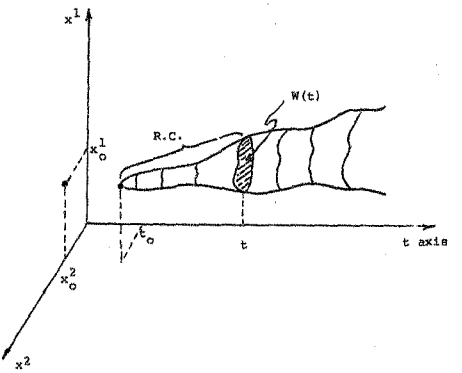
If $u \in U$ and $t_o < \tau < t$, assume that the restriction of $u(\cdot)$ to $[t_o,\tau]$ belongs to U. Hence, if $x(t,u) \in W(t)$, it follows that $x(\tau,u) \in W(\tau)$ for each τ, $t_o < \tau < t$. Hence the reachable sets $W(t)$ are 'slices' or 'cross'sections' of the reachable cone, as shown in Fig. 2.2.

Closure and convexity of the sets R.C. and $W(\tau)$, $t_o < \tau < t$, may be used to establish the existence of solutions to the optimal control problem. See Lee and Markus [40], Roxin [54] and Neustadt [49].

The Principle of Optimal Evolution: If $v(\cdot)$ is an optimal control function, then for every $t \in [t_o,t_f]$ the state $\bar{x}(t,v)$ belongs to the boundary of the set $W(t)$. i.e., the response $\bar{x}(\cdot,v)$ lies on the boundary BRC of R.C. on $[t_o,t_f]$.

This principle was developed independently and in slightly different directions by Halkin [21] and Roxin [53].

Figure 2.2



3. THE HISTORICAL DEVELOPMENT OF THE MAXIMUM PRINCIPLE.

One of the fundamental problems of control theory is that of transforming a system from one state to another in minimum time or at minimum cost of resources in general. There are many variants of this fundamental problem and many powerful approaches are now available for solving it. In its classical form, the question can be treated by means of the calculus of variations. It can also be handled quite directly by means of dynamic programming (Bellman [7]).

The school of optimal control whose approach has been largely mathematical has its inception in a 1952 Ph.D. Thesis by D.W. Bushaw, [14]. Bushaw restricted himself to one controlled variable. He confined his study to bang-bang systems only, and for the special systems studied, proved the existence of the time optimal control for bang-bang systems. In 1953 A. Feldbaum, [17], made the classical formulation of the time optimal control problem for systems subject to saturation. In the same year La Salle, [35], made the observation that the best of all bang-bang systems, if it exists, is then the best of all systems operating from the same power source.

Bellman, Glicksberg and Gross, [8], considered the system

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{3.1}$$

where Ω is the cube $|u^i| \leq 1$, and restricted themselves to the problem of starting at x_o and reaching the origin in minimum time. The $n \times n$ constant matrix A was assumed to have negative real parts, B was assumed to be a constant nonsingular

$n \times n$ matrix. They proved the existence of an optimal steering function. The form for an optimal steering function is given in the proof. However, the form given for an optimal steering function does not imply that there is a bang-bang optimal steering function. Gamkrelidze [18], [19] considered the same problem, but he removed the restriction that B be nonsingular. He proved the existence and uniqueness of an optimal control for 'normal' systems. The form of the optimal control was the same as that given by Bellman, Glicksberg and Gross, and hence, in that case, it could be concluded that the optimal control is bang-bang.

In a series of papers beginning in 1957, Krasovskii [32], [33], [34] treated the more general control problem of hitting a moving particle and studied the more general control system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + f(t) \tag{3.2}$$

where Ω is the cube $|u^i| \leq 1$, $A(t)$ is an $n \times n$ matrix function, $B(t)$ is an $n \times m$ matrix function, x and f are n-dimensional vector functions. Using results of Krein on the L-problem in abstract spaces, he proved the existence of an optimal steering function for 'proper' control systems. La Salle [36] remarks that if Krein's results are to be used without modification, the restriction to proper control systems seems to be necessary.

La Salle [36], [37] considered the general problem (3.2) stated above. He extended the results of [35], and at the same time established the bang-bang principle for all control systems where the controlled elements are linear. For the control system

(3.2), the state $x(t,u)$ of the system at time t is given by

$$x(t,u) = \Phi(t)x_o + \Phi(t)\int_{t_o}^{t} Y(\tau)u(\tau)d\tau + \Phi(t)\int_{t_o}^{t} \Phi^{-1}(\tau)f(\tau)d\tau$$

where $t\in[t_o,t_f]$, Φ is the fundamental matrix solution to the equation $\dot{\Phi}(t) = A(t)\Phi(t)$, and $Y(t) = \Phi^{-1}(t)B(t)$. It is required at some time t to have $x(t) = z(t)$. i.e. to have

$$w(t) = \int_{t_o}^{t} \Phi^{-1}(\tau)u(\tau)d\tau$$

where

$$w(t) = \Phi^{-1}(t)z(t) - x_o - \int_{t_o}^{t} \Phi^{-1}(\tau)f(\tau)d\tau.$$

$A(t)$, $B(t)$ and $f(t)$ are assumed to be continuous for $t_o \le t < \infty$.

Ω is the set of all m-dimensional vector functions, measurable on $[t_o,t_f]$ with $|u^i(t)| \le 1$. Let Ω^o be the subset of functions in Ω with $|u^i(t)| = 1$. Define

$$K(t) \triangleq \{ \int_{t_o}^{t} Y(\tau)u(\tau)d\tau : u \in U \},$$

$$K^o(t) \triangleq \{ \int_{t_o}^{t} Y(\tau)u^o(\tau)d\tau : u^o \in U^o \}, \quad t\in[t_o,t_f],$$

where

$$U \triangleq \{ u : [t_o,t_f] \to \Omega, \text{ where } u \text{ is admissible } \},$$

$$U^o \triangleq \{ u : [t_o,t_f] \to \Omega^o, \text{ where } u \text{ is admissible } \}.$$

The set $K(t)$ is related to the reachable set $W(t)$, $t\in[t_o,t_f]$. We see that the state $z(t)$ can be reached in time t using the allowable steering if and only if $w(t) \in K^o(t)$. It can be proved

$$K(t) = K^o(t)$$

and $K(t)$ is closed and convex. This means that anything can be

done by an admissible steering function can also be done by a bang-bang function. Hence, for the system described by (3.2), if there is a steering function u in Ω such that $x(t,u) = z(t)$ for some $t > t_o$, then there is an optimal steering function in Ω^o. Moreover, all optimal steering functions v are of the form

$$v(t) = \operatorname{sgn}[\eta X(t)], \quad t \in [t_o, t_f], \tag{3.3}$$

where η is some n-dimensional vector, and $a = \operatorname{sgn} b$ means that $a^i = 1$ when $b^i > 0$, and $a^i = -1$ when $b^i < 0$.

A fairly straightforward proof of the above result is given by La Salle [37], [38] using a theorem on the range of an abstract vector measure, due originally to Liapunov [42] and subsequently simplified by Halmos [29] and extended by Blackwell [10]. Using a generalization of the theorem of Liapunov, given in [22], Halkin [23] extended La Salle's result to $A(\cdot)$, $B(\cdot)$ piecewise analytic. For a good text on the material covered so far, the reader should consult Straus [90].

At the 1958 International Congress of Mathematicians in Edinburgh, L.S. Pontryagin announced the 'Maximum Principle',[50], which was the start of an even more general theory. This principle, leading to the solution of the general problem of finding a control process, optimum for rapid action, was hypothesized in 1956 by Pontryagin on the basis of the results of work performed by himself, V.G. Boltyanskii and R.V. Gamkrelidze. The Maximum Principle was verified at first for individual types of systems, and in particular, was proved in [19] for linear

systems. Boltyanskii fully proved that the Maximum Principle was a necessary condition for optimality in relation to rapid action. Gamkrelidze [18], [19] proved theorems of existence and uniqueness and examined the problem of synthesizing controls for linear systems, optimum for rapid action. The Maximum Principle was then extended to the general case of minimizing an arbitrary functional of the integral function of variable systems.

In [51] there is a detailed presentation of the basic results obtained by Pontryagin and his associates at that stage. The control vector u is required to lie in a closed set, the terminal state x_f is a prescribed vector, and the terminal time is arbitrary. Pontryagin applied the Maximum Principle to the linear controlled system of the form (3,1), where Ω is a general convex polyhedron having the origin as an interior point. In a series of articles, Rozonoer [56] extended the Maximum Principle to problems in which the terminal time t_f is fixed and x_f is free. Rozonoer also established the connection between the Maximum Principle and the method of dynamic programming, and formulated and proved the principle for optimum processes in linear discrete-time systems, analogous to the Maximum Principle. A detailed account of the results of the various papers written by Pontryagin, Boltyanskii, Gamkrelidze and Mischenko is presented in [12]. The authors' results were finally collated in [52]. The Maximum Principle is proved for autonomous systems. Certain classes of non-autonomous systems are transformed into autonomous systems, where it is required that the functions be continuously differentiable w.r.t. t. The necessary conditions are derived for the problem where the trajectory must remain in some closed domain

of the phase space. A detailed analysis of linear, autonomous time-optimal problems is given. It is also shown that the basic necessary conditions of the classical variational calculus with ordinary derivitives follow from the Maximum Principle.

The proof of Pontryagin and his associates is based on the construction of the special variations which were developed initially by Mc Shane [44]. These special variations lead to the constuction of some convex cones. While Mc Shane uses both strong and weak variations to construct the cone, Pontryagin et al use only strong variations. A heuristic outline of the proof of the Maximum Principle given in [52] is presented in the next section.

Boltyanskii [11] uses piecewise continuous functions to give a simpler but similar proof to that in [52]. However, it only applies to problems for which

$$f^{o}(x,u) > 0.$$

Diliberto [16] generates the cones of Pontryagin by using vectors with not only positive, but also negative coefficients. He gives a nonlinear generalization of the bang-bang principle.

The problem with performance index

$$J = \int_{t_o}^{t_f} f^{o}(x,u,t)dt + g(x(t_f),t_f)$$

where $g(\cdot,\cdot)$ is defined on a set of terminal states, was considered by Berkovitz, [9], where a set of necessary conditions applicable to a wider class of such problems is derived. This result is based on Mc Shane's proof of the multiplier rule for the abnormal case of the problem of Bolza. Using a device of

Valentine [58] he obtains an equivalent problem of Bolza. Necessary conditions from the Bolza problem are translated into necessary conditions for the optimal control problem. However, this proof presupposes an extensive knowledge of the calculus of variations.

It should be noted that a similar Maximum Principle was also obtained by Hestanes [30], using a generalized Lagrange multiplier rule. The development given is also an outgrowth of the method introduced by Mc Shane and later modified and extended to optimal control theory by Pontryagin et al.

A Maximum Principle for the problem with performance functional

$$J = g(x_g(t_f), x(t_f))$$

where g is a scalar, and

$$x_g(t) = \int_{t_o}^{t} f_g(x(\tau), u(\tau), \tau)\, d\tau$$

while

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

was obtained by A. Wierzbicki [63]. This is done by modifying the proof of Pontryagin and assuming that g is semiconvex. Since a function can be semiconvex even if it is not continuous, this represents a considerable generalization of the previous results.

Warga [61] suggests that a system that does not satisfy the property that

$$\bar{f}(x,\Omega,t) = \{ \bar{f}(x,u,t) : u \in \Omega \}$$

is convex, should be 'relaxed', by enlarging the set of allowed values of $\dot{\bar{x}}(t) = (\dot{x}^o, \dot{x})$ from $\bar{f}(x,\Omega,t)$ to the closure of the convex hull of $\bar{f}(x,\Omega,t)$. He then shows that the solutions of the relaxed problem can be uniformly approximated by the solutions of the original problem. These relaxed curves must satisfy necessary conditions which are generalizations of the Maximum Principle. A recently published book by the same author, [62], contains Maximum Principles for relaxed and ordinary solutions of problems defined by differential and functional integral equations, these being a considerable extension of the results in [61]. Mc Shane [66] extends the results of Warga [61] to apply to unbounded controls, and also to time and state dependent control regions $\Omega = \Omega(x,t)$. An article by Baum and Cesari, [6], gives a simplified proof of Pontryagin's necessary conditions by imposing some additional convexity constraints on the sets $\bar{f}(x,\Omega(t),t)$.

An attempt at the unification of the intimitely connected concepts of Young's generalized curves, [64], and relaxed variational problems is given by Gamkrelidze in [20]. This is done by introducing the related concepts of quasi-convexity and 'chattering' or 'generalized' controls. A generalized control assigns to almost every t, a probability measure $\mu(\cdot,t)$, defined on the Lebesque subsets of Ω. If Ω is compact, the generalized controls may be considered elements of the dual space $\beta_0^*(\Omega)$. These reduce to the conventional control u if the measure is wholly concentrated at u. Using this, Gamkrelidze derives an integral form of Pontryagin's Maximum Principle, which includes the Maximum Principle for conventional controls as a special case. If only convexity is required, then the intregral form of

the Maximum Principle is not directly applicable to conventional controls, but can be replaced by a version of the Maximum Principle with generalized controls, [64].†

Another recent approach is the use of the 'epsilon' technique in optimal control. The epsilon technique is at once a 'relaxation' method and a 'penalty function' method. Consider

$$\int_{t_o}^{t_f} f^o(x(t),u(t),t)\,dt$$

subject to the state dynamics

$$\dot{x}(t) = f(x(t),u(t),t), \quad x(t_o) = x_o$$

where $x(\cdot)$ may be subject to terminal conditions and $u(\cdot)$ is constrained. The epsilon technique is to replace the problem by a sequence of nondynamic problems:

For each $\varepsilon > 0$ minimize

$$\frac{1}{2\varepsilon}\int_{t_o}^{t_f} \| \dot{x}(t) - f(x(t),u(t),t) \|\,dt + \int_{t_o}^{t_f} f^o(x(t),u(t),t)\,dt$$

over the class of state functions $x(\cdot)$, absolutely continuous and satisfying the stipulated initial and final conditions, and over the class of control functions $u(\cdot)$ constrained as in the original problem. For each fixed epsilon, this 'free' (free of the dynamic constraint) problem is simpler theoretically and faster computationally. As epsilon goes to zero, we expect the solution, if it exists, to approximate the solution to the original problem. This is in fact what happens, [2]. The relaxed problem does not always have a solution. In fact, if the optimal control problem does not have a solution, the epsilon problem

† Much of the material covered so far in this section on the generalizations and extensions of Pontryagin's Maximum Principle has been obtained from Horn [31].

does not have a solution either. The epsilon technique has provided a constructive derivation of the Maximum Principle, [2], [3], [5], and in particular, a constructive method for obtaining the Lagrange multipliers, whose existence alone is usually proved.

Mersky [47] has presented an extension of the epsilon technique of [5]. He outlines the results obtained in [48]. He works in the class of generalized controls in the sense of Young [64] and Mc Shane [45]. Ω is assumed to be a compact set in R^m. So a weak* topology is used for the topology of U, the class of generalized controls generated by the compact set Ω. Mersky replaces the constrained problem (P) with the epsilon problem (P_ε), and proves, working in the class of generalized controls, that there exists a solution $(x_\varepsilon(\cdot), v_\varepsilon(\cdot))$ to the epsilon problem. Also, this solution satisfies the 'ε-Maximum Principle'. Then, as $\varepsilon \to 0$, $x_\varepsilon(\cdot)$ converges uniformly to $x(\cdot,v)$, v_ε converges weak* to v, where $(x(\cdot,v),v)$ is a solution to the problem (P). Furthermore, the Maximum Principle for the problem (P) can be established, provided that a certain matrix, arising out of the constraint equations, has full rank along the optimal solution.

A completely different approach to deriving the Maximum Principle has as its basis, the Principle of Optimal Evolution. This has been used by Roxin [55] to look at Pontryagin's necessary conditions from a geometric point of view. Halkin has used the Principle of Optimal Evolution to extend the proof of La Salle [37] to the general optimal control problem. Halkin does not require that non-autonomous systems need to be

transformed into autonomous systems. This means that the functions are not required to satisfy the property that they must be differentiable w.r.t. t. The proofs of Halkin are based on special variations, which are different from the variations of Mc Shane, and which lead to the construction of some convex sets. The convex cones of Mc Shane are spanned by these convex sets. Halkin [24] develops the Maximum Principle for a class of problems which depends only on the control system with initial conditions, and not on the particular control problem. It is required that the functions are sufficientl' differentiable. We outline a generalization of this proof, [25], in a later section, where the differentiability conditions are relaxed. A certain knowledge of measure theory is required. In a later result, Halkin [27] dispenses with measure theoretical results, by considering the class of piecewise continuous controls and piecewise continuous functions. Subsequently, Bryant and Mayne [13] improved on Halkin's proof by extending U to be the class of extended piecewise continuous controls, and at the same time they simplified the mathematical content.

In the rest of this essay, we shall concentrate on outlining the proofs of the Maximum Principle by Pontryagin et al [52], Halkin [25], [27] and Bryant and Mayne [13] in greater detail and attempt to give the reader an intuitive insight into the approaches used in these proofs.

4. THE PROOF OF PONTRYAGIN, BOLTYANSKII, GAMKRELIDZE AND MISCHENKO OF THE MAXIMUM PRINCIPLE, [52].

In this section we propose to give an intuitive outline of the approach used by Pontryagin et al to derive the Maximum Principle. For a more detailed 'heuristic' presentation of the Maximum Principle, we refer the reader to Athans and Falb, [1].

Assumptions.

Let

(i) $\Omega \subset R^m$ be time invariant

(ii) U, the set of admissible controls, be the class of all bounded measurable vector valued functions u of t which satisfy the condition $u(t) \in \Omega$, $\forall\ t \in [t_o, t_f]$.

(iii) $f(x,u)$, $f_x(x,u)$ be continuous on $R^n \times \bar{\Omega} \times T$, where $\bar{\Omega}$ is the closure of Ω.

Statement of Problem.

The control problem has been formulated in Chapter 2. However, Pontryagin et al consider the autonomous problem with system equation of the form

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

Two special cases of the control problem are considered.

Special Problem 1 : The target set S is of the form

$$S = \{x_f\} \times [t_o, t_f],$$

where $G(t) \equiv x_f$ for $-\infty < t < \infty$.

x_f is a fixed element of R^n. Thus special problem 1 is a

fixed-end-point, free-time problem.

Special Problem 2 : The target set S is of the form

$$S = S_1 \times [t_o, t_f],$$

where S_1 is either a smooth k-fold in R^n, or all of R^n. Thus special problem 2 is also a free-time problem.

The only difference between the two special problems is the difference between the forms of the target set. We shall first consider special problem 1 and outline the basic Maximum Principle for the problem, and then establish the transversality conditions which characterize special problem 2.

The Principle of Optimality.

This is a quite elementary, but very useful result. If $v(\cdot)$ is an admissible optimal control on $[t_o, t_f]$, and $\bar{x}(\cdot, v)$ the corresponding trajectory of the autonomous system which starts at $\bar{x}_o = \bar{x}(t_o)$ and ends at $\bar{x}_f = \bar{x}(t_f, v)$, then $v(\cdot)$ is also optimal on an interval $[t_1, t_2] \subset [t_o, t_f]$. Loosely stated, it says that any portion of an optimal control is also optimal.

In the alternative formulation of the control problem given in Chapter 2, this means that all the $\bar{x}(\cdot, u)$ which meet L must 'lie above' the trajectory generated by $v(\cdot)$.

A Small Change in the Initial Conditions and its Consequences.

Let $v(\cdot)$ be the optimal control and $x(\cdot, v)$ be the solution to the system with initial condition $x(t_o) = x_o$ corresponding to this control. Consider the new initial point

$$\bar{x}_o + \varepsilon \delta_o + o(\varepsilon),$$

where ξ_o is some constant n-vector (i.e. not dependent on ε), ε is small, and $o(\varepsilon)$ a vector such that

$$\lim_{\varepsilon \to 0} \left|\left| \frac{o(\varepsilon)}{\varepsilon} \right|\right| = 0.$$

The solution starting from the new initial point, generated by $v(\cdot)$ has the form

$$\bar{x}(\cdot, v) + \varepsilon \delta \bar{x}(\cdot) + o(\varepsilon)$$

where $\bar{x}$ is the solution of the following linear system

$$\dot{\delta \bar{x}}(t) = \left\langle \frac{\partial \bar{f}}{\partial \bar{x}}(x(t,v), v(t)), \delta \bar{x}(t) \right\rangle \qquad t \in [t_o, t_f] \quad (4.1)$$

$$\delta \bar{x}(t_o) = \xi_o.$$

We can view the vector ξ_o as being attached to $\bar{x}_o = (0, x_o)$. Then, $\xi_t = \delta \bar{x}(t)$ can be considered the result of moving ξ_o along the optimal trajectory at time $t \in [t_o, t_f]$. Let $\phi(t, t_o)$ be the $(n+1) \times (n+1)$ fundamental matrix associated with the linear system (4.1). Then

$$\xi_t = \delta \bar{x}(t) = \phi(t, t_o) \xi_o.$$

Thus (4.1) can be viewed as defining a linear transformation of the space of n+1 vectors 'attached' to $\bar{x}_o$ into the space of n+1 vectors 'attached' to $\bar{x}(t, v)$, $t \in [t_o, t_f]$. Since small changes in the initial conditions cause corresponding small changes in the solution to the differential equation, all trajectories of our system starting with the new initial point, which meet L to within first order in ε, will lie approximately above the trajectory.

Moving Hyperplanes.

Consider the adjoint of the linear system (4.1). This gives

$$\begin{pmatrix} \dot{p}^o(t) \\ \dot{p}(t) \end{pmatrix} = - \left(\begin{array}{c|c} 0 & 0 \\ \hline \frac{\partial f^o}{\partial x}(x(t,v),v) & \frac{\partial f^T}{\partial x}(x(t,v),v) \end{array} \right) \begin{pmatrix} p^o(t) \\ p(t) \end{pmatrix} \tag{4.2}$$

because $\partial \bar{f}^o / \partial x^o = 0$. Equivalently

$$\begin{aligned} \dot{p}^o(t) &= 0 \\ \dot{p}(t) &= - \frac{\partial f^o}{\partial x}(x(t,v),v) p^o(t) - \frac{\partial f}{\partial x}(x(t,v),v) p(t) \end{aligned} \tag{4.3}$$

Therefore, $p^o(t)$ is constant. i.e., $p^o(t) = p \quad \forall\, t \in T$. Defining

$$H(x,u,p,p^o) = p^o f^o(x,u) + \langle p, f(x,u) \rangle \tag{4.4}$$

we have $\dot{p}(t) = - \frac{\partial H}{\partial x}(x(t,v),v,p(t),p^o)$.

Now it can be shown by differentiating w.r.t. t that

$$\left\langle \begin{pmatrix} p^o(t) \\ p(t) \end{pmatrix}, \delta \bar{x}(t) \right\rangle = \text{constant}.$$

Let L_o be the hyperplane passing through $\bar{x}_o$ with equation

$$\left\langle \begin{pmatrix} \pi_o \\ \pi \end{pmatrix}, \bar{x} - \bar{x}_o \right\rangle = 0.$$

If (4.3) has initial condition $p^o(t_o) = \pi_o$, $p(t_o) = \pi$, then

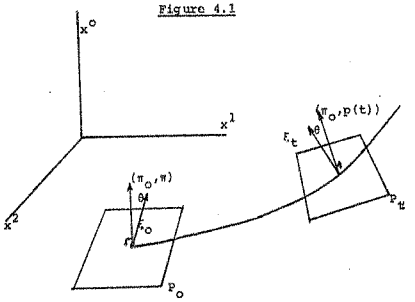
$$\left\langle \begin{pmatrix} \pi_o \\ p(t) \end{pmatrix}, \bar{x} - \bar{x}(t,v) \right\rangle = 0$$

defines a hyperplane L_t which passes through $\bar{x}(t,v)$ and has the same scalar product with the vector ξ_t as the vector ξ_o

makes with L_o. In particular, $\xi_o \perp L_o$ implies that $\xi_t \perp L_t$. The hyperplane L_t can be viewed as the result of moving the hyperplane L_o along the optimal trajectory, as is illustrated in Fig. 4.1.

We wish to determine the hyperplane passing through the particular point $\bar{x}(t,v)$. If $\bar{x}(t,v)$ is a boundary of a convex set C, then we know that there is a support hyperplane of C passing through $\bar{x}(t,v)$. So we now propose to introduce a suitable convex set which has $\bar{x}(t_f,v)$ as a boundary point and which has a support hyperplane that will be used to prove the Maximum Principle. These convex cones with vertices at $\bar{x}(t_f,v)$ will be constructed by considering a special class of variations of the control $v(\cdot)$. These cones will always lie on the same side of the hyperplane determined by the costate variables, and will be called the cones of attainability.

Figure 4.1



Variations of the Optimal Control.

Let $v(t)$ be the optimal control defined on t_o, t_f. Choose instants $\tau_1, \ldots, \tau_s, \tau$ satisfying the inequalities $t_o < \tau_1 \leq \ldots \leq \tau_s \leq t < t_f$ which are regular points of $v(t)$. (For the definition of a regular point, see [52]. Every point of continuity of $v(t)$ is regular). Further, choose arbitrary non-negative numbers $\delta t_1, \ldots, \delta t_s$, an arbitrary real number δt, and arbitrary (not necessarily different) points. Construct the semi-interval

$$I = \{ t \mid \tau_i + \varepsilon l_i < t \leq \tau_i + \varepsilon(l_i + \delta t_i) \},$$

where $l_i = \delta t - (\delta t_i + \ldots + \delta t_s)$ if $\tau_i = \tau$

$= \quad - (\delta t_i + \ldots + \delta t_s)$ if $\tau_i = \tau_s$

$= \quad - (\delta t_i + \ldots + \delta t_j)$ if $\tau_i = \tau_{i+1} = \ldots = \tau_j < \tau_{j+1}$, $j < s$.

The length of I_i is equal to $\varepsilon\delta t_i$. When $\delta t_i = 0$, the corresponding semi-interval I_i is 'empty', i.e., absent. Given a sufficiently small ε, the semi-intervals $I_1, \ldots, I_s$ do not intersect each other, and are all situated in the closed interval $t_o \leq t \leq t_f$, to the left of $\tau + \varepsilon\delta t$. Assuming that ε satisfies this condition enables us to define a control $u^*(t)$ on $t_o \leq t \leq \tau + \varepsilon\delta t$

as $u^*(t) = v_i \quad$ for $t \in I_i$

$= v(t), \quad t \notin \bigcup_{i=1}^{s} I_i,$

where the v_i, $i = 1, \ldots s$ are arbitrary points contained in Ω. Then $u^*(t)$ is an admissible control and is dependent on ε.

The Variation of the Trajectory.

Let ξ_o be the tangent vector to the trajectory $\bar{x}(\cdot, v)$ at the point $\bar{x}_o$. We now wish to find the position of the perturbed

trajectory which corresponds to the perturbed control $u^*(\cdot)$. $u^*(\cdot)$ differs from $v(\cdot)$ on a set with measure

$$\sum_{i=1}^{s} \mu(I_i) = \varepsilon(\delta t_1 + \ldots + \delta t_s).$$

For ε sufficiently small, the trajectory $\bar{x}(\cdot,u^*)$ is defined on $[t_o,\tau+\varepsilon\delta t]$. Then it can be shown, using induction over s, that

$$\bar{x}(\tau+\varepsilon\delta t,u^*) = \bar{x}(\tau,u^*) + \varepsilon\Phi(\tau,t_o)\xi_o + \varepsilon\Delta\bar{x} + o(\varepsilon),$$

where $\Delta\bar{x}$ is a vector independent of ε, defined by the equation

$$\Delta\bar{x} = \bar{f}(x(\tau,v),v)\delta t + \sum_{i=1}^{s} \Phi(\tau,\tau_i)\{\bar{f}(x(\tau_i,v),v_i) - \bar{f}(x(\tau_i,v),v(\tau_i))\}\delta t_i.$$

The Fundamental Construction.

The vector $\Delta\bar{x}$ is a function of $v_i,\tau_i,\tau,\delta t,\delta t_i$. Let

$$a \triangleq \{\tau_i,v_i,\tau,\delta t,\delta t_i\},$$

and $\Delta\bar{x} = \Delta\bar{x}_a$. We can regard $\Delta\bar{x}_a$ as a related vector issuing from the point $\bar{x}(\tau)$. If we take all possible symbols a (τ is fixed), the vectors $\Delta\bar{x}_a$ fill some set K_τ. It can be shown that the set K_τ is a convex cone with vertex $\bar{x}(\tau,v)$ - the cone of attainability. See Fig. 4.2.

Let τ be a regular point of the optimal control $v(\cdot)$, with associated trajectory $\bar{x}(\cdot,v)$, and Γ some curve issuing from the point $\bar{x}(\cdot,v)$, having a tangent ray μ at the point $\bar{x}(\tau,v)$. Then it can be shown that if the ray μ belongs to the interior of the cone K_τ, there exists some control $u^*(\cdot)$ such that the corresponding trajectory $\bar{x}(\cdot,u^*)$ with initial point $\bar{x}_o$, passes through a point different from $\bar{x}(\tau,v)$ of the curve Γ. See Fig. 4.3.

This means that the ray μ_τ issuing from the point $\bar{x}(\tau,v)$

in the direction of the negative semi-axis does not belong to the cone K_τ. Otherwise, taking as the curve Γ (and the ray μ), the ray μ_τ, we can constuct some control $u^*(\cdot)$ with corresponding trajectory which would pass through μ_τ at a point different from $\bar{x}(\tau, v)$. This leads to a contradiction of the optimality of $v(\cdot)$, $\bar{x}(\cdot, v)$.

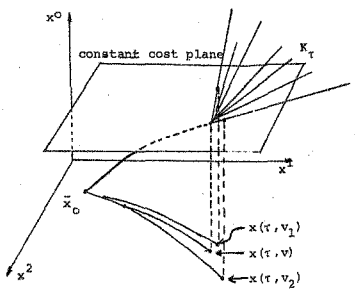


Figure 4.2

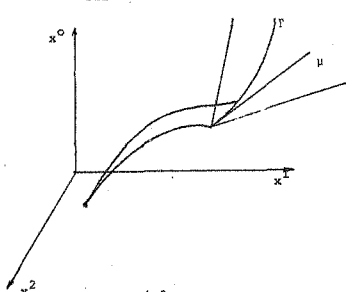


Figure 4.3

Properties of the Hamiltonian.

This section contains those properties of the Hamiltonian defined in (4.4) which will later be refined to yield the Maximum Principle. Because the ray μ_τ does not belong to the interior of the cone K_τ, K_τ does not fill the whole of R^{n+1}. Hence, K_τ has a hyperplane of support at its vertex with equation

$$\langle \bar{d}, \bar{x} \rangle = 0.$$

This hyperplane divides R^{n+1} into two halfspaces, the halfspace

$$\langle \bar{d}, \bar{x} \rangle \leq 0$$

containing K_τ (by a trivial modification, if necessary, of the constants d^i). In particular, $\Delta\bar{x} \in K_\tau$, so that

$$\langle \bar{d}, \Delta\bar{x} \rangle \leq 0 \qquad \forall\, \Delta\bar{x} \in K_\tau. \tag{4.5}$$

Putting $\delta t_1 = \delta t_2 = \ldots = \delta t_s = 0$ in the defining relation of $\Delta\bar{x}$ we obtain

$$\Delta\bar{x} = \bar{f}(x(\tau,v), v(\tau))\,\delta t$$

and hence

$$\langle \bar{d}, \bar{f}(x(\tau,v), v(\tau))\,\delta t \rangle \leq 0.$$

Since δt can be positive or negative we have

$$\langle \bar{d}, \bar{f}(x(\tau,v), v(\tau)) \rangle = 0,$$

or

$$H(x(\tau,v), v(\tau), d, d^0) = 0. \tag{4.6}$$

Let us write $\bar{p}(t,\bar{d})$ for the solution of the system of equations (4.2) with the initial conditions $\bar{p}(\tau,\bar{d}) = \bar{d}$. Then the solution $\bar{p}(t,\bar{d})$ is defined throughout the interval $t_0 \leq t \leq t_f$, since the system is linear. If the vector $\bar{d}$ satisfies condition (4.5), then it can be proved that

$$H(x(t,v),v,p(t,\bar{d}),p^{o}(t,\bar{d})) = M(x(t,v),\bar{p}(t,\bar{d})) \qquad (4.7)$$

at every regular point of control $v(\cdot)$, where

$$M(x(t,v),\bar{p}(t,\bar{d})) = \sup_{u\in\Omega} H(x(t,v),u,p(t,\bar{d}),p^{o}(t,\bar{d})).$$

From (4.6), (4.7) we have that if $\bar{d}$ satisfies (4.5) then

$$M(x(\tau,v),\bar{p}(\tau,\bar{d})) = 0 \qquad (4.8)$$

Finally, it is proved that if the absolutely continuous function $\bar{p}(t)$ satisfies equations (4.2) a.e. on closed interval I and the relationship

$$H(x(t,v),v,p(t),p^{o}(t) = M(x(t,v),\bar{p}(t)),$$

then $M(x(t,v),\bar{p}(t)) = \text{constant for all } t\in I. \qquad (4.9)$

The Limiting Cone.

If τ, τ' are regular points of the control $v(\cdot)$, where $t_o < \tau' < \tau < t_f$, then

$$\Phi(\tau,\tau')K_{\tau'} \subset K_{\tau}.$$

Let be any regular point of $v(\cdot)$ lying in (t_o,t_f). Define

$$K_{t_f}^{(\tau)} \triangleq \Phi(t_f,\tau)K_{\tau}.$$

Since $\Phi(t_f,\tau)$ is a linear mapping, $K_{t_f}^{(\tau)}$ is a convex cone. The cones $K_{t_f}^{(\tau)}$ form an increasing sequence: if $\tau' < \tau$ are regular points,

$$K_{t_f}^{(\tau')} = \Phi(t_f,\tau')K_{\tau'} = \Phi(t_f,\tau)\,\Phi(\tau,\tau')K_{\tau'} \subset \Phi(t_f,\tau)K_{\tau} = K_{t_f}^{(\tau)}.$$

Hence the union (w.r.t. all regular points τ of the interval (t_o,t_f)) of all cones $K_{t_f}^{(\tau)}$ is again a convex cone (possibly not closed),

with vertex at $\bar{x}(t_f,v)$. This cone will be denoted by K_{t_f}, and we shall call it the limiting cone. It can be shown that the ray μ_{t_f}, issuing from the point $\bar{x}(t_f,v)$ in the direction of the negative semi-axis x^0, does not belong to the interior of the cone K_{t_f}.

The Maximum Principle.

We now combine our results to obtain the Maximum Principle. Let $(x(\cdot,v),v(\cdot))$ be an optimal process. Let $\bar{c} \neq 0$ be a (n+1) vector. Then the cone K_{t_f} lies in the halfspace

$$\langle \bar{c},\bar{x} \rangle \leq 0$$

and the ray μ_{t_f} lies in the halfspace

$$\langle \bar{c},\bar{x} \rangle \geq 0. \tag{4.10}$$

Now, because the vector $(-1,0,\ldots,0)$ lies in the halfspace defined in (4.10), we have

$$c^0 \leq 0.$$

Suppose $\bar{p}(t)$ is a solution of the adjoint system given in (4.2) with initial condition $\bar{p}(t_f) = \bar{c}$. Then the function $\bar{p}(t)$ is unique and continuous on $[t_o,t_f]$. The initial condition implies

$$p^0(t_f) = c^0 \leq 0,$$

but since the right hand side of (4.2) is independent of x^0,

$$p^0(t) = \text{constant} \leq 0 \quad \forall\, t\in[t_o,t_f]. \tag{4.11}$$

Furthermore, since $\bar{p}(t)$ is a solution of (4.2), and $\bar{c} \neq 0$, it follows that $\bar{p}(t) \neq 0$. The cone $\Phi(t_f,\tau)K_\tau$ lies in the halfspace

$$\langle \bar{c},\bar{x} \rangle \leq 0,$$

or equivalently, $\langle \bar{p}(t_f), \bar{x} \rangle \leq 0$.

Since the transformation $\phi^{-1}(t_f,\tau)$ is linear and homogeneous, this implies that the cone K_1 lies in the halfspace

$$\langle \bar{p}(\tau), \bar{x} \rangle \leq 0.$$

In other words, the vector $\bar{d} = \bar{p}(\tau)$ satisfies the condition

$$\langle \bar{d}, \Delta\bar{x} \rangle \leq 0 \quad \Delta\bar{x} \in K_1.$$

But $\bar{p}(t,\bar{d})$ satisfies the initial condition $\bar{p}(\tau,\bar{d}) = \bar{d}$, so that the functions $\bar{p}(t)$ and $\bar{p}(t,\bar{d})$ are evidently identical. Moreover, from (4.6) - (4.9) it follows that

(i) $H(x(t,v),v,p(t),p^0) = M(x(t,v),\bar{p}(t))$ a.e. on $[t_0,t_f]$

(ii) $M(x(t,v),\bar{p}(t)) = 0$ for $t \in [t_0,t_f]$

where, from (4.11),

(iii) $p^0(t) = \text{constant} \leq 0$ for $t \in [t_0,t_f]$.

Conditions (i),(ii),(iii) are the necessary conditions for the optimality of the process $(x(\cdot,v),v(\cdot))$ for the special problem 1.

The Transversality Conditions.

The major difference between special problem 1 and special problem 2 is that the latter problem is a variable endpoint problem, whereas the former is a fixed endpoint problem. In the latter case we require relationships from which the position of the point x_f on the manifold S_1 can be defined. Such relationships are provided by the transversality conditions. These conditions enable

us to write k relationships in terms of the co-ordinates of the end-point x_{t_f}. Since, on the other hand, the number of unknown parameters (in comparison with the problem with fixed ends) is also increased by k (because the position of the point $x_{t_f} \in S_1$ is characterized by k parameters), in conjunction with the Maximum Principle, the transversality conditions form a 'sufficient' system of conditions for the solution of the optimal problem with moving ends.

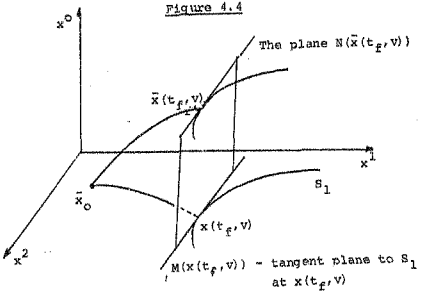
Let S_1 be our smooth k-fold in R^n and $x(t_f,v)$ be the terminal point of the optimal trajectory in R^n, so that $x(t_f,v) \in S_1$. Then the tangent plane

$$M(x(t_f,v)) \triangleq \{ x : \langle (\partial g_i/\partial x)(x(t_f,v)), x - x(t_f,v) \rangle = 0 \quad \text{for } i = 1,2,\ldots,n-k \}$$

is well defined and is a k-dimensional plane in R^n passing through $x(t_f,v)$. Denote the k-dimensional plane passing through the point $\bar{x}(t_f,v)$ as illustrated in Fig. 4.4, by

$$N(\bar{x}(t_f,v)) \triangleq \{ \bar{x} : \bar{x} = (x^0(t_f,v), x), \quad x \in M(x(t_f,v)) \}.$$

Figure 4.4



Every element of $N(\bar{x}(t_f,v))$ can be written as

$$\bar{x} = \bar{x}(t_f,v) + \begin{bmatrix} 0 \\ \hat{x} \end{bmatrix}$$

where $\hat{x}$ is an n-vector with the property that

$$< \frac{\partial q}{\partial x}(x(t_f,v)), \hat{x} > = 0 \quad i = 1,\ldots,n-k.$$

Let $\hat{M} \triangleq \{ \hat{x} :$ there is an $\bar{x}$ in $N(\bar{x}(t_f,v))$ such that $\bar{x} - \bar{x}(t_f,v) = (0,\hat{x}) \}$, then we can show that $\hat{M}$ is a subspace, and the set $\hat{N} \triangleq \{ \hat{x} : \hat{x} = (0,\hat{x}) : \hat{x} \in \hat{M} \}$ is a subspace of R^{n+1}. $\hat{M}$ and $\hat{N}$ are both of dimension k. Also

$$N(\bar{x}(t_f,v)) = \bar{x}(t_f,v) + \hat{N}.$$

Since $\hat{N}$ is a subspace of R^{n+1}, the set $N(\bar{x}(t_f,v))$ is a convex cone with $\bar{x}(t_f,v)$ as vertex. Let μ be the ray in the direction of the negative semi-axis x^0 issuing from the point $\bar{x}(t_f,v)$. Then

$$N(\bar{x}(t_f,v)) + \mu = \{ \bar{x} : \bar{x} = \bar{x}(t_f,v) + \hat{\bar{x}} + \beta e, \quad \hat{\bar{x}} \in \hat{N}, \ \beta \geq 0, \quad e = (-1,0,\ldots,0)^T \}$$

is a convex cone with vertex at $\bar{x}(t_f,v)$. It can be shown that the convex cones $N(\bar{x}(t_f,v)) + \mu$ and K_{t_f} are seperated. It follows that there is a nonzero n+1 vector $(p^0,p(t_f))^T$ such that

$$< \begin{bmatrix} p^0 \\ p(t_f) \end{bmatrix}, \bar{x} - \bar{x}(t_f,v) > \geq 0 \quad \forall \ x \in N(\bar{x}(t_f,v)) + \mu$$

i.e.

$$< \begin{bmatrix} p^0 \\ p(t_f) \end{bmatrix}, \hat{\bar{x}} + \beta e > \geq 0 \quad \forall \ \hat{\bar{x}} \in \hat{N}, \tag{4.12}$$

and

$$< \begin{bmatrix} p^0 \\ p(t_f) \end{bmatrix}, \Delta\bar{x} > \leq 0 \quad \forall \ \Delta\bar{x} \in K_{t_f}.$$

Since $\hat{N}$ is a subspace of R^{n+1}, 0 is an element of $\hat{N}$, and equation (4.12) implies that

$$< \begin{bmatrix} p^0 \\ p(t_f) \end{bmatrix} , e > \geq 0.$$

On the other hand, letting $\beta = 0$ in (4.12) implies that

$$< \begin{bmatrix} p^0 \\ p(t_f) \end{bmatrix} , \hat{x} > \geq 0 \quad \forall \ \hat{x} \in \hat{N}.$$

As $\hat{N}$ is a subspace of R^{n+1}, $\hat{x} \in \hat{N}$ implies that $-\hat{x} \in \hat{N}$. Thus

$$< \begin{bmatrix} p^0 \\ p(t_f) \end{bmatrix} , \hat{x} > = 0 \quad \forall \ \hat{x} \in \hat{N},$$

and we may conclude that if S_1 is a smooth k-fold in R^n that

(iv) $\langle p(t_f), x - x(t_f, v) \rangle = 0$ for all $x \in M(x(t_f, v))$.

If S_1 is all of R^n, N is the entire half space 'below' the hyperplane $x^0 - x^0(t_f, v) = 0$. As the cone K_{t_f} must lie 'above' this hyperplane, we can see that the vector $[p^0, p(t_f)]^T$ must be of the form $[p^0, 0]^T$. i.e.

(v) $p(t_f)$ must be the zero vector.

Conditions (iv) and (v) are known as the transversality conditions.

Extensions of the Maximum Principle to other Problems.

We have outlined the Maximum Principle for the two special problems mentioned earlier. The above proofs can be adapted or extended to obtain a Maximum Principle to many different optimal control

problems. Most optimal control problems, in particular the non-autonomous case, can be transformed into one of the two problems which we have considered above. For non-autonomous control problems it is required that the functions be continuously differentiable w.r.t. t. The reader should consult, for example Pontryagin et al [52], Lee and Markus [41], Horn [31], to see how the Maximum Principle as presented above can be extended to various optimal control problems.

Varaiya, [59], remarks that the book by Pontryagin et al contains many extensions and examples, and is still an important source. However, the derivation of the Maximum Principle given in the book by Lee and Markus [41] is more satisfactory. The approach used by Lee and Markus is closely related to the approach used by Pontryagin et al which we have outlined above.

5. HALKIN'S PROOF OF THE MAXIMUM PRINCIPLE, [27].

Halkin obtains the necessary conditions by a method fundamentally different to the method used by Pontryagin and his associates. Halkin notes that he avoids some unresolved topological difficulties encountered in the reasoning in [52].

In the remaining sections, for notational simplicity, we shall call $\tilde{x}$, $\tilde{f}$ just x, f etc. In order to be consistant with the original derivations we shall consider maximizing the x^n co-ordinate, where x is an n-vector (as opposed to minimizing x^0 where $\tilde{x}$ is an (n+1)-vector). The minimization problem can be transformed into an equivalent maximization problem. $T \triangleq [0,1]$.

Assumptions.

Let

(i) $\Omega \subset R^m$ be time invariant.

(ii) U be the class of all bounded measurable vector functions which satisfy the condition $u(t) \in \Omega$, $\forall\ t \in [0,1]$.

The totality of the function space U is not considered as the stategy space because strong assumptions on the function $f(x,u,t)$ have to made to ensure that for every $u \in U$, the solution to

$$\dot{x}(t) = f(x(t),u(t),t) \quad \text{a.e. } t \in [0,1], \tag{5.1}$$

$$x(t_0) = x_0 \tag{5.2}$$

exists and is unique.

(iii) $U^* \subset U$ be defined as the set of all $u \in U$ for which there exists an x, continuous and a.e. differentiable, satisfying (5.1), (5.2).

(iv) There exists $\varepsilon > 0$ such that $f(x,u,t)$, f_x are defined, measurable w.r.t. u and t, uniformly equicontinuous w.r.t. x and uniformly bounded for all $(x,t,u) \in N(x,\varepsilon) \times \Omega^*$, where

$$N(x,\varepsilon) \triangleq \{ (\bar{x},\bar{t}) \mid |\bar{x} - x|^2 + |t - \bar{t}|^2 < \varepsilon \},\ t\in[0,1],$$

and Ω^* is a bounded subset of Ω.

The conditions on the differentiability and boundedness of the function $f(x,u,t)$ are much weaker than the assumptions made by Pontryagin et al [52]. In an earlier publication, [24], Halkin, using the same method as outlined below, obtained the same results for a more restrictive class of problems. In particular, $f(x,u,t)$ was required to be sufficiently differentiable on some subset of $X \times \Omega \times [0,1]$.

Statement of Problem.

The problem studied by Halkin can be stated as follows: Consider a line B in $X \times [0,1]$, parallel to the x^n axis, and determined by its projection x_1^i, $i = 1,\ldots,n-1$ and $t = 1$ on the other axis. More precisely, B is the set

$$\{ (x,t) : x^i = x_1^i,\ i = 1,\ldots,n-1,\ x^n\in R,\ t = 1 \}. \quad (5.3)$$

Hence, if $v\in U^*$ is optimal, then

$$(x(1,v),1) \in B. \quad (5.4)$$

Furthermore, for any $u\in U^*$ with the property $(x(1,u),1) \in B$, then necessarily

$$x^n(1,u) \le x^n(1,v). \quad (5.5)$$

The statement of the problem given here is made up of two parts: the control system with initial conditions, and the optimal control problem for the control system with initial conditions, defined by (5.3) - (5.5).

The developments which follow depend only on the control system with initial conditions, not the particular control problem. This allows Halkin to dispense with the formal transformations required in order to apply the Maximum Principle, and to dispense with the consideration of transversality conditions which, after such transformations, are needed to obtain a non-trivial set of necessary conditions for an optimal solution.

Remarks on the Structure of $f(x,u,t)$.

In contrast to Pontryagin, Halkin allows $f(x,u,t)$ to be dependent on the variable x^n to be maximized at time $t = 1$. To assume $f(x,u,t)$ independent of x^n leads to very little simplification in the proof of the Maximum Principle. Also, many practical problems show a natural dependence on the variable to be maximized.

Halkin does not require the differential system to be time independent. The problem may be transformed into the problem treated by Pontryagin by the introduction of an appropriate artificial variable. However, this transformation cannot be done if the time dependence of the d.e.'s is not continuous. This is in contrast to Halkin's very weak assumptions on the time dependence of the differential equations; he only requires measurability w.r.t. time.

The Comoving Co-ordinate Space along a Trajectory.

Let $v \in U^*$ be the optimal control (which we assume exists). Denote by Y the n-dimensional space, and introduce the non-singular mapping $G(t,v) \in R^{n \times n}$ from $X \times [0,1]$ into $Y \times [0,1]$ by

$$y(t,u) = G(t,v)\,(x(t,u) - x(t,v)) \qquad u \in U^*$$

where $G(t,v)$ is continuous w.r.t. t and satisfies

$$\dot{G}(t,v) = - G(t,v)\, A(t) \qquad \text{a.e. } t \in [0,1]$$

$$G(1,v) = I$$

where $$A(t) \triangleq f_x(x(t,v),v(t),t)$$

and I is the $n \times n$ identity matrix. $Y \times [0,1]$ is called the comoving co-ordinate space along the trajectory $x(t,v)$. We have

$$y(t,u) = \int_0^t G(\tau,v)\,(\phi(\tau,u) + k(\tau,u)\,d\tau$$

where

$$\phi(t,u) = f(x(t,v),u,t) - f(x(t,v),v,t)$$

$$k(t,v) = f(x(t,u),u,t) - f(x(t,v),v,t) - \phi(t,u) - A(t)\,(x(t,u) - x(t,v)).$$

The Approximate Trajectory.

Let $y(t,u)$ have approximation

$$y^+(t,u) = \int_0^t G(\tau,v)\ \phi(\tau,u)\,d\tau$$

$$y^+(0,u) = 0.$$

Note that $y(t,v) = y^+(t,v) \equiv 0 \quad \forall\, t \in [0,1]$. We have

$$y(t,u) - y^+(t,u) = \int_0^t G(\tau,v)k(\tau,u)\,d\tau,$$

$$y(0,u) - y^+(0,u) = 0.$$

Reachable Sets.

Consider the set $W(1)$ defined before and given by

$$W \triangleq \{ x(1,u) : u \in U^* \}.$$

Consider the intersection of the set of all reachable events from the initial event $y = 0$ by the trajectory $y(t,u)$, $u \in U^*$ with the hyperplane $t = 1$:

$$W(v) \triangleq \{ y(1,u) : u \in U^* \}.$$

Similarly, consider the set $W^+(v)$ which is the intersection of the set of all reachable events from the initial event $y^+ = 0$ by the trajectory $y^+(t,u)$, $u \in U^*$ with the hyperplane $t = 1$:

$$W^+(v) \triangleq \{ y^+(1,u) : u \in U^* \}.$$

For a linear system, $W(v) = W^+(v)$.

Convexity of the Range of a Vector Integral over Borel Sets.

Halkin proves the well known result in measure theory, Liapunov's theorem:

If f is a Lebesque integrable function from $[0,1]$ into R^n and if B is the class of all Borel subsets of $[0,1]$, then the set $\{ \int_E f(t)dt : E \in B \}$ is convex.

Hence, it can be shown that

$$W^+(v) = \{ \int_0^1 \phi(t,u)dt : u \in U^* \}$$

is convex, where

$$\phi(t,u) \triangleq G(t,v) (f(x(t,v),u,t) - f(x(t,v),v,t)).$$

Reachable Sets.

Consider the set $W(1)$ defined before and given by

$$W \triangleq \{ x(1,u) : u \in U^* \}.$$

Consider the intersection of the set of all reachable events from the initial event $y = 0$ by the trajectory $y(t,u)$, $u \in U^*$ with the hyperplane $t = 1$:

$$W(v) \triangleq \{ y(1,u) : u \in U^* \}.$$

Similarly, consider the set $W^+(v)$ which is the intersection of the set of all reachable events from the initial event $y^+ = 0$ by the trajectory $y^+(t,u)$, $u \in U^*$ with the hyperplane $t = 1$:

$$W^+(v) \triangleq \{ y^+(1,u) : u \in U^* \}.$$

For a linear system, $W(v) = W^+(v)$.

Convexity of the Range of a Vector Integral over Borel Sets.

Halkin proves the well known result in measure theory, Liapunov's theorem:

If f is a Lebesque integrable function from $[0,1]$ into R^n and if B is the class of all Borel subsets of $[0,1]$, then the set $\{ \int_E f(t)dt : E \in B \}$ is convex.

Hence, it can be shown that

$$W^+(v) = \{ \int_0^1 \phi(t,u)dt : u \in U^* \}$$

is convex, where

$$\phi(t,u) \triangleq G(t,v) \{ f(x(t,v),u,t) - f(x(t,v),v,t) \}.$$

We are now in a position to outline Halkin's proof of the Maximum Principle. The guiding idea behind the derivation of the necessary conditions can be understood as follows: Halkin derives easily a set of necessary conditions for optimality of $x(\cdot,v)$. He then proves that his conclusions are still valid when y^+ is a close enough approximation to y.

1. The Principle of Optimal Evolution. See Chapt. 2.

If v is an optimal control function, then for every $t\in[0,1]$ the state $x(t,v)$ belongs to the boundary of the set

$$W(t) \triangleq \{ x(t,u) : u \in U^* \},$$

where $W(1) \equiv W$.

2. The Properties of the Boundary of $W(v)$.

We have that

$$y(1,u) = G(1,v)(x(1,u) - x(1,v)),$$

where $G(1,v)$ is the identity matrix. Hence $W(v)$ is a simple translation of W. This translation conserves the topological properties of the points in W; in particular, to a boundary point of W corresponds a boundary point of $W(v)$, and conversely. Therefore, $y(1,v) = 0$ is a boundary point of $W(v)$.

3. The Properties of the Boundary of $W^+(v)$.

Firstly we recall that $W^+(v) = W(v)$ for a system which is linear in x. For the nonlinear system assume that $0 \in$ interior of $W^+(v)$. Then it can be shown, using amongst other results,

Brouwer's Fixed Point Theorem†, that $0 \notin$ interior $W(v)$. This contradicts 2, so 0 is a boundary point of $W^+(v)$.

4. Properties of the Support Hyperplane to $W^+(v)$.

We have shown previously that the set $W^+(v)$ is convex. Hence, if $y = 0$ is a boundary point of $W^+(v)$, there exists a nonzero constant vector defining a hyperplane

$$\langle \pi(v), y \rangle = 0$$

$$\text{s.t.} \quad \langle \pi(v), y \rangle \leq 0 \qquad \forall\, y \in W^+(v).$$

Now assume that there is a $u \in U$ such that

$$\langle \pi(v), G(t,v)(f(x(t,v),u,t) - f(x(t,v),v,t)) \rangle \geq \varepsilon > 0$$

for $t \in E$ where $E \in B$ with $\mu(E) > 0$. Introducing the vector valued function

$$u^*(t) = v(t) + \chi(E)(u(t) - v(t)) \qquad \forall\, t \in T \quad ††$$

gives $$\langle \pi(v), y^+(1,u^*) \rangle \geq \varepsilon\mu(E) > 0. \qquad (5.6)$$

But $y^+(1,u^*) \in W^+(v)$. This contradicts (5.6). Therefore,

$$\langle \pi(v), G(t,v)(f(x(t,v),u,t) - f(x(t,v),v,t)) \rangle \leq 0.$$

Defining $$p(t) = G^T(t,v)\,\pi(v),$$

we have that $p(t)$ is nonidentically zero and continuous over T.

---

† Brouwer's Fixed Point Theorem states that if a continuous function maps a closed convex subset of an Euclidean space R^n into itself, then it has a fixed point. In other words, if $h(x)$ is a continuous function defined on the closed convex set A such that $h(x) \in A$ for all $x \in A$, then there is an $x \in A$ such that $h(x) = x$.

†† $\chi(t) = 1$ if $t \in E$,
$= 0$ if $t \notin E$.

Then it can be shown that

$$\langle p(t), f(x(t,v),u,t) - f(x(t,v),v,t)\rangle \leq 0 \qquad (5.7)$$

and

$$\begin{aligned} \dot{p}(t) &= \dot{G}^T(t,v)\,\pi(v) && \text{a.e. } t\in T \\ &= \left(-G(t,v)A(t)\right)^T \pi(v) && \text{a.e. } t\in T \\ &= -A^T(t)\,p(t) && \text{a.e. } t\in T. \end{aligned} \qquad (5.8)$$

5. Proof of the Maximum Principle.

If the element v of U^* is optimal, then we have from 1 that $x(1,v)$ is a boundary point of W. Then from 2, 3, $y = 0$ is a boundary point of the set $W^+(v)$. From 4 we have that there exists a vector $p(t)$, continuous and nonidentically zero on T such that conditions (5.7), (5.8) are satisfied. Defining

$$H(x,u,p,t) = \langle p, f(x,u,t)\rangle$$

we obtain the Maximum Principle of Pontryagin.

It was stated earlier that the results obtained did not depend on the control problem, but the initial conditions. Necessary conditions have been derived for an element $v\in U^*$ where $x(1,v)$ is a boundary point of the set W. If the dimension of the set W is less than n, then the results are trivial, since for any $u\in U^*$, the point $x(1,u)$ will be a boundary point of W.

In the next section we shall discuss a proof by Halkin of the necessary conditions which dispenses with measure theoretical results.

6. THE MAXIMUM PRINCIPLE : HALKIN [27], BRYANT AND MAYNE [13].

A fundamental result required to establish the necessary conditions of optimality in the last section was Liapunov's Theorem. Halkin, [26], proved a new result of the same type which enabled him to simplify the previous approach mathematically and does not contain any measure theoretical concepts.

Assumptions.

Let

(i) $\Omega \subset R^m$ be time invariant

(ii) U be the class of piecewise continuous controls from $[0,1]$ into Ω.

This means that each function in the class U is continuous at every point $[0,1]$ except at a finite number of points where it has finite left and right limits, and has a finite right limit at 0 and a finite left limit at 1. The preceding definition implies, in particular, that every function in the class U is bounded. For $u \in U$, let $\theta(u)$ be defined as the set of points in $[0,1]$ at which the function u and the function $f(x,u,t)$ are continuous w.r.t. t.

(iii) The vector valued function $f(x,u,t)$ is defined for all $x \in R^n$, $u \in R^m$, $\forall\, t \in [0,1]$.

(iv) $f(x,u,t)$ is $2\times$ continuously differentiable w.r.t. x, continuous w.r.t. u, and piecewise continuous w.r.t. t.

Consider the system described by

$$\dot{x}(t,u) = f(x(t,u),u(t),t) \qquad \forall\, t \in \theta(u)$$
$$x(0,u) = 0.$$

To prevent a finite escape time, assume

(v) there exists a constant $M < \infty$ such that

$$|f(x,u,t)| \leq M(|x| + 1) \quad \forall\, x \in R^n,\ u \in \Omega,\ \forall\, t \in [0,1].$$

Statement of Problem.

Consider maximizing $x^n(1,u)$ subject to the control constraint $u \in U$ and endpoint constraints of the form

$$x^i(1,u) = s_i \qquad i = 1,\ldots,r \leq n-1.$$

Define $$S \triangleq \{\, x : x \in R^n,\ x^i = s_i,\ i = 1,\ldots,r \,\}.$$

Hence if $v \in U$ is optimal, then

$$x(1,v) \in S,$$

and furthermore, if $u \in U$ and $x(1,u) \in S$, then necessarily,

$$x^n(1,u) \leq x^n(1,v).$$

This formulation corresponds to the Mayer Problem of the classical calculus of variations. As is well known, problems involving integral cost components and other forms of endpoint constraints can be transformed to this form by the introduction of extra states. We still allow $f(x,u,t)$ to be dependent on x^n and don't require the differential system to be time independent.

The Variation Trajectory.

Let $v \in U$ be the optimal control (which we assume exists). Define the variational trajectory for the control u w.r.t. the control function v :

$$y(t,u) = x(t,u) - x(t,v).$$

So $y(t,v) = 0.$

Let $\theta^*(u) = \theta(u) \cap \theta(v)$. Then $y(t,u)$ is continuous w.r.t. t $\forall\ t\in[0,1]$, differentiable w.r.t. t $\forall\ t\in\theta^*(u)$, and

$$\dot{y}(t,u) = f(x(t,u),u,t) - f(x(t,v),v,t) \quad \forall\ t\in\theta^*(u).$$

Then defining

$$\phi(t,u) = f(x(t,v),u,t) - f(x(t,v),v,t)$$

$$k(t,u) = f(x(t,u),u,t) - f(x(t,v),v,t) - \phi(t,u) - A(t)\,(\,x(t,u) - x(t,v)\,),$$

where $A(t) \triangleq f_x(x(t,v),v,t)$,

we have

$$\dot{y}(t,u) = A(t)y(t,u) + \phi(t,u) + k(t,u) \quad \forall\ t\in\theta^*(u). \tag{6.1}$$

Let $\Phi: T \times T \to R^{n\times n}$ denote the transition matrix for the system (6.1). Hence $\Phi(t,\tau)$ is the unique solution of the matrix differential equation

$$\dot{\Phi}(t,\tau) = A(t)\Phi(t,\tau)\ ; \qquad \Phi(\tau,\tau) = I,$$

and we have

$$y(t,u) = \int_0^t \Phi(t,\tau)\,(\,\phi(\tau,u) + k(\tau,u)\,)\,d\tau \quad \forall\ t\in[0,1]$$

$$y(0,u) = 0.$$

For the system which is linear in x, i.e.

$$\dot{x} = Ax + \phi(t,u),$$

$k(t,u) \equiv 0$. For the nonlinear problem, $k(t,u)$ can be shown to have the following properties:

(i) There exists $K_1 < \infty$, such that

$$|k(t,u)| \le K_1 |y(t,u)|^2 \quad \forall\ t\in[0,1] \text{ s.t. } u(t) = v(t).$$

(ii) There exists $K_2 < \infty$ such that $\forall\ u\in U$, $\forall\ t\in[0,1]$,

$$|k(t,u)| \le K_2|y(t,u)|.$$

It can also be shown that there exists $N < \infty$ $\forall\ u\in U$ and $E\in\mathcal{A}$ with $u(t) = v(t)$ $\forall\ t\in[0,1]\sim E$ such that

$$|x(\tau,u) - x(\tau,v)| \le N\mu(E) \quad \forall\ \tau\in[0,1].$$

where $\mathcal{A}$ is the class of all subsets of $[0,1]$ which are the union of a finite number of disjoint intervals.

In order to approximate $y(t,v)$ and the effects of strong variations in control, the following linear model is introduced.

The Approximate Trajectory.

For any $u\in U$ let $z(t,u)$ denote the unique solution to

$$\dot{z}(t,u) = A(t)z(t,u) + \phi(t,u) \qquad \forall\ t\in\theta^*(u), \tag{6.2}$$

$$z(0,u) = 0, \tag{6.3}$$

or equivalently,

$$z(t,u) = \int_0^t \Phi(t,\tau)\phi(\tau,u)\,d\tau \qquad \forall\ t\in[0,1].$$

Then

$$y(t,u) - z(t,u) = \int_0^t \Phi(t,\tau)k(\tau,u)\,d\tau \qquad \forall\ t\in[0,1],$$

$$y(0,u) - z(0,u) = 0.$$

Also $z(t,v) = 0$, because $\phi(t,v) = 0$.

$z(\cdot,u)$ is called the approximate trajectory for the variational trajectory $y(\cdot,u)$. This is well justified, since $y(\cdot,u) - z(\cdot,u)$ is small whenever $k(\cdot,u)$ is small, i.e., whenever $y(\cdot,u)$ is small, or whenever the comparison trajectory $x(\cdot,u)$ is close to the optimal trajectory $x(\cdot,v)$.

The Reachable Sets.

The system reachable set at $t = 1$ is defined by

$$W \triangleq \{ x(1,v) + y(1,u) : u \in U \},$$

and the reachable set at $t = 1$ for the system described by (6.2), (6.3) is defined by

$$\tilde{W} \triangleq \{ x(1,v) + z(1,u) : u \in U \}.$$

Equivalently the sets W, $\tilde{W}$ could be written as

$$W = \{ x(1,u) : u \in U \},$$
$$\tilde{W} = \{ x(1,v) + z : z \in Z \},$$
where $$Z = \{ z(1,u) : u \in U \}.$$

Define the set S^+ of all states in S with an n^{th} co-ordinate greater than $x^n(1,v)$. i.e.,

$$S^+ \triangleq \{ x : x \in S, x^n(1,u) > x^n(1,v), u \in U \}.$$

S^+ is convex, and by definition, S^+ and W have no points in common. Also, for a system linear in x, the sets W and $\tilde{W}$ are the same.

Convexity of the Range of a Vector Integral over the Class $\mathcal{R}$ of Subsets of $[0,1]$.

Consider a piecewise continuous function $f(t)$. The vector $I = \int_0^1 f(t)\,dt$ represents the average value of the function $f(t)$ on the interval $[0,1]$. The vector I is also the value at $t = 1$ of the function $g(t) = \int_0^t f(\tau)\,d\tau$. If we consider the estimation of the average I as a continuous process, then the

vector $g(t)$ may be regarded as a certain approximation of the fraction tI of the average vector I. This continuous estimation process is not very accurate if the function $f(t)$ fluctuates greatly. Instead of basing the estimation on a single 'balayage' of the interval $[0,1]$, we could consider a simultaneous balayage of each of the intervals $[0,\frac{1}{2})$ and $[\frac{1}{2},1]$ and introduce a function $g_1(t)$ defined over $[0,1]$ by

$$g_1(t) = \int_0^{\frac{t}{2}} f(\tau)\,d\tau + \int_{\frac{1}{2}}^{\frac{1+t}{2}} f(\tau)\,d\tau$$

as an approximation of the fraction tI of the average vector I. This process may be refined further: for each integer k we partition the interval $[0,1]$ into 2^k consecutive intervals of equal length $1/2^k$ and define a function $g_k(t)$ on $[0,1]$ by

$$g_k(t) = \sum_{i=1}^{2^k} \int_{(i-1)/2^k}^{(i-1+t)/2^k} f(\tau)\,d\tau,$$

or equivalently by the relation

$$g_k(t) = \int_{D_t^k} f(\tau)\,d\tau$$

where the set D_t^k is defined by

$$D_t^k = \bigcup_{i=1}^{2^k} \left[\frac{i-1}{2^k}\,,\ \frac{i-1+t}{2^k}\right).$$

Using Fourier Series methods, Halkin proves that the functions $g_1(t),\ldots,g_k(t),\ldots$ become more and more accurate approximations of the function tI as k increases. More precisely:

Balayage Theorem : If $f(t)$ is a piecewise continuous function from $[0,1] \to R^n$ and if $\varepsilon > 0$, then there exists an integer K s.t.

$$\left|\int_{D_\alpha^k} f(t)\,dt - \alpha\int_0^1 f(t)\,dt\right| \leq \varepsilon$$

for all $\alpha\in[0,1]$ and all $k \geq K$.

Consider the set

$$L(f) = \{ \int_A f(t)dt : A \in \mathcal{A} \},$$

where $\mathcal{A}$ is the class of all subsets of $[0,1]$. Using the above theorem as well as Brouwer's Fixed Point Theorem enables Halkin to prove:

Liapunov's Theorem : If the vector valued function $f(t)$ is piecewise continuous, then the set $L(f)$ is convex.

With the help of Liapunov's Theorem, Halkin proves that the reachable set for the linear system described by (6.2), (6.3), $\tilde{W}$, is convex.

We are now in a position to outline the stategy behind the proof of the Maximum Principle given in [27].

1. Principle of Optimal Evolution.

As before, we have that if v is an optimal control function, then for every $t\in[0,1]$, the state $x(\cdot,v)$ belongs to the boundary of the set $W(t)$. $W(1) = W$.

2. The Fundamental Lemma.

The fundamental lemma states that:

If there is no hyperplane seperating the convex sets S^+ and $\tilde{W}$, then the sets S^+ and W have at least one point in common.

For convenience three new sets S^+_*, W_* and $\tilde{W}_*$ are introduced.

$$S^+_* \triangleq \{ x : x + x(1,v) \in S^+ \}$$
$$W_* \triangleq \{ x : x + x(1,v) \in W \}$$

$$\tilde{W}_* \triangleq \{ x : x + x(1,v) \in \tilde{W} \}.$$

In other words, S^+_*, W_*, $\tilde{W}_*$ are translations of S^+, W, $\tilde{W}$ along the vector $x(1,v)$, respectively. The sets S^+_*, W_*, $\tilde{W}_*$ can be equivalently expressed as follows:

$$S^+_* = \{ x : x_i = 0,\ i = 1,\ldots,r \text{ and } x^n > 0 \}$$
$$W_* = \{ y(1,u) : u \in U \}$$
$$\tilde{W}_* = \{ z(1,u) : u \in U \}.$$

The fundamental lemma stated above is then equivalent to the modified fundamental lemma which states:

If there is no hyperplane seperating the convex sets S^+_* and $\tilde{W}_*$, then the sets S^+_* and W_* have at least one point in common.

Halkin proves this initially for the case $r = n-1$, making use of the Balayage Theorem and Brouwer's Fixed point Theorem. The proof is then extended to the general case $r < n-1$ by considering S^+_π, W_π, $\tilde{W}_\pi$, the projections of the sets S^+_*, W_*, $\tilde{W}_*$ on the $(r+1)$-dimensional space R^{r+1} obtained by taking the dimensions $1, 2, \ldots, r$ and n of the original Euclidean space R^n. This procedure is due to Warga [61].

As a corollary to the fundamental lemma we have:

If the sets S^+ and W have no points in common then there is a hyperplane which seperates the convex sets S^+ and $\tilde{W}$.

3. The Maximum Principle.

Let $v \in U$ be optimal. Then there exists a vector valued function $\lambda: T \to R^n$, defined, non-zero and absolutely continuous satisfying the differential equation:

(i) $\dot{\lambda}(t) = -A(t)\lambda(t) \qquad \forall\ t\in\theta(v)$

with the transversality conditions

(ii) $\lambda^i(1) = 0 \quad i = r+1,\ldots,n-1$

(iii) $\lambda^n(1) \geq 0,$

and, furthermore

(iv) $\langle\phi(t,v),\lambda(t)\rangle \geq \langle\phi(t,u),\lambda(t)\rangle \qquad \forall\ t\in\theta(v),\ \forall\ u\in U.$

By construction, S^+ and $\tilde{W}$ have no point in common. Therefore, by the above corollary there is a hyperplane seperating S^+ and $\tilde{W}$. This means that we can choose some non-zero $c\in R^n$ satisfying the following conditions of seperation:

$$\langle x,c\rangle \leq h \qquad \forall\ x\in\tilde{W}$$
$$\langle x,c\rangle > h \qquad \forall\ x\in S^+.$$

If we denote by $\bar{S}^+$ the closure of the set S^+, then the relation may be strengthened to

$$\langle x,c\rangle \geq h \qquad \forall\ x\in\bar{S}^+,$$

and since $x(1,v) \in \bar{S}^+\cap\tilde{W}$, then

$$\langle x(1,v),c\rangle = h.$$

We now verify that c satisfies the transversality conditions. If e_i is the unit vector in the direction of the i^{th} axis, then for $i = r+1,\ldots,n-1$ we have

$$(x(1,v)\pm e_i) \in \bar{S}^+.$$

Therefore, $\quad \langle(x(1,v)\pm e_i),c\rangle \geq \langle x(1,v),c\rangle.$

This implies that $\langle e_i, c\rangle = 0$, and hence $c_i = 0$, $i = r+1,..,n-1$, establishing condition (ii). Also, since

$$(x(1,v)+e_n) \in \bar{S}^+,$$
$$\langle (x(1,v)+e_n), c\rangle \geq \langle x(1,v), c\rangle.$$

Therefore, $\langle e_n, c\rangle \geq 0$, and $c^n \geq 0$, establishing (iii).

Define $\lambda: T \to R^n$ as the solution to the linear matrix differential equation (i) with the boundary condition

$$\lambda(1) = c$$

where c is the seperating hyperplane normal introduced above. Since $c \neq 0$, from the theory of linear differential equations, $\lambda(t) \not\equiv 0$, and λ is absolutely continuous on $[0,1]$.

To complete the proof, condition (iv) is verified using a contradiction argument. Assume (iv) is not satisfied. Then for some $t' \in \theta(v)$, there exists a $\bar{u} \in \Omega$ such that

$$\langle \phi(v(t'),t'), \lambda(t')\rangle < \langle \phi(\bar{u}(t'),t'), \lambda(t')\rangle.$$

Then by the continuity of $\phi(v(\cdot),\cdot)$ at t' and λ, there exists an $\eta > 0$ and $\varepsilon > 0$ satisfying

$$\langle \phi(v,\tau), \lambda(\tau)\rangle < \langle \phi(\bar{u},\tau), \lambda(\tau)\rangle - \eta, \text{ for } |t'-\tau| < \varepsilon.$$

Define control u^* by

$$u^*(t) = \bar{u}(t) \quad \text{for all } t \text{ such that } |t'-t| < \varepsilon$$
$$= v(t) \quad \text{otherwise.}$$

Therefore, for $u^* \in U$ we have

$$\langle x(1,v) + z(1,u^*), c\rangle = \langle (x(1,v) + \int_0^1 \phi(1,\tau)\phi(u^*,\tau)d\tau), c\rangle$$

$$= \langle x(1,v), c\rangle + \int_0^1 \langle \phi(u^*,\tau), \lambda(\tau)\rangle d\tau$$

$$\geq \langle x(1,v), c\rangle + \int_0^1 \langle \phi(v,\tau), \lambda(\tau)\rangle d\tau + 2n\varepsilon$$

$$= \langle x(1,v), c\rangle + 2n\varepsilon \qquad > h.$$

where ϕ is the transition matrix associated with the linearized system (6.2). Hence we obtain a contradiction that

$$(x(1,v) + z(1,u^*)) \in \tilde{W}.$$

Thus condition (iv) is satisfied. Since

$$\phi(t,v) = 0$$

and $$\phi(t;u) = f(x(t,v),u,t) - f(x(t,v),v,t)$$

defining $$H(x,u,\lambda,t) = \langle f(x,u,t), \lambda(t)\rangle$$

condition (i) - (iv) translate into the conditions of the Maximum Principle for nonlinear systems.

Bryant and Mayne [13] give a proof of the Maximum Principle similar to that given by Halkin [27]. Their main objective was to replace the use of Brouwer's Fixed Point Theorem by an easily proven contraction mapping theorem.

The first use of Brouwer's fixed point theorem in [27] is to prove the convexity of the reachable set of the linearized system (6.1), $\tilde{W}$. This is replaced by a simple constructive proof of Liapunov's theorem. This requires a slightly more general class of admissible controls than the piecewise continuous class employed by Halkin; the class of extended piecewise continuous controls and functions. They first prove that:

If $f \in P^n$ (the class of extended piecewise continuous functions, i.e. discontinuous on a countable set), then for any scalar k satisfying $0 \leq k \leq 1$ there exists a set $A \subset \mathcal{A}$, the set of all subsets of $[0,1]$, such that

$$k \int_0^1 f(\tau)\,d\tau = \int_0^1 f(\tau)\chi_A(\tau)\,d\tau$$

where $\chi_A = 1 \quad t \in A$

$= 0$ otherwise.

This represents a simplified form of Liapunov's Theorem and enables them to prove that $\tilde{W}$ is convex.

The second application of Brouwer's fixed point theorem is in proving a fundamental lemma, making use of the Balayage theorem. This is eliminated by a strengthened Balayage result and the use of an additional property of the linearized model of the nonlinear system, permitting a contraction mapping theorem to be employed.

The Interval Set (D_N^λ).

Given $N > 0$, partition the interval $[0,1]$ into N equal intervals I_j $(j = 1,\ldots,N)$. Within each I_j select a subinterval I_j^λ say, of length λ/N $(0 \leq \lambda \leq 1)$. Define

$$D_N^\lambda = \bigcup_{j=1}^{N} I_j^\lambda.$$

Using this partition, Bryant and Mayne prove:

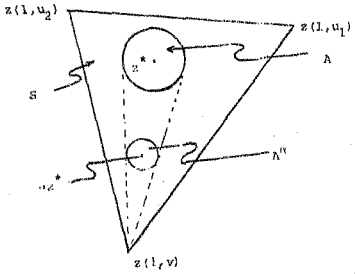
Strengthened Balayage Theorem : For any admissible u and any $\varepsilon > 0$, there exists an $N < \infty$ such that, $\forall\, \lambda \in [0,1]$, $j > N$,

$$\|\lambda z(1,u) - \int_{D_j^\lambda} \Phi(1,\tau)\phi(\tau,u)\,d\tau\| < \varepsilon\lambda.$$

Proof of the Fundamental Lemma.

Suppose $\tilde{W}$ is of dimension n. If not, project S^+ onto the linear manifold containing $\tilde{W}$ and proceed with the construction in the reduced dimension space. If $\tilde{W}$ and S^+ are not seperated, then there exists at least one point, say z^*, contained in $r\,int(S^+) \cap int(W)$. Since $z(1,v) \in \tilde{W}$ and $z^* \in \tilde{W}$, there exists a simplex S with vertices $z(1,v)$, $z(1,u_1)$, ..., $z(1,u_n) \in \tilde{W}$ containing z^* as an interior point. Select a closed sphere A int S with centre z^* and let η denote its radius. Then for $0 < \alpha < 1$ we can define a closed sphere A^α with centre αz^* and radius $\alpha\eta$, where $A^\alpha \subset$ int S by convexity.

Figure 6.1



Now, given $a \in A^\alpha$, any integer N, Bryant and Mayne construct an approximate control $u^{a,N}$ associated with the point a and the integer N. Using the strengthened Balayane theorem, it can be shown that for any $\varepsilon > 0$, there exists an $N_0 < \infty$ such that

$$\|a - z(1,u^{a,N})\| \leq \varepsilon\alpha$$

for all $a \epsilon A^{\alpha}$ and $N \geq N_o$.

Consider a mapping $h^N: R^n \to R^n$ defined as follows:

$$h^N(a) \triangleq a - y^{a,N} - z^*,$$

where $y^{a,N} \triangleq x(1,u^{a,N}) - x(1,v)$.

Then using the strengthened Balayage theorem, as well as an interesting property of the linearized model (linearized in state), they show that an α and N can be chosen such that h is a contraction mapping† in the closed set A^{α} and mapping A^{α} into itself. Hence there exists an $a \epsilon A^{\alpha}$ satisfying

$$h(a) = a,$$

implying $$y^{a,N} = \alpha z^*.$$

Since W and S^+ are assumed not separated, then $z^* \epsilon \text{rint}(S^+)$. Also, $0 \epsilon S^+$. Therefore, $\alpha z^* \epsilon \text{rint}(S^+)$ for $\alpha \epsilon [0,1]$, and hence $\alpha z^* \epsilon S^+$. But $\alpha z^* = y^{a,N} \epsilon W$. Since S^+ is by definition disjoint from W, this is a contradiction, proving the fundamental lemma.

The rest of the proof of the Maximum Principle follows the same approach as the preceding approach of Halkin's.

The derivation by Bryant and Mayne represents a very useful advance on the derivation due to Halkin, [27]. Not only are the class of extended functions still of practical interest to the engineer, but the mathematical content is a good deal simpler.

---

† A simple form of the standard contraction mapping theorem states that if a function H maps a closed subset A of R^n into itself, and there exists an α $(0 \leq \alpha < 1)$ such that $\|H(x)-H(\bar{x})\| \leq \alpha\|x-\bar{x}\|$, for all $x, \bar{x} \epsilon A$, then there exists an unique point $\bar{x} \epsilon A$ which is a fixed point of H.

7. CONCLUSION.

The purpose of this essay has been to give a broad outline of the development of the necessary conditions that a process $(x(\cdot),v),v(\cdot))$ has to satisfy in order to be optimal for a certain class of problems. The emphasis in this presentation has been to give an 'intuitive understanding' of the approaches used by various authors to obtain necessary conditions, rather than an attempt at mathematical completeness. Of the major approaches covered, the proof by Bryant and Mayne was the least demanding mathematically, and yet still applied to a fairly general class of problems. The approach of Halkin, [27], and the epsilon approach of Balakrishnan have been used to formulate computational techniques to solve for the optimal control in [28] and [4] respectively.

The Maximum Principle is a set of necessary conditions for the optimality of a control. Suppose for some control problem we have found a control and a response satisfying the conditions of the Maximum Principle. Is this control optimal? More generally, when are the Maximum Principle's necessary conditions also sufficient? Some answers to these two questions have been given by Lee [39] and Mangasarian [43].

As can be expected, there exist other necessary conditions for optimal controls other than the Maximum Principle. For example, Vincent and Goh [60] obtain a new necessary condition to be satisfied at the endpoints of the trajectory.

In conclusion we remark that the Maximum Principle has been extended to a wider class of problems where the dynamics considered are often other than ordinary differential equations.

REFERENCES.

1. Athans, M., and P.L. Falb, Optimal Control, McGraw-Hill, 1966.

2. Balakrishnan, A.V., 'On a new computing technique in optimal control', SIAM J. Control, May 1968.

3. Balakrishnan, A.V., 'On a new computing technique in optimal control and the maximum principle', Proc. Nat. Acad. Science, Feb. 1968.

4. Balakrishnan, A.V., 'On a new computing technique in optimal control and its applications to minimum time flight profile optimization, J. Opt. Theory and Appl., July 1969.

5. Balakrishnan, A.V., 'The epsilon technique : A constructive approach to optimal control', in Control Theory and the Calculus of Variations, (ed. A.V. Balakrishnan), Academic, 1969.

6. Baum, R.F., and L. Cesari, 'On a recent proof of Pontryagin's necessary conditions', SIAM J. Control, 10, 56 - 75, 1972.

7. Bellman, R., Dynamic Programming, Princeton Univ. Press, Princeton, 1957.

8. Bellman, R., I. Glicksberg and O. Gross, 'On the bang-bang control problem', Quart. Appl. Math., 14, 11 - 18, 1956.

9. Berkovitz, L.D., 'Variational methods in problems of control and programming', J. Math. Anal. Appl., 3, 145 - 169, 1961.

10. Blackwell, D., 'The range of certain vector integrals', Amer. Math. Soc. Proc., 2, 390 - 395, 1951.

11. Boltyanskii, V.G., Mathematical Methods of Optimal Control, transl. from Russian by K.N. Trirogoff; Holt, Rinehart and Winston, New York - London, 1971.

12. Boltyanskii, V.G., R.V. Gamkrelidze and L.S. Pontryagin, 'The theory of optimal processes', Izv. Akad. Nauk SSSR Ser. Math., 24, 3 - 42, 1960.

13. Bryant, G.F., and D.Q. Mayne, 'The maximum principle', Pub. No. 73/16, Dept. Comp. Control, Imperial College, June 1973.

14. Bushaw, D.W., 'Differential equations with a discontinuous forcing term, Experimental Towing Tank', Report. No. 469, Stevens Inst. Tech., New Jersey, Jan. 1953.

15. Coddington, E.A., and N. Levinson, Theory of Ordinary Differential Equations, McGraw-Hill, 1970.

16. Diliberto, S., 'The Pontryagin maximum principle', in Topics in Optimization (ed G. Leitman), Academic Press, 1967.

17. Fel'dbaum, A.A., 'Optimal processes in automatic control systems', Aut. Telemek., 14, No. 6, 712 - 728, 1953.

18. Gamkrelidze, R.V., 'Theory of time-optimal processes for linear systems', Izv. Akad. Nauk. SSSR Ser. Math., 22, 449 - 474, 1958.

19. Gamkrelidze, R.V., 'On theory of optimal processes in linear systems', Doklady Akad. Nauk SSSR, 116, No. 1, 1957.

20. Gamkrelidze, R.V., 'On some extremal problems in the theory of differential equations with applications in the theory of optimal control', J. Soc. Ind. Appl. Math. Ser. A Control, 3, 106 - 128, 1965.

21. Halkin, H., 'The principle of optimal evolution', Int. Symp. on Nonlinear D.E. and Nonlinear Mechanics (ed J.P. La Salle, S. Lefschetz), Academic Press, 1961.

22. Halkin, H., 'On a generalization of a theorem of Liapunov', J Math. Anal. Appl., 10, 325 - 329, 1965.

23. Halkin, H., 'A generalization of La Salle's bang-bang principle', SIAM J. Control, 2, 199 - 202, 1964.

24. Halkin, H., 'Liapunov's theorem on the range of a vector measure and Pontryagin's maximum principle', Arch. Rational Mech. Anal., 10, No. 4, 296 - 304, 1962.

25. Halkin, H., 'On the necessary conditions for optimal control of nonlinear systems', J. D'analyse Math., 12, 1 - 82, 1964.

26. Halkin, H., 'Some further generalizations of a theorem of Liapunov', Arch. Rational Mech. Anal., 17, 272 - 277, 1964.

27. Halkin, H., 'Mathematical foundations of system optimization', in Topics in Optimization (ed. G. Leitman), Academic, 1967.

28. Halkin,H., 'Method of convex ascent', in Computing Methods in Optimization Problems (ed. A.V. Balakrishnan and L.W. Neustadt), Academic Press, 211 - 239, 1964.

29. Halmos, P.R., 'The range of a vector measure', Bull. Am. Math. Soc., 54, 416 - 421, 1948.

30. Hestenes, M.R., Calculus of Variations and Optimal Control Theory, Wiley, New York, 1966.

31. Horn, E.J., The Maximum Principle of Pontryagin in Optimal Control Theory, M.Sc. Diss., Univ. of South Africa, 1974.

32. Krasovskii, N.N., 'Concerning the theory of optimal control', Aut. Telemek. 18, No.11, 960,- 970, Nov. 1957.

33. Krasovskii, N.N., 'On one optimal control problem', Prikl. Mat. Mekh., 21, No. 5, 670 - 677, Oct. 1957.

34. Krasovskii, N.N., 'On a problem of optimal control of nonlinear systems, Prikl. Mat. Mekh., 23, No. 2, 209 - 229, Apr. 1959.

35. La Salle, J.P., 'Study of the basic principle underlying the bang-bang servo', Goodyear Aircraft Corp., Rep. CER-5518, 1953.

36. La Salle, J.P., 'Time optimal control systems', Proc. Nat. Acad. Sciences, 45, 573 - 577, 1959.

37. La Salle, J.P., 'The time-optimal control problem', in Contributions to the Theory of Nonlinear Oscillations, Vol V, 1959.

38. La Salle, J.P., 'The bang-bang principle', Proc. 1st Intern. Cong. of Inter. Feder. Aut. Control, Moscow, 493 - 497, 1960.

39. Lee, E.B., 'A sufficient condition in the theory of optimal control', SIAM J. Control, 1, 241 - 245, 1963.

40. Lee, E.B., and L. Markus, 'Optimal control for nonlinear processes', Arch. Rational Mech. Anal., 8, 36 - 58, 1961.

41. Lee, E.B., and L. Markus, Foundations of Optimal Control Theory, Wiley, 1967.

42. Liapunov, 'Sur les fonctions-vecteurs completement additives', Izv. Akad. Nauk SSSR Ser Mat., 8, 465 - 478, 1940.

43. Mangasarian, O.L., 'Sufficient conditions for the optimal control of nonlinear systems', SIAM J control, 4, 1966.

44. Mc Shane, E.J., 'On multipliers for Lagrange problems', Amer. J. Math., 61, 809 - 319, 1939.

45. Mc Shane, E.J., 'Optimal controls, relaxed and ordinary', in Mathematical Theory of Control (ed A.V. Balakrishnan and L.W. Neustadt), Academic Press, 1967.

46. Mc Shane, E.J., 'Relaxed controls and variational problems', SIAM J Control, 5, 438 - 485, 1967.

47. Mersky, J.W., 'Constrained Optimal control problems', 5th Conference on Opt. Techn., Part 1, Springer-Verlag, 1973.

48. Mersky, J.W., An Application of the Epsilon Technique to Control Problems with Inequality Constraints, Disser., Univ. Calif., Los Angeles, 1973.

49. Neustadt, L.W., 'The existence of optimal controls in the absence of convexity conditions', J. Math. Anal. Appl., 7, 110 - 117, 1963.

50. Pontryagin, L.S., 'Optimal processes of regulation', Proc. Int. Math. Conf., Edinburgh, 1958.

51. Pontryagin, L.S., 'Optimal control processes', Uspekh Mat. Nauk, 14, No. 1, 3 - 20, 1959.

52. Pontryagin, L.S., V.G. Boltyanskii, R.V. Gamkrelidze and E. Mischenko, The Mathematical Theory of Optimal Processes, Interscience (Wiley), New York, 1962.

53. Roxin, E., 'Reachable zones in autonomous differential systems, Bol. Soc. Mat. Mex., 125 - 135, 1960.

54. Roxin, E., 'On the existence of optimal controls', Michigan J. Math, 9, 109 - 119, 1962.

55. Roxin, E., 'A geometric interpretation of Pontryagin's maximum principle', Int. Symp. on Nonlinear D.E. and Nonlinear Mechs (ed J.P. La Salle and S Lefschetz), Academic Press, 1961.

56. Rozoncer, L.I., 'Pontryagin's maximum principle in the theory of optimum systems', Aut. Telem., 20; Part.I, No. 10, 1320 - 1334; Part II, No. 11, 1441 - 1458; Part III, No. 12, 1561 - 1578, 1959.

57. Strauss, A., An Introduction to Optimal Control Theory, Lecture notes in Operations Research and Math. Econ., 3, (ed. M. Beckmann and H.P. Künzi), Springe-Verlag, 1968.

58. Valentine, F.A., 'The problem of Lagrange with differential inequalities as added side conditions', in Contributions to the Calculus of Variations, 407 - 448, Univ. of Chicago Press, 1937.? 1937.

59. Varaiya, P.P., Notes on Optimization, Notes on System Sciences, Van Nostrand Reinhold, New York, 1972.

60. Vincent, T., and B.S. Goh, 'Terminality, normality and transversality conditions', J. Opt. Theory Appl., 9, 32 - 50. '2.

61. Warga, J., 'Relaxed variational problems', J. Math. Anal. Appl., 4, No. 1, 111 - 142, 1962.

62. Warga, J., Optimal Control of Differential and Functional Equations, Academic Press, 1972.

63. Wierzbicki, A., 'Maximum principle for semiconvex performance functionals', SIAM J. Control, 10, 444 - 459, 1972.

64. Young, L.C., Lectures on the Calculus of Variations and Optimal Control Theory, Saunders, 1969.

THE CONSTRUCTION OF LIAPUNOV FUNCTIONS FOR AUTONOMOUS SYSTEMS.

Kendal Clive Jordt

THE CONSTRUCTION OF LIAPUNOV FUNCTIONS FOR AUTONOMOUS SYSTEMS.

Kendal Clive Jordi

An Essay Submitted to the Faculty of Science, in partial fulfilment of the requirements for the degree of Master of Science.

University of the Witwatersrand,
Johannesburg.
1975.

ABSTRACT.

The theory of stability in the sense of Liapunov for autonomous (explicit time independence) systems is developed. Analytic methods for constructing Liapunov functions are outlined. In particular, the methods of Aizerman, Szego and Lur'e and Letov, the variable gradient method of Schultz and Gibson, and the Format method due to Peczkowski are discussed.

INTRODUCTION.

Though originally introduced near the end of the last century, the second method of Liapunov may be broadly classified as one of the 'modern' approaches to the solution of stability problems in automatic control.

The direct, or second method of Liapunov attempts to make statements on the stability of the equilibrium state without any knowledge of the solutions of the differential equations, i.e., without using the explicit form of the perturbed motions. A function V, called the Liapunov function, is examined as a function of time. To prove stability, it is sufficient to show that this V function approaches zero as time approaches infinity. The required V function is always available for linear systems, but is sometimes difficult to find for nonlinear systems.

The purpose of this essay is to outline the most important analytic means that are now available for generating the required Liapunov function for an autonomous system. This essay is based to a large extent on the article by Schultz in [22].

The first two chapters give the necessary background, definitions and theorems. The choice of the coordinate system in which the system equations are expressed is discussed in the third chapter. Also covered are the Routh-Hurwitz stability criteria for linear systems, the transfer function, and systems with a single nonlinearity.

The section on the actual generation of Liapunov functions for autonomous systems comprises eight major divisions. In the introductory section we discuss, very briefly, the Zubov construction procedure.

However, this approach is largely numerical, and is not covered any further. Linear systems are then considered, followed by a general approach based on norm type V functions. Although this is rather impractical, the discussion serves to demonstrate that this general approach is the theoretical justification for the usual procedure of linearization about an operating point. Aizerman's approach is then shown to be an excellent method, that is not only simple, but often gives good results. Following this, the procedure developed by Szego is outlined, and this is shown to extend and compliment the Aizerman approach. The well-known methods of Lur'e and the very important variable gradient method of Schultz and Gibson are then described. The last method covered is the Format method due to Paczkowski.

1. SYSTEM REPRESENTATION AND DEFINITIONS.

SYSTEM REPRESENTATION.

Consider the system described by the equation

$$\dot{x}(t) = f(x(t),t), \quad x(t_o) = x_o. \tag{1.1}$$

$x(t) \in R^n$, $f : R^n \times T \to R^n$ a nonlinear function of its arguments. A solution at time t with given initial condition x_o at t_o is denoted by $x(t;x_o,t_o)$ or simply $x(t)$ if no confusion can arise. The whole solution (i.e. whole function) is denoted $x(\cdot,x_o,t_o)$ or simply $x(\cdot)$.

Definition 1.1: The system described by equation (1.1) is called stationary or autonomous if f does not depend explicitly on t.

In this essay we are interested in autonomous systems. The trajectories of an autonomous system are invariant under a translation of time.

i.e. $\quad x(t;x_o,t_o) = x(t+T;x_o,t_o+T) \qquad \forall\, t, x_o, t_o, T.$

Definition 1.2: If $f(x_e(t)) = 0$, $\forall\, t$, then $x(t;x_e,t_o) = x_e$ for any t_o and x_e is called an equilibrium solution.
If x_e is equal to zero, $x(t;x_e,t_o)$ is called the null solution.
x_e is called the equilibrium state.

Definition 1.3: A solution $x(\cdot;x_o,t_o)$ is said to have a finite escape time if $||x(t;x_o,t_o)|| \to \infty$ as $t \to t_e < \infty$.

We assume that $x(t)$ has unique solution defined for all

$t \in [t_0, \infty)$ for each $x_0 \in R^n$, (i.e. there is no escape time). Sufficient conditions for this are :

(i) f is linear in x,

(ii) f is continuously differentiable in x.

DEFINITIONS OF STABILITY.

The concept of stability for general time-dependent non-linear systems is very complex. A large number of definitions exist; only the most useful ones, with reference to autonomous systems, will be discussed in this section. These definitions deal with the stability of an equilibrium or a fixed motion, with respect to the initial conditions, and were discussed rigorously by the Russian mathematician A. Liapunov. The definitions given here follow Willems [26].

Let x_e be an equilibrium state of the free dynamic system

$$\dot{x}(t) = f(x(t)) \tag{1.2}$$

with

$$f(x_e(t)) = 0 \quad \text{for all } t. \tag{1.3}$$

Definition 1.4: The equilibrium state x_e or the equilibrium solution $x(t) = x_e$, $t \in [t_0, \infty)$ is called stable, if for any given t_0 and $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that

$$\|x_0 - x_e\| < \delta \Rightarrow \|x(t;x_0,t_0) - x_e\| < \varepsilon, \quad \forall\, t \geq t_0.$$

Definition 1.5: The equilibrium state x_e is called convergent if for any t_0 there exists δ_1 such that

$$\|x_0 - x_e\| < \delta_1 \Rightarrow \lim_{t \to \infty} x(t;x_0,t_0) = x_e, \quad \forall\, t_0.$$

Definition 1.6: The equilibrium state x_e is called asymptotically stable if it is convergent and stable.

Definition 1.7: The equilibrium state is called bounded if there exists $B(x_o)$ such that

$$||x(t;x_o,t_o)|| \leq B, \quad \forall\ t \geq t_o.$$

Definition 1.8: The equilibrium state is called asymptotically stable in the large if :

(i) it is stable,

(ii) it is bounded,

(iii) it is convergent.

The definitions given above are usually referred to as uniform. However, all stability properties of autonomous systems are uniform, so we omit reference to the 'uniform' for brevity. The definitions that we have given are compatible with the definitions given by La Salle [7], for example.

POSITIVE DEFINITE FUNCTIONS.

Let $V(x)$ be a real scalar function of the vector x and let S be a closed bounded region in R^n containing the origin.

Definition 1.9: The function $V(x)$ is positive semi-definite in S if, for all $x \in S$,

(i) $V(x)$ has continuous partial derivitives w.r.t the components of x,

(ii) $V(0) = 0$,

(iii) $V(x) \geq 0$.

Definition 1.10: The positive semi-definite function $V(x)$ above is positive definite in S if condition (iii) of Definition 1.9 is strengthened to

(iii) $V(x) > 0, \quad x \neq 0.$

Condition (iii) of Definition 1.10 can be replaced by (See Willems [26]),

(iii)' $V(x) \geq \Psi(r)$, for all $x \in S$, where $r \triangleq ||x||$, and the function $\Psi(r)$ is continuous and strictly increasing, and $\Psi(0) = 0$.

Example: $x \in R^2$.

$V(x) = \frac{1}{2}x_1^2$ is a positive semi-definite function

$V(x) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2$ is a positive definite function.

When V is in quadratic form, expressible as

$$V(x) = x^T A x \tag{1.4}$$

where A is a symmetric, square matrix with constant coefficients, the usual means of determining the definiteness of the form is through the application of Sylvester's Theorem.

Sylvester's Theorem: In order that the quadratic form of the equation (1.4) be positive definite, it is necessary and sufficient that the determinants of the principle minors, that is, the magnitudes

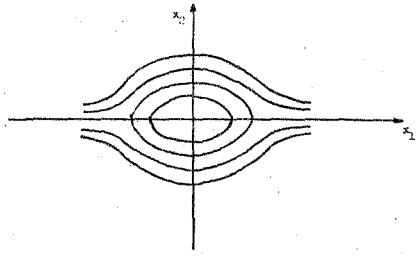
$$|a_{11}|, \quad \begin{vmatrix} a_{11}, a_{12} \\ a_{12}, a_{22} \end{vmatrix} \text{ and } \begin{vmatrix} a_{11}, a_{12}, \cdots, a_{1n} \\ a_{12}, \cdots \\ \vdots \\ a_{1n}, \cdots \quad a_{nn} \end{vmatrix}$$

be positive.

Definition 1.11: A scalar function is radially unbounded if, as $||x|| \to \infty$, $V(x) \to \infty$, i.e. for any given $M > 0$, there exists $N > 0$ such that $V(x) > M$, $\forall\, x$ such that $||x|| > N$.

Closely allied to the concept of definiteness and radial unboundedness, is the concept of a simple closed curve or surface. A surface enclosing the origin is simple if it does not intersect itself, and closed if it intersects all paths that lead from the origin to infinity. That is, a simple closed surface is topologically equal to the surface of an n-dimensional sphere. If V is a positive definite function, then equations $V = K_1$, $K_1 < K_2 < K_3 \ldots$ represent a set of nested closed surfaces about the origin in a sufficiently small region. In order to ensure that the region extends to infinity, it is necessary to ensure that the curve $V = K$ is closed for sufficiently large K. The closure of the curves $V = K$ is assured if we require $V(x)$ to be positive definite and radially unbounded, See Figure 1.1

FIGURE 1.1. The curves $V(x) = \text{constant}$ for a positive definite function which is not radially unbounded.



An alternative method of investigating the region in which $V = K$ remains closed is to examine the gradient of V, which is defined as

$$\nabla V = \begin{bmatrix} \partial V/\partial x_1 \\ \cdot \\ \cdot \\ \cdot \\ \partial V/\partial x_n \end{bmatrix}$$

As long as ∇V is not zero anywhere in a region Ω containing the origin, except at the origin, then $V = K$ represents a closed surface in Ω. If ∇V is zero only at the origin, then Ω includes the whole space and the function is radially unbounded.

As an example of a curve that is positive definite, and yet closed only for values of $K < 1$, Letov [9] considers the function

$$V = x_1^2 + \frac{x_2^2}{1 + x_2^2} .$$

Here, $V(x)$ is not radially unbounded. Also, the gradient is zero not only at $x_1 = x_2 = 0$, but at $x_1 = 0$, $x_2 = \infty$.

2. LIAPUNOV STABILITY THEOREMS.

In this section we consider the stability properties of the null solution to the autonomous system

$$\dot{x}(t) = f(x(t)) \tag{2.1}$$

with

$$f(0) = 0.$$

A large number of theorems exist which are related to the second method of Liapunov; for example, Donalson [2] lists 32. The original theorems due to Liapunov, Theorems 2.1, 2.2 are applicable only to arbitrarily small regions about the origin.

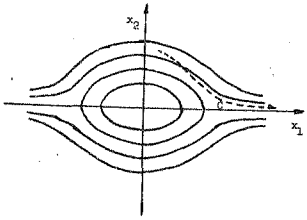
Theorem 2.1 [26]: The null solution, or the equilibrium state at the origin of system (2.1), is stable if there is some neighbourhood S of the origin where a positive definite function $V(x)$ exists such that its derivitive $\dot{V}(x)$ w.r.t. the solutions of (2.1) is negative semi-definite in that region.

Theorem 2.2 [12],[26]: The null solution of system (2.1) is asymptotically stable, if in some neighbourhood S of the origin there is a positive definite function $V(x)$ such that its derivitive $\dot{V}(x)$ is negative definite in that region.

Theorem 2.3 [26]: If the conditions of Theorem 2.2 are satisfied for all x, and if $V(x)$ is radially unbounded, then the null solution of (2.1) is asymptotically stable in the large.

The condition that $V(x)$ be radially unbounded, to obtain global asymptotic stability, cannot be waived, as is illustrated in Fig. 2.1.

FIGURE 2.1.



$V(x)$ is positive definite, but not radially unbounded. Although it decreases along the trajectory C, the motion is unbounded.

The above theorems are overly restrictive. The requirement that dV/dt be negative definite, rather than semi-definite, causes difficulties when we attempt to generate suitable Liapunov functions for nonlinear systems. This shortcoming was overcome by La Salle [7], [8].

Theorem 2.4: The null solution of system (2.1) is asymptotically stable if, in some neighbourhood S of the origin, there is a positive definite function $V(x)$ such that its derivitive $\dot{V}(x)$ along the solutions of (2.1) is negative semi-definite in S, and such that $\dot{V}(x)$ does not vanish identically along any solution of (2.1) in S, other than the null solution.

Theorem 2.5: The null solution of system (2.1) is asymptotically stable in the large if the assumptions of Theorem 2.4 hold in the entire space, and if $V(x)$ is radially unbounded.

Ingwerson [6] has proved a theorem which is closely related to Theorems 2.4 and 2.5.

Theorem 2.6: The null solution of the system (2.1) is asymptotically stable in some closed region S containing the origin, if there exists a positive definite function $V(x)$ in S, such that

(i) one of the surfaces $V = K$ bounds S,

(ii) the gradient of V, ∇V is not zero anywhere in S, except at $x = 0$,

(iii) dV/dt is negative or negative semi-definite in S,

(iv) dV/dt is not identically zero on a solution of the system other than the null solution.

The conditions of the above Theorem are stronger than the conditions of Theorem 2.4, but will enable us to get some idea of the size of the region of asymptotic stability. This will prove useful later on. If ∇V does not equal zero except at $x = 0$, then the region S extends to the entire space, and Theorems 2.5 and 2.6 are equivalent.

In order to ensure that the solution to the equation $dV/dt = 0$ is not also a solution of equation (2.1), it is only necessary to substitute the solution of this equation back into equation (2.1). In practice this is often a trivial problem. One note of caution: in cases where the original system has more than one equilibrium point and dV/dt is semi-definite, the existance of the other equilibrium points is easily overlooked, and a wrong conclusion is reached.

THEOREMS ON INSTABILITY.

Liapunov's technique also yields theorems on the instability of equilibrium states of dynamic systems. These theorems are of limited importance, since one is almost always interested in determining stability properties. However, they are sometimes useful to avoid a waste of effort trying to prove stability. The following theorems are examples of instability criteria:

Theorem 2.7 [26]: The null solution of the autonomous system (2.1) is not asymptotically stable if there exists a scalar function $V(x)$ with the following properties in some closed neighbourhood S of the origin.

(i) $V(x)$ vanishes at the origin, has continuous partial derivitives, and assumes negative values arbitrarily close to the origin; and

(ii) the derivitive $\dot{V}(x)$ along the solutions of system (2.1) is negative semi-definite.

The null solution is unstable if, in addition, either $\dot{V}(x)$ is negative definite or does not vanish along any solution of (2.1), except the null solution.

If in the above theorem, the functions $V(x)$ and $\dot{V}(x)$ are negative definite in R, then the null solution is completely unstable.

Theorem 2.8 [26]: Suppose that there exists a scalar function $V(x)$ with continuous partial derivitives in some neighbourhood S of the origin, such that:

(i) $V(0) = 0$, and

(ii) $V(x)$ assumes negative values arbitrarily close to the origin.

If the derivitive of $V(x)$ along the solution of (2.1) can be expressed as

$$\dot{V}(x) = aV(x) + V_1(x)$$

where $V_1(x)$ is negative semi-definite in R, then

(i) the null solution of (2.1) is not asymptotically stable if a is a non-negative constant; and

(ii) the null solution of (2.1) is unstable if a is a positive constant.

3. AUTONOMOUS SYSTEMS.

N^{th} ORDER DIFFERENTIAL EQUATIONS AND LINEAR SYSTEMS.

Let a free, linear autonomous system be governed by an n^{th}-order homogeneous differential equation with constant coefficients

$$p(D)x = 0 \tag{3.1}$$

where D denotes the operator d/dt. Let

$$p(D) = D^n + p_{n-1}D^{n-1} + \ldots + p_0. \tag{3.2}$$

Then, with $x_1 = x$, the system can be described by the first-order vector differential equation

$$\dot{x} = Ax \tag{3.3}$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \\ -p_0 & -p_1 & -p_2 & \ldots & -p_{n-1} \end{bmatrix} \tag{3.4}$$

i.e.

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ &\vdots \\ \dot{x}_n &= -p_0x_1 - p_1x_2 \ldots - p_{n-1}x_n. \end{aligned} \tag{3.5}$$

This particular choice of state variables is referred to as the phase variables.

Theorem 3.1: The null solution of system (3.3) is asymptotically stable if and only if the polynomial $p(s)$ has only zeros with negative real parts.

Theorem 3.2: The null solution of system (3.3) is stable if and only if the polynomial $p(s)$ has no zeros with positive real parts, and if the zeros of $p(s)$ with zero real parts are simple.

THE TRANSFER FUNCTION.

Single input-single output systems are often described by their impulse response $w(t)$; i.e. the output to a unit impulse input $\delta(t)$ (Rosenbrock and Storey [19]) when the system is initially in the equilibrium state. For simplicity it is assumed that this equilibrium state corresponds to zero output. The output of a system with impulse response $w(t)$ to an arbitrary input $u(t)$ is then

$$y(t) = \int_{-\infty}^{+\infty} w(\tau)u(t-\tau)d\tau \qquad (3.6)$$

when the system is initially in its equilibrium state.

A very common concept for describing single input-single output systems is the transfer function. It is defined as the Laplace transform of the impulse response

$$G(s) = \int_{-\infty}^{+\infty} w(t)\exp(-st)dt \qquad (3.7)$$

where $s = \sigma + i\omega$ is a complex number. The transfer function is defined in the region of convergence of (3.7).

Theorem 3.3: A necessary and sufficient condition for the input-output stability of the equilibrium state of a single input-single output system is that the region of convergence of its transfer fuction includes the imaginary axis. If the system is non-anticipative (consequently the impulse response vanishes for negative t), then

$$G(s) = \int_0^{+\infty} w(t)\exp(-st)dt$$

and the input-output stability condition is that the transfer function has no singularities in the right half-plane or on the imaginary axis.

We have therefore that the transfer function is only a complete description of the system if the common factors of numerator and denominator are not cancelled. This can only be achieved if the transfer function is determined from the differential equation of the system. Only then can asymptotic stability of the system be checked by means of the transfer function. The common zeros of numerator and denominator are called the hidden modes of the system. The asymptotic stability and the input-output stability of a system are only equivalent if there are no hidden modes with positive or zero real parts.

It is apparent that a system should never be designed so that a right half-plane pole is cancelled mathematically by a right half-plane zero. This cancellation must either consist of the removal of the associated equipment, or must be achieved physically, for instance by means of feedback.

THE ROUTH HURWITZ STABILITY TESTS.

As has been shown in the previous sections, the condition for asymptotic stability of (3.1) is that the characteristic polynomial $p(s)$ has only zeros with negative real parts. Algebraic criteria have been developed independently by Routh [20] and Hurwitz [8].

The Hurwitz Stability Criterion.

Consider the polynomial

$$p(s) = p_n s^n + p_{n-1} s^{n-1} + \cdots + p_0.$$

If the zeros $z_1, z_2, \cdots, z_n$ of $p(s)$ have negative real parts, then all coefficients $p_0, p_1, \cdots, p_n$ have the same sign. Indeed

$$\frac{p_{n-1}}{p_n} = -\sum_i z_i,$$

$$\frac{p_{n-2}}{p_n} = +\sum_{i,j} z_i z_j,$$

$$\vdots$$

$$\frac{p_0}{p_n} = (-1)^n z_1 z_2 \cdots z_n$$

are positive. This condition is necessary but not sufficient. Consider the square matrix of the n^{th} order

$$H = \begin{pmatrix} p_{n-1} & p_{n-3} & p_{n-5} & \cdots & 0 \\ p_n & p_{n-2} & p_{n-4} & \cdots & 0 \\ 0 & p_{n-1} & p_{n-3} & \cdots & 0 \\ 0 & p_n & p_{n-2} & \cdots & 0 \\ 0 & 0 & p_{n-1} & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & & p_0 \end{pmatrix}$$

H is called the Hurwitz matrix. The index of the coefficients increases by one along a column and decreases by two along a row. The element h_{ij} is

$$h_{ij} = p_{n+1-2j} \quad \text{if } 0 \leq 2i - j \leq n,$$
$$= 0 \quad \text{if } 2i < j \text{ or } 2i - j > n.$$

The principal minors $D_1, D_2, \ldots, D_n$ are the determinants

$$D_k = \begin{vmatrix} p_{n-1} & p_{n-3} & \cdots & \\ p_n & p_{n-2} & \cdots & \\ 0 & p_{n-1} & \cdots & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ 0 & 0 & \cdots & p_{n+k} \end{vmatrix} \quad k = 1,2,\ldots,n.$$

They are called the Hurwitz determinants. A necessary and sufficient condition for the polynomial $p(s)$ to have only zeros with negative real parts is

$$(p_n)^k D_k > 0 \quad \text{for } k = 1,2,..,n.$$

This the Hurwitz stability criterion. We can assume in the sequel, without loss of generality, that p_n is positive. Indeed, if this were not true, then the polynomial $-p(s)$ should be considered, which has the same zeros as $p(s)$. Then the stability criterion states that the n Hurwitz determinants of $-p(s)$ should be non-zero and positive. For a proof of this criterion the reader should see Willems.

The Routh Stability Criterion.

A different, although equivalent (see Willems), procedure was developed by Routh [20]. It requires the evaluation of the following array, which is called the Routh array of the polynomial $p(s)$:

$$
\begin{array}{lll}
p_n & p_{n-2} & p_{n-4} \quad \cdots \\
p_{n-1} & p_{n-3} & p_{n-5} \quad \cdots \\
a_1 & a_2 & a_3 \quad \cdots \\
b_1 & b_2 & b_3 \quad \cdots \\
c_1 & c_2 & c_3 \quad \cdots \\
\vdots & \vdots & \vdots
\end{array}
$$

The first and the second row contain the coefficients of the odd and the even parts of $p(s)$. The elements of the third and following rows are computed by means of the following algorithm

$$
a_1 = \frac{p_{n-1}p_{n-2} - p_n p_{n-3}}{p_{n-1}}, \qquad a_2 = \frac{p_{n-1}p_{n-4} - p_n p_{n-5}}{p_{n-1}}, \qquad \cdots
$$

$$
b_1 = \frac{a_1 p_{n-3} - a_2 p_{n-1}}{a_1}, \qquad b_2 = \frac{a_1 p_{n-5} - a_3 p_{n-1}}{a_1}, \qquad \cdots
$$

$$
c_1 = \frac{a_2 b_1 - a_1 b_2}{b_1}, \qquad c_2 = \frac{a_3 b_1 - a_1 b_3}{b_1}, \qquad \cdots
$$

The total array contains $(n+1)$ rows.

The necessary and sufficient condition for $p(s)$ to have only zeros with negative real parts, is that all the elements of the first column of its Routh array have the same sign, and that none vanishes.

SINGLE - NONLINEARITY SYSTEMS.

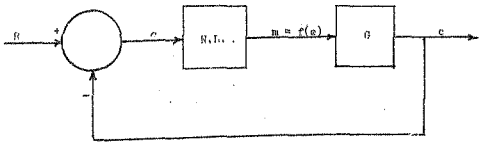


FIGURE 3.1. Single-nonlinearity system.

Consider the single-loop system as shown in Fig. 3.1. Gibson comments that this is not a limited special case. Any system with any number of loops and interconnections but with only one nonlinearity may be reduced to this form.

ALTERNATIVE CHOICE OF SYSTEM VARIABLES.

For systems of order higher than two, the output and its n-1 derivatives, i.e. the phase variables, are often a poor choice for system state variables. It is shown here that the choice of more natural system variables overcomes these disadvantages. Consider the system based on Fig. 3.2. Note that the example chosen includes a zero in the forward transfer function. Choice of phase variables as the particular set of phase variables, as in Fig. 3.2 (b), leads to a second-order equation involving the derivative of the nonlinearity, as

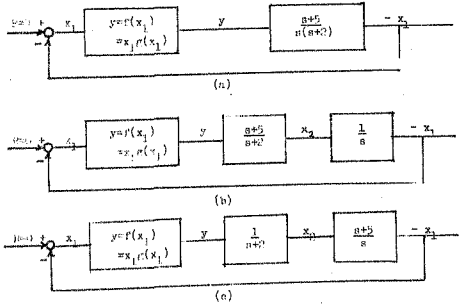


FIGURE 3.2.

$$\ddot{x}_1 + (2 + x_1 \frac{\partial g(x_1)}{\partial x_1} + g(x_1))\dot{x}_1 + 5g(x_1)x_1 = 0. \tag{3.8}$$

The basic difficulty with phase variables in this example can be anticipated from a knowledge of linear systems. If this problem is thought of in terms of the usual Routh-Hurwitz conditions for linear systems, the coefficient of $\dot{x}_1$ must always be greater than zero. To avoid terms involving the slope of the nonlinearity from appearing as a coefficient of $\dot{x}_1$, an alternative choice of state variables is made, as in Fig. 3.2 (c). From the block diagram of Fig. 3.2 (c), the system differential equations are written

$$\begin{aligned} \dot{x}_1 &= -(\dot{x}_2 + 5x_2) \\ \dot{x}_2 &= -2x_2 + y \end{aligned} \tag{3.9}$$

where $y = f(x_1) = x_1 g(x_1)$. These equations may be rearranged by substitution for $\dot{x}_2$ to yield

$$\begin{aligned} \dot{x}_1 &= -3x_2 - g(x_1)x_1 \\ \dot{x}_2 &= -2x_2 + g(x_1)x_1. \end{aligned} \tag{3.10}$$

Schultz illustrates, by means of examples, how the phase variables can lead to trouble when trying to determine regions of stability by the second method of Liapunov. However, satisfactory results may be obtained for certain problems through the use of phase variables. He comments that the quality of the answer obtained when trying to generate the regions of stability is thus not merely a function of the investigators ability to manipulate one or other methods of generating Liapunov functions, but also on the co-ordinate system in which he chooses to operate.

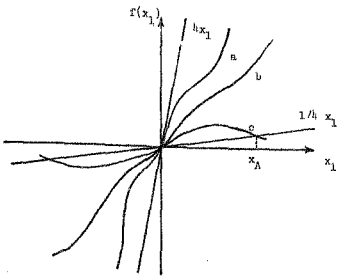


FIGURE C.1. Possible nonlinearities associated with Ex. C.1.

nonlinearity must satisfy to ensure global stability. The preferred function is the one which gives the largest range.

e.g. Constraining dV/dt to initially be

$$\frac{dV}{dt} = -2(x_1^2 + 6x_2^2)$$

yields a range of

$$0.15 < g(x_1) < 6.51.$$

For further insight into the example assume that $y = f(x_1)$ is approximated with sufficient accuracy by $y = x_1 - x_1^3$ in the region of interest. Then

$$g(x_1) = 1 - x_1^2, \qquad \text{and}$$

$$\frac{dV}{dt} = -6\left(x_1^2 - x_1^4 + \frac{4}{3} x_1^3 x_2 + x_2^2 \right).$$

$\frac{dV}{dt} = 0$ at the origin.

Along $x_2 = 0$, dV/dt changes sign when $x_1^2 = x_1^4$, i.e., $x_1 = \pm 1$. In order to divide the state space into regions where dV/dt is positive and negative, the equation $dV/dt = 0$ must be solved in terms of x_1 and x_2. As x_2 appears as a squared term, it may be determined by the quadratic formula to be

$$x_2 = -\frac{2}{3} x_1^3 \pm \left(\frac{4}{9} x_1^6 + x_1^4 - x_1^2 \right)^{\frac{1}{2}}.$$

Values of x_2 are determined by substituting values for x_1 into the above formula. In this example, values of x_1 between -0.866 and 0.866 yield complex values for x_2. For $|x_1| \geq 0.866$, the curve $dV/dt = 0$ is plotted in Fig. C.2. dV/dt is negative in the region surrounding the origin and bounded on the left and right by $dV/dt = 0$.

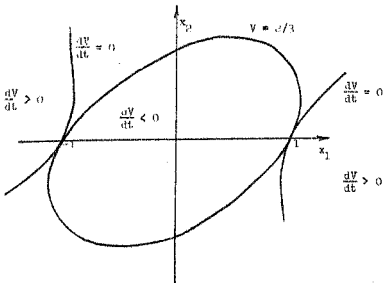


FIGURE C.2. Regions of stability for $f(x_1) = x_1(1-x_1^2)$ in Ex. C.1.

In order to determine the region of asymptotic stability, the largest V curve that fits into the region of dV/dt must be found. The curve $V = \frac{2}{3}$, is such a curve, and is illustrated in Fig. C.2.

The Method of Aizerman determined a V function that proved to be suitable for a variety of nonlinearities. In the above example, the regions of stability for the given system were determined for nonlinearities as diverse as those pictured in Fig. C.1.

The Method of Aizerman is easy to understand and to apply. However, it has been seen that the bounds on $g(x_1)$ were determined by the solution of a quadratic equation in the second-order case. For higher order cases, direct solution is not always possible because of the problem of finding the n roots of an n^{th} order polynomial. Also, if the coefficients of the given differential equations had not been given as numbers, then, for $n > 3$, no solution would be available at all.

D. THE METHOD OF SZEGO [25].

In the Aizerman approach the nonlinearity is represented by the "best" constant K. The equation, or nonlinear expression, need not have been specified, other than $y = f(x_1) = x_1 g(x_1)$, and one V function serves to define a region of stability for a variety of $f(x_1)$. The Szego approach will be to approximate the nonlinearity by a polynomial and seeks to find the one best V function for a given specific case. The hope is that since the V function is tailored directly to one unique nonlinear function, the resulting region of stability will be as large as possible.

Consider the system described by the equation

$$\dot{x}(t) = A(x(t))x(t), \qquad x(t_0) = x_0,$$

where A is some nonlinear function of its arguments. For the ensuing discussion, it is assumed that A is not a function of one of the variables x_j. Let the variable be x_n. This is often a satisfactory assumption, particularly in nonlinear gain type problems, where x_n is not involved in any nonlinear terms.

Allow the generating V function to be a general quadratic form with variable coefficients:

$$V(x) = x^T S(x)\, x$$

where $S(x) = (s_{ij}(x_j, x_i)), \quad s_{ij} = s_{ji}$,

and the variable coefficients are not allowed to be functions of x_n. This restriction is made necessary by the manner in which Szego constrains dV/dt. We have

$$\frac{dV}{dt} = x^T(A^T(x)S(x_i,x_j) + S(x_i,x_j)A(x) + \frac{dS}{dt}(x_i,x_j)) x.$$

In this particular case, the above expression can be written as

$$\frac{dV}{dt} = x^T(A^T(x)S^*(x_i,x_j) + S^{*T}(x_i,x_j)A(x)) x, \qquad (D.1)$$

where the elements of the matrix $S^*(x_i,x_j)$ have the form

$$s^*_{ij}(x_i,x_j) = s_{ij}(x_i,x_j) + \varepsilon_{ij}(s_{ij}(x_i,x_j))_{x_i} x_i$$

where $\varepsilon_{ij} = \frac{1}{2} \quad i = j$

$= 1 \quad i \neq j,$

and $s^*_{ij}(x_i,x_j) \neq s^*_{ji}(x_i,x_j).$

Since $s_{ij}(x_i,x_j)$ are polynomial functions, the form of $s^*_{ij}(x_i,x_j)$ is the same as the form of $s_{ij}(x_i,x_j)$, i.e., they are both polynomials with the same terms in x_i. Therefore, $S(x_i,x_j)$ has the same form as $S^*(x_i,x_j)$. Consequently, it is possible to investigate, instead of the exact expression (D.1), the auxiliary equation

$$\Psi(x) = x^T(A^T(x)S(x_i,x_j) + S(x_i,x_j)A(x)) x.$$

Now, if $A(x)$ has no terms involving x_n, then since the coefficients of $S(x_i,x_j)$ do not contain x_n, $\Psi(x)$ is always an algebraic equation of second degree in x_n. As a consequence, the equation $\Psi(x) = 0$ can always be solved for x_n by the *quadratic formula*. The solutions to this equation define two surfaces in the phase space, and the sign of $\Psi(x)$ changes as these surfaces are crossed. If these surfaces are forced to coincide, $\Psi(x)$ will not change sign in the whole phase space.

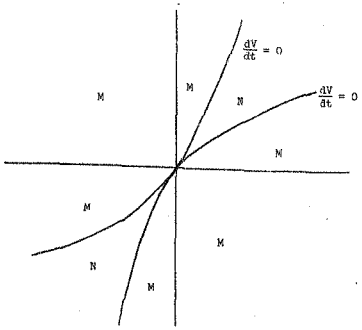


Figure D.1. Szego's method of constraining dV/dt to be negative semi-definite by forcing the solutions to the equation dV/dt = 0 to coincide.

This fact is most easily seen in two dimensions, as in Fig. D.1. Here the solution of the curve $\Psi(x) = 0$ is simply two lines in the plane as shown. It is assumed arbitrarily that, in regions M of Fig. D.1, $\Psi(x)$ is negative, and in regions N it is positive. As the curves are brought closer together, the regions N shrink, so that when the curves coincide, $\Psi(x)$ is negative on the whole space.

If $\Psi(x)$ is arranged as a quadratic in x_n in the form

$$\Psi(x) = Ax_n^2 + Bx_n + C,$$

then if A, B, C are treated as constants, the roots of the equation $\Psi(x) = 0$ can be made to coincide, if the radical in the usual quadratic formula can be made equal to zero.

i.e. $B^2 - 4AC = 0$.

Note: This can be done by setting $A = B = 0$ or $B = C = 0$, for example.

However, $\Psi(x)$ is not the function of interest. The function of interest is dV/dt, but dV/dt does have the same form as $\Psi(x)$. Hence, it is reasonable to expect that a V function of the same form that was used in connection with $\Psi(x)$ might also yield a dV/dt that could be constrained to be at least negative semi-definite, as $\Psi(x)$ is constrained to be at least negative semi-definite.

Thus the problem is started over, this time with a V function of the form determined from the consideration of the auxiliary equation $\Psi(x)$. The coefficients of this new V function are left arbitrary, and they are determined by constraints on dV/dt which make it at least negative semi-definite.

Thus dV/dt is negative or zero in the whole space, and the region of stability is determined by the largest V surface that remains closed. We now demonstrate the method by means of two examples.

Example D.1: Consider the system described by the following equations:

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -x_2 - x_1^3.$$

Choose V to be

$$V = s_{11}(x_1)x_1^2 + 2\,s_{12}(x_1)\,x_1x_2 + s_{22}x_2^2.$$

Choose $s_{22} = 1$. This is equivalent to choosing a scale factor for V. The auxiliary equation is

$$\Psi(x) = x_2^2(\,2s_{12}(x_1) - 2\,) + x_1(\,2s_{11}(x_1)x_1 - 2s_{12}(x_1)x_1 - 2x_1^3\,) - 2s_{12}(x_1)x_1^4.$$

which is arranged as a quadratic in x_2. The roots of the equation $\Psi(x) = 0$ can be made to coincide by setting

$$B^2 = 4AC,$$

where

$$A = 2s_{12}(x_1) - 2$$
$$B = 2s_{11}(x_1)x_1 - 2s_{12}(x_1)x_1 - 2x_1^3$$
$$C = -2s_{12}(x_1)x_1^4.$$

Constrain A *and* B to be zero. Then,

$$s_{12}(x_1) = s_{12} = 1, \quad s_{11}(x_1) = 1 + x_1^2.$$

Hence the V associated with $\Psi(x)$ is known and the form of V associated with dV/dt is also known. The problem is now started over, under the assumption that V is

$$V = ax_1^4 + bx_1^2 + cx_1x_2 + x_2^2.$$

Here a, b, c, and are arbitrary constants. For $a = \frac{1}{4}$, $b = 1$, $c = 2$,

$$V = \tfrac{1}{4}x_1^4 + x_1^2 + 2x_1x_2 + x_2^2,$$

$$\frac{dV}{dt} = -2x_1^4.$$

Here V is positive definite and dV/dt negative semi-definite. Theorem 2.6 applies, since dV/dt is not zero along a trajectory, as $x_1 = 0$ is not a solution of the given equations, except when x_2 also equals zero. Ω is the entire space, and thus the given equations are globally asymptotically stable.

The difficulty as well as the flexibility of the Szego method is shown up in the next example.

Example D.2: Consider the system described by the following equations:

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = x_3$$
$$\dot{x}_3 = -(x_1 + cx_2)^3 - bx_2.$$

Then defining V to be

$$V = s_{11}(x_1)x_1^2 + 2s_{12}(x_1,x_2)x_1x_2 + 2s_{13}(x_1)x_1x_3 + s_{22}(x_2)x_2^2 + 2s_{23}(x_2)x_2x_3 + s_{33}x_3^2.$$

$\frac{1}{2}\dot{V}(x)$ is found to be

$$\begin{aligned}\tfrac{1}{2}\dot{V}(x) = &- x_3^2(s_{33}b - s_{23}(x_2)) \\ &- x_3\{s_{33}(x_1 + cx_2)^3 + bs_{23}(x_2)x_2 + bs_{13}(x_1)x_1 \\ &- s_{22}(x_2)x_2 - s_{13}(x_2)x_2 - s_{12}(x_1,x_2)x_1\} \\ &- s_{23}(x_2)x_2(x_1 + cx_2)^3 + s_{13}(x_1)x_1(x_1 + cx_2)^3 \\ &- s_{12}(x_1,x_2)x_2^2 - s_{11}(x_1)x_1x_2.\end{aligned}$$

To constrain the surfaces resulting from the equation $\dot{V}(x) = 0$ to coincide, set $B = C = 0$. In C a term in x_1^4 results which cannot be cancelled unless s_{13} is zero, Since one coefficient is always arbitrary, set $s_{23}(x_2) = 1$. Then $C = 0$ results in

$$x_2(x_1 + cx_2)^3 = s_{12}(x_1,x_2)x_2^2 + s_{11}(x_1)x_1x_2,$$

or $$x_1^3 + 3cx_1^2x_2 + 3c^2x_1x_2^2 + c^3x_2^3 = s_{12}(x_1,x_2)x_2 + s_{11}(x_1)x_1.$$

If $s_{11}(x_1) = x_1^2$, then

$$s_{12}(x_1,x_2) = 3cx_1^2 + 3c^2x_1x_2 + c^3x_2^2.$$

When these known coefficients are substituted into the equation $B = 0$,

$$3cx_1^3 + 3c^2x_1^2x_2 + c^3x_2^2x_1 + s_{22}(x_1)x_2$$
$$= bx_2 + s_{33}x_1^3 + 3s_{33}cx_1^2x_2 + 3s_{33}c^2x_1x_2^2 + s_{33}c^3x_2^3.$$

If terms in like powers and like variables are equated, we have

$$3cx_1^3 = s_{33}x_1^3 \tag{D.2}$$
$$3c^2x_1^2x_2 = 3s_{33}cx_1^2x_2 \tag{D.3}$$
$$c^3x_1^2x_2^2 = 3s_{33}c^2x_1x_2^2 \tag{D.4}$$
$$s_{22}(x_2)x_2 = bx_2 + s_{33}c^3x_2^3.$$
$$\text{i.e.} \quad s_{22}(x_2) = b + s_{33}c^3x_2^2.$$

However, if equation (D.2) is solved for s_{33}, then the result $s_{33} = 3c$ does not satisfy the remaining equations. In each case it can be seen that s_{33} should be of the form $s_{33} = kc$, k constant. The fact these coefficients do not cancel means that $\Psi(x)$ is not negative semi-definite, however, the form of the coefficients is satisfactory, where

$$s_{11}(x_1) = x_1^2, \quad s_{13} = 0, \quad s_{23} = 1,$$
$$s_{12}(x_1,x_2) = 3cx_1^2 + 3c^2x_1x_2 + c^3x_2^2,$$
$$s_{22} = b + s_{33}c^3x_2^2, \quad s_{33} = kc.$$

We overlook this in the hope that the terms will actually cancel when the form of V determined from $\Psi(x)$ is applied to $\frac{1}{4}dV/dt$. The problem is now reworked with

$$V = a_1x_1^4 + a_2x_1^3x_2 + a_3x_1^2x_2^2 + a_4x_1x_2^3 + a_5x_2^4 + bx_2^2 + 2x_2x_3 + a_6x_3^2.$$

Then evaluating dV/dt, we obtain as the coefficients of:

$$x_3^1: \quad a_2x_1^3 + 2a_3x_1^2x_2 + 3a_4x_1x_2^2 + 4a_5x_2^3 - 2a_6(x_1 + cx_2)^3$$
$$x_3^0: \quad 4a_1x_1^3x_2 + 3a_2x_1^2x_2^2 + 2a_3x_1x_2^3 + a_4x_2^4 - 2x_2(x_1 + cx_2)^2.$$

It can be easily checked that these can be forced to zero by choosing

$$a_1 = \tfrac{1}{2},\ a_2 = 2c,\ a_3 = 3c^2,\ a_4 = 2c^3,\ a_5 = \tfrac{1}{2}c^4,\ a_6 = c,$$

so $$\frac{dV}{dt} = -2x_3^2(bc - 1)$$

and $$V = bx_2^2 + 2x_2x_3 + cx_3^2 + \tfrac{1}{2}(x_1 + cx_2)^4.$$

V is positive definite for $c > 0$ and dV/dt negative semi-definite in a manner such that Theorem 2.6 applies. Thus the system is globally asymptotically stable for $bc < 1$.

The resulting V proved to have a form which was successful in determining a suitable dV/dt to demonstrate global asymptotic stability. Schultz remarks that this seems to be the characteristic of the Szego approach, i.e., it works. The method of constructing $\Psi(x)$ to be negative semi-definite may be overly restrictive, but, as has been seen in the example above, that is not a vital element of the Szego technique.

Schultz comments that any objection to the basic assumption (that the nonlinearity be represented by a polynomial) on the grounds that it may be impossible to prove global asymptotic stability for a system, which is in fact globally asymptotically stable, simply because the assumption of the nonlinearity in polynomial form produces an unbounded output for large x, is not valid. In the range of physical applicability of the system in question, polynomial representation of a nonlinearity is almost always possible.

E. THE FIRST CANONIC FORM OF LUR'E.

One approach to the generation of Liapunov functions is the use of standard or canonic forms. Two canonic forms have been proposed by Lur'e [11] and a third by Letov [9]. Gibson [3] remarks that while the second and third canonic forms are perfectly valid from a mathematical point of view, they do not appear to be motivated by, nor applicable to, a wide range of physical control systems. The first canonic form, however, appears to apply to many nonlinear control systems. We shall therefore restrict ourselves to a preliminary discussion of the first canonic form of Lur'e. For a comprehensive treatment of the methods of Lur'e and Letov, the reader should refer to Letov [9]. Schultz remarks that extensions due to Rekasius [18] are given adequate coverage in the nonlinear text by Gibson [3].

Popov [16] gives three variations of canonic forms, all of which reduce to the first canonic form. Gibson remarks that the advantage of the use of such canonic forms is not merely that V functions have been developed for them, but also that simplified stability criteria are available, which may significantly simplify the analysis. A number of simplified stability criteria have been reported in the literature, however, it is not our intention to discuss these here. Gibson lists ten of the more well-known simplified criteria.

The first canonic form may be defined by the following relationships:

$$\frac{dz_i}{dt} = -\lambda_i z_i + f(\sigma) \qquad i = 1,\ldots,n \tag{E.1}$$

$$\sigma = \sum_{i=1}^{n} \alpha_i z_i \tag{E.2}$$

$$\frac{d\sigma}{dt} = \sum_{i=1}^{n} \beta_i z_i - rf(\sigma) \tag{E.3}$$

where $z_i \triangleq$ canonic variables

$e \triangleq$ variable at input to nonlinearity

$m \triangleq$ variable at output to nonlinearity (used below)

α_i shown to be negative residues at poles of $G(s)$

$-\lambda_i$ shown below to be poles of $G(s)$

$\beta_i \triangleq -\alpha_i\lambda_i$ set of constants (see below)

$r \triangleq -\sum_{i=1}^{n} \alpha_i$ a constant (see below).

To show the relationship of equations (E.1) to (E.2) to a nonlinear system such as Fig. 3.1, one may write equation (E.1) as

$$(p + \lambda_i)z_i = m \qquad i = 1,\ldots,n,$$

where $p \triangleq d/dt$ and $m = f(e)$. Solving for z_i and substituting into equation (E.2) yields

$$\frac{e}{m} = \sum_{i=1}^{n} \frac{\alpha_i}{p+\lambda_i}.$$

Since $G(s) = -e(s)/m(s)$ (see Gibson), the above equation is

$$G(s) = -\sum_{i=1}^{n} \frac{\alpha_i}{s+\lambda_i} \tag{E.4}$$

if the Laplace variable may replace the differential operator. This demonstrates that α_i and λ_i are defined as above. Now if equation (E.2) is differentiated with respect to time and equation (E.1) is substituted, we obtain equation (E.3) with

$$\beta_i = -\alpha_i\lambda_i, \quad r = -\sum_{i=1}^{n} \alpha_i \qquad i = 1,\ldots,n. \tag{E.5}$$

Rekasius [18] has shown that any differential equation in the first canonic form can be represented by a block diagram as shown in Fig. 3.1

but that the converse is true only for *systems with simple poles*.

Example E.1: Consider the system of Fig. 3.1 and find the equations of the first canonic form for

$$G(s) = \frac{s+1}{(s+2)(s+3)(s+5)}.$$

A partial-fraction expansion yields

$$G(s) = -\frac{1/3}{(s+2)} + \frac{1}{(s+3)} - \frac{2/3}{(s+5)}.$$

By means of equation (E.4), α_i and λ_i are identified, and equation (E.1) may be written as

$$\dot{z}_1 = -2z_1 + f(e)$$
$$\dot{z}_2 = -3z_2 + f(e)$$
$$\dot{z}_3 = -5z_3 + f(e).$$

From equation (E.2),

$$\dot{e} = -\frac{1}{3}z_1 + 3z_2 - \frac{10}{3}z_3.$$

A V function suitable for analysis of the first canonic form has been given by Lur'e as

$$V = \int_0^e f(e)de + \Phi(z_1,z_2,\ldots,z_n) + F(a_1z_1,\ldots,a_nz_n) \qquad (E.6)$$

where

$$\Phi(z_1,z_2,\ldots,z_n) = \tfrac{1}{2}(A_1z_1^2 + A_2z_2^2 + \ldots + A_nz_n^2) + (C_1z_{n+1}z_{n+2} + C_3z_{n+3}z_{n+4} + \ldots + C_{n-s-1}z_{n-1}z_n)$$

and $F(a_1z_1,\ldots,a_nz_n) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{a_ia_jz_iz_j}{\lambda_i+\lambda_j}$,

The constants λ_i can be either real or complex. The real λ_i's in the above equations are designated as $\lambda_1, \lambda_2, \ldots, \lambda_s$. The corresponding canonic variables may also be shown to be real. The remaining $(n - s)$ λ_i are complex conjugate pairs and are designated λ_{s+1}, $\lambda_{s+2}, \ldots, \lambda_{n-1}, \lambda_n$. The corresponding canonic variables $z_{s+1}, \ldots, z_n$ likewise appear as complex conjugate pairs. Consequently the quadratic forms Φ and F can only take on real values, and Φ can only be nonnegative for positive values of the constants A_i and C_i. The quadratic form F will take on only nonnegative values if

(i) $\mathrm{Re}(\lambda_i) > 0$ for all $i = 1, \ldots, n$, i.e., all the poles of $G(s)$ are in the LHP, and

(ii) the constants a_i are real for real λ_i's and complex conjugate pairs for complex conjugate pairs of λ_i.

Under these restrictions the V function of equation (E.6) will take on only nonnegative values if the nonlinear element satisfies

$$\int_0^e f(e)de \geq 0 \quad \text{for all} \quad |e| > 0 \qquad f(0) = 0.$$

Differentiating equation (E.6) and substituting equations (E.1) and (E.2) gives

$$\begin{aligned}\frac{dV}{dt} = {} & -f(e)^2(e) - \Big\{ \sum_{i=1}^{n} a_i z_i \Big\}^2 - \Big\{ A_1\lambda_1 z_1^2 + \cdots + A_n\lambda_n z_n^2 \\ & + C_1(\lambda_{s+1} + \lambda_{s+1})z_{s+1}z_{s+2} + \cdots + C_{n-1}(\lambda_{n-1} + \lambda_n)z_{n-1}z_n \Big\} \\ & + f(e)\Big\{ \sum_{i=1}^{n} z_i \left(A_i + B_i - 2a_i \sum_{j=1}^{n} \frac{a_j}{\lambda_i + \lambda_j} \right) \\ & + \sum_{i=1}^{n-1} z_{s+i} \left(C_{2i-1} + B_{s+i} - 2a_{s+i} \sum_{j=1}^{n} \frac{a_j}{\lambda_j + \lambda_{s+i}} \right) \Big\}. \end{aligned} \tag{E.7}$$

The constants A_i, C_i and a_j can always be selected so as to make

the sum of the terms of equation (E.7) not containing f(e) either positive or negative definite, depending on the sign of r. It is frequently possible to select A_i, C_i and a_j in such a way as to make the term containing f(e) equal to zero, at the same time retaining the definiteness of the sign of dV/dt. This can be accomplished by setting

$$\sum_{j=1}^{n} \frac{2a_i a_j}{\lambda_i + \lambda_j} = B_i + A_i \qquad i = 1,\ldots,s$$

and

$$\sum_{j=2}^{n} \frac{2a_{s+i}a_j}{\lambda_{s+i}+\lambda_j} = B_{s+i} + C_{2i-1} \qquad i = 1,\ldots,n-s \tag{E.8}$$

The solution of the above equations yields sufficient conditions for the stability of a system described by the first canonic form of differential equations. It is common practice and seems quite appropriate to refer to equations, such as equations (E.8), that yield sufficient conditions, as stability equations.

Example E.2: Consider the previous example E.1. Apply the V function of equation (E.6) to establish the limits of stability. dV/dt can be found by substituting into equation (E.7).

$$\frac{dV}{dt} = -(A_1 z_1^2 + A_2 z_2^2 + A_3 z_3^2)^2 - 2A_1 z_1^2 - 3A_2 z_2^2 - 5A_3 z_3^2$$
$$+ f(e)\Big[(A_1 + \tfrac{2}{3} - \tfrac{1}{2} a_1^2 + \tfrac{2}{5} a_1 a_2 - \tfrac{2}{7} a_1 a_3) z_1$$
$$+ (A_2 - 3 + \tfrac{2}{5} a_1 a_2 - \tfrac{1}{3} a_2^2 - \tfrac{1}{4} a_2 a_3) z_2$$
$$+ (A_3 + \tfrac{10}{3} - \tfrac{2}{7} a_1 a_3 - \tfrac{1}{4} a_2 a_3 - \tfrac{1}{5} a_3^2) z_3 \Big].$$

It can be seen that dV/dt may be made negative definite by setting the terms of f(e) equal to zero and by selecting the values of the

constants A_1, A_2 and A_3 as sufficiently small positive numbers. In order to prove stability, however, it is sufficient to let

$$A_1 = A_2 = A_3 = 0.$$

From equation (E.8) one obtains three expressions involving a_1, a_2 and a_3, which, when solved simultaneously, give

$$a_1 = \frac{10}{3}, \quad a_2 = -12, \quad a_3 = 11\frac{2}{3}.$$

Note: In general, such a solution is not necessarily simple. For n constants, n simultaneous nonlinear equations must be solved. For n = 2 the problem is not usually difficult.

Thus equation (E.6) becomes

$$V = \int_0^e f(e)de + \frac{25}{9} z_1^2 - 16\, z_1 z_2 + \frac{100}{9} z_1 z_3 + 24\, z_2^2 - 35\, z_2 z_3 + \frac{245}{18} z_3^2.$$

From equation (E.7)

$$\frac{dV}{dt} = -\left(\frac{10}{3} z_1 - 12\, z_2 + 11\frac{2}{3} z_3\right)^2.$$

Since V is positive definite and dV/dt is negative semi-definite, the system is globally asymptotically stable provided that the nonlinearity satisfies $\int_0^e f(e)de \to \infty \quad \forall\, |e| > 0$, $f(0) = 0$.

If the constant r in the canonic equation is positive, (E.7) can be negative definite under much weaker restrictions than the requirement that the term containing f(e) be zero. To show this, write (E.7) as

$$\frac{dV}{dt} = -\Big(A_1\lambda_1 z_1^2 + \ldots + A_s\lambda_s z_s^2 + C_1(\lambda_{s+1} + \lambda_{s+2})z_{s+1}z_{s+2} + \ldots + C_{n-s-1}(\lambda_{n-1} + \lambda_n)z_{n-1}z_n \Big) - \Big(\sum_{i=1}^{n} a_i z_i + \sqrt{r}\, f(e) \Big)^2 \tag{E.9}$$

$$+ f(e)\Big\{ \sum_{i=1}^{s} z_i \Big(A_i + \beta_i + 2a_i\Big(\sqrt{r} - \sum_{j=1}^{n} \frac{a_j}{\lambda_j + \lambda_i} \Big) \Big) + \sum_{i=1}^{n-s} z_{s+i} \Big(C_{2i-1} + \beta_{s+i} + 2a_{s+i}\Big(\sqrt{r} - \sum_{j=1}^{n} \frac{a_j}{\lambda_j + \lambda_{s+i}} \Big) \Big) \Big\}.$$

In order that dV/dt be negative definite, the constants A_i and C_i must be positive, and

$$Re(\lambda_i) > 0, \quad i = 1,\ldots,n.$$

The constant r must be nonnegative, and

$$A_i + \beta_i + 2a_i\Big(\sqrt{r} - \sum_{j=1}^{n} \frac{a_j}{\lambda_j + \lambda_i} \Big) = 0, \qquad i = 1,\ldots,s$$
$$C_{2i-1} + \beta_{s+i} + 2a_{s+i}\Big(\sqrt{r} - \sum_{j=1}^{n} \frac{a_j}{\lambda_j + \lambda_{s+i}} \Big) = 0 \quad i = 1,\ldots,n-s \tag{E.10}$$

Under the above restrictions, the V function of (E.6) will be positive definite even if the quadratic form $\Phi(z_1,\ldots,z_n)$ is omitted. The derivitive dV/dt will then be negative semi-definite. Equation (E.10) may be replaced by

$$2a_i\sqrt{r} - 2a_i \sum_{j=1}^{n} \frac{a_j}{\lambda_j + \lambda_i} + \beta_i = 0 \quad i = 1,\ldots,n \tag{E.11}$$

The solution of equation (E.11) and the constraint $\int_0^e f(\sigma)d\sigma \geq 0$ for all $|e| > 0$, $f(0) = 0$, yield sufficient conditions for asymptotic stability for many conditions. These are summarized by Lur'e in:

Theorem E.3: If a system described by the first canonic form satisfies the following conditions:

(i) λ_i and β_i are real for $i \leq s$,

λ_i and β_i are each complex conjugate pairs for $s < i \leq n$,

(ii) $\mathrm{Re}(\lambda_i) > 0$ for $i = 1,\ldots,n$,

(iii) $r > 0$,

(iv) $\int_0^\infty f(a)da \geq 0$ for all $|a| > 0$ and $f(0) = 0$,

then this system is globally asymptotically stable if there exists at least one solution of the set of stability equations (E.11) with the roots $a_1,\ldots, a_s$ being real, and the roots $a_{s+1},\ldots, a_n$ being complex conjugate pairs.

Rozenvasser [21] has shown that requirement (iii) in the above theorem can be replaced by $r \geq 0$, a significant generalization.

We have already seen that the solution procedure becomes increasingly complex as n increases. Schultz [22] comments that with a large number of terms in V, the central problem of constraining dV/dt to be at least negative semi-definite becomes prohibitive. We will now describe two methods which overcome this problem by assuming a simple quadratic V (as did Szego), where the coefficients of V can contain higher order terms as necessary.

F. THE VARIABLE GRADIENT METHOD.

We now discuss the variable gradient method of Schultz and Gibson [23]. For a detailed discussion of the variable gradient method, also see Schultz [22], Gibson [3] and Willems [26]. As the name implies, the variable gradient method is based upon the assumption of a vector VV with n undetermined components. The following lemma is a useful guide in the selection of the gradient.

Lemma F.1: A necessary and sufficient condition such that a continuous vector function $g(x)$ be the gradient of a scalar function is that the matrix

$$M = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_2}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_1} \\ \frac{\partial g_1}{\partial x_2} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_n}{\partial x_2} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\partial g_1}{\partial x_n} & \frac{\partial g_2}{\partial x_n} & \cdots & \frac{\partial g_n}{\partial x_n} \end{pmatrix}$$

be symmetric, where $g_1, g_2, \ldots, g_n$ denote the components of the vector function $g(x)$.

Hence, if $g(x) = VV(x)$, we have, from the lemma, that the following $(n-1)n/2$ equations must be satisfied:

$$\frac{\partial VV_i}{\partial x_j} = \frac{\partial VV_j}{\partial x_i} \qquad i,j = 1,\ldots n. \tag{F.1}$$

Since $g(x)\cdot dx = VV(x)\cdot dx = dV(x)$, the integral

$$\int_{x_A}^{x_B} g(x)\cdot dx$$

is independent of the particular path joining the points A and B of the state space. The function V may be determined from ∇V as a line integral

$$V = \int_0^x \nabla V \cdot dx \qquad (F.2)$$

along any path joining the origin of the state space to the point x.

In order to emphasize the role of the gradient function, Theorem 2.6 is restated as:

Theorem F.2: If for the system of equations $\dot{x} = f(x)$, with $f(0) = 0$, there exists a region Ω and a real vector function ∇V, with elements ∇V_i such that

(i) $(\partial \nabla V_i/\partial x_j) = (\partial \nabla V_j/\partial x_i)$,

(ii) ∇V is not zero at any point in Ω, except at the origin,

(iii) $dV/dt = \nabla V \dot{x} \leq 0$ for $x \neq 0$, and $dV/dt = 0$ for $x = 0$,

(iv) dV/dt is not identically zero in Ω on a solution of the system, other than at $x = 0$,

and if the scalar function V, formed by a line integral of ∇V, as

$$V = \int_0^x \nabla V \cdot dx$$

has the following properties:

(i) it is positive definite

(ii) one of the surfaces $V = K_1$, bounds Ω,

then the given system $\dot{x} = f(x)$ is asymptotically stable in Ω.

This is simply a restatement of Theorem 2.6 to indicate a shift in emphasis.

We now outline the formal application of the variable gradient method.

1. Assume a gradient of the form

$$\nabla V = \begin{bmatrix} \nabla V_1 \\ \nabla V_2 \\ \vdots \\ \nabla V_n \end{bmatrix} = \begin{bmatrix} \alpha_{11}(x)x_1 + \alpha_{12}(x)x_2 + \ldots + \alpha_{1n}(x)x_n \\ \alpha_{21}(x)x_1 + \alpha_{22}(x)x_2 + \ldots + \alpha_{2n}(x)x_n \\ \vdots \\ \alpha_{n1}(x)x_1 + \alpha_{n2}(x)x_2 + \ldots + \alpha_{nn}(x)x_n \end{bmatrix}$$

(The α_{ij}'s may be written as constants until the need arises to allow them to be more complicated.)

2. From the variable gradient, form dV/dt as $dV/dt = \nabla V \dot{x}$.

3. In conjunction with and subject to the requirements of the generalized curl equations (F.1), constrain dV/dt to be at least negative semi-definite.

In general, an attempt is made to make dV/dt negative semi-definite in as simple a way as possible. This may be accomplished if indefinite terms in dV/dt are eliminated, and then in the simplest case,

$$\frac{dV}{dt} = -K x_i^2 \qquad K > 0$$

where K is initially assumed a constant. If dV/dt is constrained as above, the remaining terms in dV/dt must be forced to cancel. This is accomplished by grouping terms of similar state variables and choosing the α_{ij}'s to force cancellation.

4. From the known gradient, determine V and the region of closedness of V.

This procedure will now be illustrated with examples.

Example F.3: Consider the second order system

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -x_2 - x_1^3 ,$$

1. Take
$$\nabla V = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix}.$$

2. dV/dt is determined as

$$\frac{dV}{dt} = x_1x_2(a_{11} - a_{21} - a_{22}x_1^2) - (a_{22} - a_{12})x_2^2 - a_{21}x_1^4.$$

3. To satisfy the curl equations

$$x_1\frac{\partial a_{11}}{\partial x_2} + a_{12} + x_2\frac{\partial a_{12}}{\partial x_2} = a_{21} + x_1\frac{\partial a_{21}}{\partial x_1} + x_2\frac{\partial a_{22}}{\partial x_1} .$$

To constrain dV/dt to be negative definite, make a_{12}, a_{21}, a_{22} constant, and set $a_{22} > a_{12}$, $a_{21} > 0$, $a_{11} = a_{21} + a_{22}x_1^2$.

The curl equation gives

$$a_{12} = a_{21},$$

and finally we have

$$a_{22} > a_{12} = a_{21} > 0.$$

4. Compute $V(x)$ and check for definiteness.

$$V(x) = \frac{1}{4}a_{22}x_1^4 + \frac{1}{2}a_{21}x_1^2 + a_{21}x_1x_2 + \frac{1}{2}a_{22}x_2^2 ,$$

which is positive definite for $a_{22} > a_{21} > 0$.

This establishes global asymptotic stability of the null solution of the above system of equations.

Example F.4: Consider the system described by equations (3.10).

$$\dot{x}_1 = -3x_2 - g(x_1)x_1$$
$$\dot{x}_2 = -2x_2 + g(x_1)x_1.$$

1. $$\nabla V = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix}$$

2. $$\frac{dV}{dt} = -x_1^2(a_{11}g(x_1) - a_{21}g(x_1)) + x_1x_2(-3a_{11} - a_{12}g(x_1) - 2a_{21} + a_{22}g(x_1)) - x_2^2(2a_{22} + 3a_{12}).$$

3. The simplest manner in which dV/dt may be constrained to be negative is if $a_{12} = a_{21} = 0$. Note that this ensures that the curl equations are automatically satisfied. Then we have

$$\frac{dV}{dt} = -a_{11}g(x_1)x_1^2 - 2a_{22}x_2^2 + x_1x_2(-3a_{11} + a_{22}g(x_1)).$$

The indefinite terms in x_1x_2 may be completely eliminated if

$$a_{11} = \tfrac{1}{3}a_{22}\,g(x_1).$$

Thus the gradient is known to be

$$\nabla V = \begin{pmatrix} \frac{1}{3}a_{22}\,g(x_1)x_1 \\ a_{22}x_2 \end{pmatrix}$$

and $$\frac{dV}{dt} = -\tfrac{1}{3}a_{22}g(x_1)^2x_1^2 - 2a_{22}x_2^2.$$

4. $$V = \int_0^{x_1} \tfrac{1}{3}a_{22}g(x_1')x_1'dx_1' + \int_0^{x_2} a_{22}x_2'dx_2'.$$

With $a_{22} = 6$ as a choice of scale factor to eliminate fractions

$$V = 2\int_0^{x_1} g(x_1')x_1'dx_1' + 3x_2^2.$$

As long as $x_1 g(x_1) = f(x_1) = y$ lies in the first and third quadrants, V is positive definite. If the integral goes to ∞ as $x_1 \to \infty$, then V is not only positive definite, but represents a closed surface in the whole space. dV/dt is negative definite. On the basis of Theorem 2.6 or F.2 the system is globally asymptotically stable. This is understandably a better answer than was realized using Aizerman's procedure, since both V and dV/dt reflect the effect of the specific nonlinear function $x_1 g(x_1)$.

We have given a resumé here of the powerful variable gradient method. For our purposes we have considered fairly simple, straightforward examples. The reader should refer to the excellent discussion given in the article by Schultz [22], where fairly complex examples are considered, and the flexibility of this approach is demonstrated. The variable gradient method reduces the second method of Liapunov to a practical working tool for the analysis of many nonlinear systems.

In the next section we consider a method which does not appear to be entirely unrelated to the variable gradient method. However, not much information or coverage is given to this method in the literature.

G. THE FORMAT METHOD, [17].

Definition G.1: A diagonal matrix has elements on the principle diagonal only.

Definition G.2: A skew-symmetric matrix P has no elements on the principle diagonal, and in addition

$$P_{ji} = - P_{ij}.$$

For a skew-symmetric matrix,

$$P + P^T = 0$$

and $$Pf \cdot f = 0.$$

The format method was developed by J.L. Peczkowski [14],[15] and is based upon the following theorem.

Theorem G.3: If D(x) is a diagonal $n \times n$ matrix such that

$$Df \cdot f = \dot{V}(x)$$

and if P(x) is a real skew-symmetric matrix such that

$$VV = (D + P)f$$

then, if the curl equations (F.1) are satisfied we have

$$\dot{V}(x) = VV \cdot f.$$

When D(x) has only one element, $d_{ii} = L(x)$, then

$$\dot{V}(x) = Df \cdot f = d_{ii} f_i^2 = L(x) f_i^2.$$

Because f_i^2 is positive, $\dot{V}(x)$ will be semi-definite or definite if

$L(x)$ is sign definite.

To generate Liapunov functions by the Format method:

1. Write the vector format

$$\dot{V} = (D + P)f.$$

2. Choose the arbitrary functions p_{ij} and d_{ii} to satisfy the curl equations (F.1) and then obtain P and D. Because $D(x)$ may have different forms, it is customary to assume first that $D(x)$ contains only the d_{ii} element (i.e. $\dot{V}(x) = d_{11}f_1^2$). If this does not yield a satisfactory Liapunov function, then assume that $D(x)$ contains only the d_{22} element. After exhausting all the single element possibilities, different forms are then tried.

3. Construct the functions $V(x)$ and $\dot{V}(x) = Df.f$.

We now demonstrate the method by means of an example.

Example 6.6: Consider example F.1 described in the previous section.

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -x_1^3 - x_2.$$

1. Write the vector format.

$$\dot{V} = (D + P)f$$

$$= \begin{bmatrix} L & -p \\ p & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ -x_1^3 - x_2 \end{bmatrix}$$

$$= \begin{pmatrix} Lx_2 + px_2 + px_1^3 \\ px_2 \end{pmatrix}.$$

2. The curl equations give

$$\frac{\partial}{\partial x_2}(Lx_2 + px_2 + px_1^3) = \frac{\partial}{\partial x_1}(px_2).$$

i.e. $$\frac{\partial}{\partial x_2}(Lx_2 + px_2) + x_1^3 \frac{\partial p}{\partial x_2} = x_2 \frac{\partial p}{\partial x_1}.$$

In the first trial assume that p is constant. Therefore,

$$\frac{\partial}{\partial x_2}(Lx_2 + px_2) = 0$$

The curve equation is satisfied if $L = -p$ is constant. Thus

$$\nabla V = \begin{pmatrix} px_1^3 \\ px_2 \end{pmatrix}.$$

Integrating to obtain the Liapunov functions gives

$$V(x) = \int_0^x \nabla V \, dx$$

$$= \int_0^{x_1} px_1^3 \, dx_1 + \int_0^{x_2} px_2 \, dx_2$$

$$= \frac{px_1^4}{4} + \frac{px_2^2}{2}.$$

The derivitive along the trajectories is

$$\dot{V}(x) = -px_2^2.$$

For $p > 0$, $V(x)$ is positive definite and $\dot{V}(x)$ is negative def, for $p < 0$, $V(x)$ is negative definite and $\dot{V}(x)$ is positive def. Thus the system is globally asymptotically stable for all values of p. It is to be noted that although the Liapunov function of section F and that given above are different, they both yield the same result.

$$= \begin{pmatrix} Lx_2 + px_2 + px_1^3 \\ px_2 \end{pmatrix}.$$

2. The curl equations give

$$\frac{\partial}{\partial x_2}(Lx_2 + px_2 + px_1^3) = \frac{\partial}{\partial x_1}(px_2),$$

i.e.

$$\frac{\partial}{\partial x_2}(Lx_2 + px_2) + x_1^3 \frac{\partial p}{\partial x_2} = x_2 \frac{\partial p}{\partial x_1}.$$

In the first trial assume that p is constant. Therefore,

$$\frac{\partial}{\partial x_2}(Lx_2 + px_2) = 0$$

The curve equation is satisfied if $L = -p$ is constant. Thus

$$VV = \begin{pmatrix} px_1^3 \\ px_2 \end{pmatrix},$$

Integrating to obtain the Liapunov functions gives

$$V(x) = \int_0^x VV \, dx$$

$$= \int_0^{x_1} px_1^3 \, dx_1 + \int_0^{x_2} px_2 \, dx_2$$

$$= \frac{px_1^4}{4} + \frac{px_2^2}{2}.$$

The derivitive along the trajectories is

$$\dot{V}(x) = -px_2^2.$$

For $p > 0$, $V(x)$ is positive definite and $\dot{V}(x)$ is negative def, for $p < 0$, $V(x)$ is negative definite and $\dot{V}(x)$ is positive def. Thus the system is globally asymptotically stable for all values of p. It is to be noted that although the Liapunov function of section F and that given above are different, they both yield the same result.

CONCLUSION.

We have shown that there are an infinite number of V functions, and their corresponding time derivatives, which are capable of proving the stability for linear systems. For linearized systems whose eigenvalues have negative real parts, a general approach was indicated that always gave an estimate, although often a poor one, of the region of stability. This general approach was inflexible, in that the forms of V and dV/dt were dictated by the manner in which the state variables were chosen.

The Aizerman approach returned to the infinite choice of dV/dt, and therefore V, of the linear system case by approximating the nonlinearity by a "best linear" gain K. Here the equation or nonlinear expression need not been specified, other than $y = f(x_1) = x_1 g(x_1)$, and one V function served to define a region of stability for a variety of $f(x_1)$. Contrasted with this was the Szego approach, in which the nonlinearity was approximated by a polynomial and V was tailored to fit the specific case. This better fit was accomplished by allowing V to contain terms of higher order than the second-order terms dictated by the usual quadratic-form choice for V. As noted, the Szego method retained the flexibility of the Aizerman approach by allowing dV/dt to be constrained to be at least negative semi-definite in a variety of ways.

We then considered techniques that contained the advantages of the Aizerman and Szego methods to allow for consideration of nonlinearities which are expressible as polynomials or general functions of x_i, such as $f(x_i)$. This extended capability was made possible by including integrals in V. Lur'e and Letov assumed that V has a

quadratic form plus an integral. However, Schultz comments that in order to make V general enough to cover a large variety of situations, too many terms have to included in V. With a large number of terms in V, the central problem of constraining dV/dt to be at least negative semi-definite becomes prohibitive.

Szego overcame this problem by starting with a simple quadratic V, but allowing each of the coefficients to contain a number of higher order terms, as necessary. A similar procedure was adopted in the Variable Gradient and Format methods. Both methods allowed the coefficients of VV to be nonlinear, when necessary, which were required to satisfy the curl equations.

Although we have only discussed methods for generating Liapunov functions for autonomous systems, and most research has been conducted in this area, Schultz [22] proposed three methods for the solution of non-autonomous systems via the second method of Liapunov.

As we have shown, there is no set way of determining the required Liapunov function upon which the second method is based. However, we have demonstrated that there are a series of systematic methods that may be used to obtain a region of stability through the use of the second method. In the final analysis, though, in doing a stability analysis for a particular system, we need to exploit the properties of the particular system under consideration.

REFERENCES.

1. Aizerman, M.A., The Theory of Automatic Control of Motors, (Russian), GITTL, Moscow, 1952.

2. Donalson, D.D., The Theory and Stability Analysis of a Model Referenced Parameter Tracking Technique for Adaptive Automatic Control Systems, Doctoral Dissertation, Univ. Calif., Los Angeles, California, 1961.

3. Gibson, J.E., Nonlinear Automatic Control, McGraw-Hill, New York, 1963.

4. Hahn, W., Theory and Application of Liapunovs Direct Method, Prentice-Hall, Englewood Cliffs, New Jersey.

5. Hurwitz, A., Ueber die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Theilen besitz, Mathematische Annalen, 46, 273 - 284, 1895. Reprinted in 'Mathematical Trends in Control Theory', edited by Bellman, R.E., and Kaleba, R., Dover Publications, New York, 1964.

6. Ingwerson, D.R., A Modified Liapunov Method for Nonlinear Stability Problems, Doctoral Dissertation, Stanford Univ., Stanford, California, 1960.

7. La Salle, J.P., Some Extensions of Liapunov's Second Method, Res. Inst. Advan. Study Tech. Report, No 60-5, Baltimore, 1960.

8. " , Asymptotic Stability Criteria, Proc. Symp. Appl. Math., Vol 13: Hydrodynamic Instability, Amer. Math. Soc., 299 - 307, 1962.

9. Letov, A.M., Stability in Nonlinear Control Systems, Princeton Univ. Press, Princeton, New Jersey, 1961.

10. Liapunov, M.A., Problème general de la stabilité du mouvement, reprinted in Annals of Mathematical Studies, 17, Princeton Univ. Press, Princeton, New Jersey, 1949.

11. Lur'e, A.I., Some Nonlinear Problems in the Theory of Automatic Control, Gostekhizdat, Moscow, 1951. English translation: Her Majesty's Stationary Office, London, 1957.

12. Malkin, I.G., Theory of Stability of Motion, (English transl.), U.S. At. Energy Comm. Transl. No. 3352, 1959.

13. Margolis, S.G., and Vogt, W.G., Control engineering applications of V.I. Zubov's construction procedure for Liapunov functions, IEEE Trans. Automatic Control, AC-9, 104 - 113, 1963.

14. Paczkowski, J.L., A Format Method for Generating Liapunov Functions, Ph.D. Thesis, Univ. of Notre Dame, 1966.

15. " , A format method for generating Liapunov functions, Trans. ASME 89 (2), Series D, June 1967.

16. Popov, E.P., Dynamic Automatic Control Systems, Akadem-Verlag, GMBA, Berlin, 1958.

17. Raven, F.H., Automatic Control Engineering, McGraw-Hill, 1968.

18. Rekasius, Z.V., Stability Analysis of Nonlinear Control Systems by the Second Method of Liapunov, Ph.D. Thesis, Purdue Univ., Lafayette, Ind., 1961.

19. Rosenbrock, H.H., and Storey, C., Mathematics of Dynamical Systems, Nelson, London, 1970.

20. Routh, E.J., A Treatise on the Dynamics of a System of Rigid Bodies, Macmillan, London, 1877.

21. Rozenvasser, E.N., Stability of Nonlinear Control Systems, Automation and Remote Control (English translation), 20, 1959.

22. Schultz, D.G., The generation of Liapunov functions, Advances in Control Systems (ed. Leondes, C.T.), 2, 1 - 64, 1965.

23. Schultz, D.G., and Gibson, J.E., The variable gradient method for generating Liapunov functions, Trans. AIEE Pt II 81, 203 - 210, 1962.

24. Szego, G.P., On a new partial differential equation for the stability analysis of time-invariant control systems, SIAM J. Control, 1, 63 - 75, 1962.

25. " , A contribution to Liapunov's second method: nonlinear autonomous systems, in 'International Symposium in Nonlinear Differential Equations and Nonlinear Mechanics', (eds. La Salle, J.P., and Lefschetz, S.), 421 - 430, Academic Press, New York, 1963.

26. Willems, J.L., Stability Theory of Dynamical Systems, Nelson, London, 1970.

27. Zubov, V.I., The Methods of Liapunov and their Applications, Leningrad Univ. Press, Leningrad, 1957.

DIFFERENTIAL DYNAMIC PROGRAMMING :

A UNIFIED APPROACH TO THE OPTIMIZATION OF

DYNAMIC SYSTEMS.

Kendal Clive Jordi

DIFFERENTIAL DYNAMIC PROGRAMMING : A UNIFIED APPROACH TO THE OPTIMIZATION OF DYNAMIC SYSTEMS.

Kendal Clive Jordi

A Dissertation submitted to the Faculty of Science in partial fulfilment of the requirements for the degree of Master of Science.

University of the Witwatersrand,
Johannesburg.
1975.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Professor D.H. Jacobson, my supervisor, for his kind assistance and most valuable guidance with my work.

Special thanks to Mrs. F. Roxton-Wiggall for typing this dissertation.

This work was done with the assistance of a grant from the C.S.I.R., Pretoria, from January 1974 through to June 1975.

ABSTRACT

The purpose of this dissertation is to present certain expressions satisfied by the predicted change in cost in an optimal control problem, and to indicate the unifying role that these expressions can play in control theory and computation. To achieve this end a paper by Mayne has been used as a basis, and the results and theme of this paper have been expanded upon.

A partial differential equation is obtained which is satisfied by the cost function for an arbitrary control or control policy. The Dynamic Programming Technique is then applied to the cost function in the neighbourhood of some nominal or reference trajectory. Provided that the deviations from the nominal trajectory are small, this leads to the differential equations derived by Mayne. These differential equations are useful for obtaining optimization algorithms including the powerful Differential Dynamic Programming (D.D.P.) algorithms.

From the differential equations certain exact expressions for the change ΔV in cost due to a change in control are obtained. These expressions, which enable two arbitrary controls to be compared are useful for obtaining conditions of optimality, particularly sufficient conditions. Further estimates for the expressions for ΔV are derived, which lead to expressions $\Delta\hat{V}$ which approximate ΔV. One of the approximations $\Delta\hat{V}$ leads to a first order algorithm with

proven convergence.

The analysis is then extended to obtain new differential equations for problems with fixed endpoint constraints and free terminal time; again the two controls compared are arbitrary. The second-order algorithm of Jacobson is then discussed. It is then shown how the problem with terminal inequality constraints can be modified so that the second-order D.D.P. algorithm of Jacobson is now applied to an unconstrained problem. The first-order algorithm is then extended to handle terminal inequality constraints.

Problems with control constraints are then considered. For the case where the optimal control is continuous for all $t \in T$, using D.D.P. on the Hamilton-Jacobi-Bellman partial differential equation, certain differential equations derived by Jacobson are obtained. The second-order D.D.P. algorithm can be modified easily to solve these equations, and a first-order algorithm emerges as a special case. For the problem where the optimal control is discontinuous (i.e. of the Bang-Bang type), the dynamic programming approach is used to derive jump conditions in the partial derivatives of the cost function. Again, the second-order D.D.P. algorithm can be modified to handle the bang-bang control problem.

Finally, the approximate expressions $\hat{\Delta V}$ for ΔV are used to demonstrate necessary conditions of optimality for State Constrained problems and necessary and sufficient conditions for the non-negativity of $\hat{\Delta V}$ for singular control problems.

CONTENTS

		Page
	INTRODUCTION	1
1.	DEFINITIONS, ASSUMPTIONS AND PRELIMINARY RESULTS	
	I. System Description	1
	II. Basic Assumptions	3
	III. Properties of Trajectories	5
2.	PROPERTIES OF THE COST FUNCTION	
	I. Differentiabilty of the Cost Function	17
	II. Partial Differential Equation for the Cost Function	20
	III. The Adjoint Differential Equation	22
	IV. Exact Differential Equations for $V_x(\tilde{x}(t),t)$, $V_{xx}(\tilde{x}(t),t)$	24
	V. Approximate Differential Equations for $V_x(\tilde{x}(t),t)$, $V_{xx}(\tilde{x}(t),t)$	35
	VI. Exact Expressions for ΔV	40
3.	SUFFICIENT CONDITIONS OF OPTIMALITY	
	I. Global Sufficiency Results	42
	II. A Local Sufficiency Theorem for Strong Variations in Control	51
4.	FURTHER ESTIMATES OF ΔV	57
5.	OPTIMIZATION ALGORITHMS	
	I. First-Order D.D.P. Algorithms	68
	II. Second-Order D.D.P. Algorithms	77

6. TERMINAL CONSTRAINTS

I. Terminal Equality Constraints with Free Terminal Time 95

II. A Second-Order Algorithm for Terminal Equality Constraints with Free Terminal Time 108

III. Terminal Inequality Constraints 116

IV. First-Order Algorithms for Terminal Inequality Constraints 120

7. CONTROL CONSTRAINTS 125

8. BANG-BANG CONTROL PROBLEMS

I. Properties of the Cost Function at Switching Points 135

II. The Optimal Cost for the State $(\tilde{x}(t),t)$ 142

III. Algorithms for Solving Bang-Bang Control Problems 146

9. STATE CONSTRAINED PROBLEMS 157

10. SINGULAR CONTROL PROBLEMS 179

CONCLUSION

REFERENCES

INTRODUCTION.

Much interest has centered on the problem of determining the optimal control for the dynamic system described by non-linear, ordinary differential equations of the form

$$\dot{x}(t) = f(x(t),u(t),t); \qquad x(t_o) = x_o.$$

The criterion of optimality is the minimization of the performance index or so-called cost functional

$$V(x_o,t_o) = \int_{t_o}^{t_f} L(x(t),u(t),t)\,dt + F(x(t_f),t_f),$$

where the properties of f, L, F are presented in Chapter 1.

Sometimes it is required that $x(t)$ and $u(t)$ satisfy some or all of the following constraints:

$$g(u(t),t) \leq 0,$$
$$\psi(x(t_f),t_f) = 0,$$
$$S(x(t),t) \leq 0,$$

where, g is a $p < m$ vector function of u at time t,
ψ is a $s < n$ vector function of x at time t_f,
S is a $q < n$ vector function of x at time t.
The object of the control problem is to choose $u(t)$, $t \in [t_o,t_f]$ such that the necessary constraints are satisfied, and the cost is minimized.

The optimal cost is given by

$$V^{o}(x,t) = \min_{\substack{u(\tau) \\ \tau\in[t,t_f]}} -[\int_{t_-}^{t_f} L(x(\tau),u(\tau),\tau)\,d\tau + F(x(t_f),t_f)\ ].$$

Applying the dynamic programming technique, [1], to the above equation results in the well known Hamilton-Jacobi-Bellman (H-J-B) partial differential equation:

$$-\frac{\partial V^{o}}{\partial t}(x^{o},t) = \min_{u(t)} [\ L(x^{o},u,t) + \langle V_x^{o}(x^{o},t), f(x^{o},u,t)\rangle\ ]$$

which can only be derived if the optimal cost V^{o} is assumed to have continuous partial derivitives w.r.t. x and t. The above partial differential equation is, in general, unsolvable analytically. The difficulty of numerical solution of the equation is enormous, primarily due to the high dimensionality of the equation, which means that storage and computational time requirements are immense.

In 1960 the L.Q.P. problem (linear dynamics with quadratic performance index) was solved, [27]. This stimulated research into second-variation type methods. Merriam [43], Mitter [44], Mc Reynolds and Bryson [40] and Noton, Dyer and Markland [45] have derived algorithms that converge to the optimal solution in one step when applied to the L.Q.P. problem. On non-linear problems, convergence, if it occurs, is rapid.

Jacobson [13] outlines the dynamic programming technique, which is applied to $V(\bar{x}+\delta x,t)$, which describes the

non-optimal cost in a neighbourhood of some nominal trajectory $\bar{x}(\cdot)$, defined by a nominal control $\bar{u}(\cdot)$. Provided the change in the state variation, δx, is kept small, $V(\bar{x}+\delta x,t)$ can be described adequately by a low order power series in δx about $\bar{x};t$. Storage requirements of the defining parameters of the truncated series are modest. Iterative techniques result that determine a new control function $\bar{u}+\delta u$ and thus a new trajectory $\bar{x}+\delta x$ such that the new cost obtained using this control is less than the old cost using $\bar{u}(\cdot)$.

Mayne [33] used a similar approach for discrete-time free-endpoint problems, though he did derive a second-order algorithm for the continuous-time system by allowing the discrete control problem formulation to tend to the continuous-time formulation. Westcott [51] referred to the method as Differential Dynamic Programming. Jacobson [13],[14] developed the notion of D.D.P. further, and showed that the second variation methods of [40], [44] are approximations to Mayne's algorithm. Mc Reynolds [41] obtained an algorithm equivalent to Mayne's.

Jacobson [13], [14], [15], [16] and Jacobson and Mayne [18] applied D.D.P. directly to continuous-time systems and obtained new algorithms by allowing global variations in the control, as opposed to weak variations. Dyer and Mc Reynolds [6], [7] obtained results which were similar to those described by Jacobson [16] for bang-bang control problems. Dyer and Mc Reynolds collated

their results in [8]. Jacobson and Mayne [18] and Gershwin and Jacobson [9] extended the algorithms described by Jacobson [13], [15] to discrete-time problems. D.D.P. has also been applied to stochastic control problems. For references, see [18].

The algorithms mentioned above generally have one feature in common. They invariably require the minimization w.r.t. u of the r.h.s. of the H-J-B equation, for some fixed state $x = \bar{x}(\cdot)$. This leads to a control which could essentially be described as the optimal control for the state $\bar{x}(\cdot)$. In [36], Mayne considered the case of comparing any two controls or control policies. He derived exact expressions for ΔV, the change in cost resulting from a change in the control. Using these expressions, he obtained, by means of a straightforward differential equation approach, generalizations of the equations given previously in the literature. He also demonstrated the usefulness of the expressions obtained for ΔV in obtaining, in particular, sufficient conditions of optimality. Approximations $\Delta\hat{V}$ to ΔV were used to obtain the necessary conditions for optimality for State Constrained problems derived by Jacobson, Lele and Speyer [26]. They were also used to show how necessary and sufficient conditions for the nonnegativity of $\Delta\hat{V}$ could be obtained for Singular control problems.

In this dissertation, we are primarily concerned with the comparison of two arbitrary controls. The basic theory is developed in a unified manner through Differential

Dynamic Programming. Consider the partial differential equation

$$-\frac{\partial V(x,t)}{\partial t} = H(x,w,V_x(x,t),t), \quad t \in [t_o,t_f],$$

where w denotes a control u, or a control policy $k(\cdot)$ which generates $u(\cdot)$. $V(x,t)$ denotes V^u or V^k, according to the context. The Dynamic Programming Technique is then applied to the above equation in a neighbourhood of some nominal trajectory $\bar{x}(\cdot)$ defined by a nominal control $\bar{u}(\cdot)$. Then, providing the state variable δx is kept small, a low order expansion can be made about $\bar{x};t$. This leads to the differential equations satisfied by $a(t) \triangleq V(\bar{x},t) - V^{\bar{u}}(\bar{x},t)$, $V_x(\bar{x},t)$, $V_{xx}(\bar{x},t)$ which were derived by Mayne, [36]. It is then shown that ignoring certain terms in the differential equations satisfied by $V_x(\bar{x},t)$ and $V_{xx}(\bar{x},t)$ respectively, introduces errors in $a(t)$. These can be restricted by allowing either weak variations in the control, or, in the case of global variations in the control, restricting the size of the δx's artificially.

From the differential equations satisfied by $a(t)$, we obtain the exact expressions derived by Mayne [35], for the change ΔV in cost due to a change in control. These expressions, which enable two arbitrary controls to be compared, enable us to obtain certain well known global sufficient conditions of optimality, as well as a local sufficient condition of optimality suggested by Jacobson

in [13] and proved by Mayne in [36]. Estimates for ΔV are then derived, which lead to expressions $\Delta\hat{V}$ which approximate ΔV; one of these is a strong version of the well known second variation formula for the cost function.

It is then shown that the differential equations used for the algorithms described by Kelley [28], Bryson and Denham [2], Jacobson [13], [14], [15], [18], Mayne [33], Mc Reynolds [41], Mc Reynolds and Bryson [40], and Mitter [44], are special cases of the equations derived. These algorithms are outlined briefly for completeness. Also, one of the estimates $\Delta\hat{V}$ for ΔV leads to the first order algorithm of Mayne and Polak, with proven convergence.

We then extend the analysis to the problem with fixed endpoint constraints and free terminal time. Again, the two controls compared are arbitrary, which leads to new differential equations satisfied by partial derivitives of the cost function. The differential equations of Jacobson [13], [18] emerge as a special case of these equations. We then outline how the second-order algorithm of Jacobson, [13], [14], [18], can be adapted to solve the fixed endpoint, free terminal time problem. It is then shown how the problem with terminal inequality constraints can be modified so that the second-order D.D.P. algorithm of Jacobson can be applied to an unconstrained problem. The first-order algorithm, [37], is then extended to handle terminal inequality constraints, [46].

Problems with control constraints are then considered. For the case where the optimal control is continuous for

all $t \in [t_o, t_f]$, we obtain certain differential equations derived by Jacobson [13], [18]. The second-order D.D.P. algorithm can be modified easily to solve these equations, and a first-order algorithm emerges as a special case. For the problem where the optimal control is discontinuous (i.e. of the bang-bang type), the dynamic programming approach is used to obtain certain equations satisfied by partial derivitives of the cost function at switching times times t_i. Algorithms for solving the bang-bang control problem are outlined, which differ in their choice of t_i and the change in switching times δt_i to obtain a new bang-bang control $u(\cdot)$. The choice of δt_i gives jump conditions for the partial derivitives of the cost function which are used in the algorithmm.

We then consider the State Constrained case with no integral cost. It is shown that one of the approximations $\Delta\hat{V}$ to ΔV leads to necessary conditions of optimality. The necessary conditions of Jacobson, Lele and Speyer [26] emerge as a special case.

Necessary and sufficient conditions for the nonnegativity of ΔV for singular problems are then outlined. It is shown, Mayne [36], how one of the expressions for $\Delta\hat{V}$ leads directly to a strong version of the necessary condition of optimality derived by Jacobson, [19]. The generalized Legendre-Clebsch necessary conditions are discussed briefly, and sufficient conditions are derived.

Finally, the concluding section seeks to place the material covered in perspective, and to suggest areas for future research.

CHAPTER 1

DEFINITIONS, ASSUMPTIONS AND PRELIMINARY RESULTS

SYSTEM DESCRIPTION

The continuous-time dynamic system studied in this dissertation is assumed to be described by the following finite set of nonlinear ordinary differential equations

$$\dot{x}(t) = f(x(t),u(t),t); \qquad x(t_o) = x_o \qquad (1.1)$$

where f is an n-dimensional function (nonlinear) of its arguments, $x(t) \in R^n$ describes the state of the system at any time $t \in [t_o,t_f]$, $u(t) \in R^m$ represents the control variables available at time $t \in [t_o,t_f]$. The solution of (1.1) is denoted $x(t;x_o,t_o,u)$. The solution due to the policy k

$$u(t) = k(x(t),t), \quad t \in [t_o,t_f] \qquad (1.2)$$

is denoted by $x(t;x_o,t_o,k)$, i.e. $x(t;x_o,t_o,k)$ is the solution of

$$\dot{x}(t) = f(x(t),k(x(t),t),t); \quad x(t_o) = x_o . \qquad (1.3)$$

The performance of the system is measured by the performance index or "cost functional" :

$$V^u(x_o,t_o) \triangleq \int_{t_o}^{t_f} L(x(t;x_o,t_o,u),u(t),t)dt$$

$$+ F(x(t_f;x_o,t_o,u),t_f) \qquad (1.4)$$

where L and F are scalar functions (nonlinear) of their arguments. The final time t_f may be given explicitly or implicitly.

Similarly, $V^k(x_o,t_o)$ is the cost due to policy k and initial condition (x_o,t_o):

$$V^k(x_o,t_o) \triangleq \int_{t_o}^{t_f} L(x(t;x_o,t_o,k),k(x(t;x_o,t_o,k),t),t)dt$$

$$+ F(x(t_f;x_o,t_o,k),t_f) \quad . \qquad (1.5)$$

The object of the control problem is to choose $u(t), t \in [t_o,t_f]$, so that certain constraints (to be discussed later) are satisfied and V^u given by (1.4) is minimized. (Similarly, for the policy k).

II BASIC ASSUMPTIONS

The following assumptions are necessary at various stages in what follows. The bracketted statements are the extra assumptions required when a policy k rather than a control u is being considered.

The set $S \subset R^n \times R^m \times T$, where $T \triangleq [t_o, t_f]$ is defined by

$S \triangleq \{x,u,T | x \in R^n, u \in \Omega, t \in T\}$ where $\Omega \subset R^m$ is bounded.

H1 $f(x,u,t), L(x,u,t), F(x,t)$ and their partial derivatives w.r.t. x (w.r.t. (x,u)) up to order s exist and are continuous in (x,u,t) for all $(x,u,t) \in S$. ($k(x,t)$ and its partial derivitives w.r.t. x up to order s exist and are continuous in (x,u,t) for all $(x,u,t) \in S$, except at a finite number of times where they have left and right limits).

H2 $||f(x,u,t)|| \leq M(||x||+1)$, $M < \infty$ for all $(x,u,t) \in S$ ($||f(x,k(x,t),t)|| \leq M(||x||+1)$, $M < \infty$ for all $x \in R^n$, $t \in T$).

H3 The partial derivitives $f_t(x,u,t)$, $L_t(x,u,t)$, $F_t(x,t)$ exist and are continuous in (x,u,t) for all $(x,u,t) \in S$.

When H1,H2 include the bracketted statements they will, when necessary, be referred to as H1A, H2A.

Let G denote the class of piecewise continuous functions from T into Ω. A function is defined to be piecewise continuous if it is continuous except at a finite number of points where it has finite left and right limits, and has a finite right limit at t_o and finite left limit at t_f.

Let $\theta(u)$ denote the set of points in T at which u is continuous.

III PROPERTIES OF TRAJECTORIES : Halkin [12]

We now state a standard result, the proof of which can be found in any good book on ordinary differential equations. See for example [4].

PROPOSITION 1.1 Let H1, H2 be satisfied, $s = 1$ for all $u \in G$ there exists an unique solution $x(t;x_o,t_o,u)$ to (1.1). $x(t;x_o t_o,u)$ is absolutely continuous and differentiable over $\theta(u)$.

NOTE: H2 prevents the possibility of a finite escape time.

The following lemma, stated without proof, will be used a great deal later, especially to show under what conditions any approximations we might make are good.

PROPOSITION 1.2 (Generalized Gronwall's Inequality). Suppose that $f(t)$ is a real-valued continuous and piecewise differentiable function defined over T, that $g(t)$ is a real-valued piecewise continuous function defined over T, and that L is a constant. If

$$\dot{f}(t) \leq Lf(t) + g(t)$$

for all $t \in T$ for which $f(t)$ is differentiable and $g(t)$ is continuous, then

$$f(t) \leq e^{Lt}(f(0) + \int_0^t e^{-L\tau} g(\tau)\,d\tau) \quad \text{for all } t \in T.$$

Proof. see [12]

PROPOSITION 1.3 Let H2 be satisfied. Let $u \in G$. Then there exists an $N \in (0,\infty)$ s.t.

$$||x(t;x_o,t_o,u)|| < N .$$

Proof. Follows from H2 and an application of the Gronwall inequality.

We have defined the trajectory $x(t;x_o,t_o,u)$ corresponding to the control function $u(t)$. Let $x(t)$ denote $x(t;x_o,t_o,u)$, $\bar{x}(t)$ denote $x(t;x_o,t_o\bar{u})$, and

$$\delta x(t) \triangleq x(t) - \bar{x}(t) \tag{1.6}$$

The vector valued function $\delta x(t)$ is called the variational trajectory for the control function $u(t)$ w.r.t. the control function $\bar{u}(t)$. We then have

$$\delta\dot{x}(t) = f(x(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t) \text{ for all } t \in \theta(u) \cap \theta(\bar{u}) \tag{1.7}$$

Define

$$\Delta f(t) \triangleq f(\bar{x}(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t) \tag{1.8}$$

$$k(t,u,\bar{u}) \triangleq f(x(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t) - \Delta f(t) - f_x(\bar{x}(t),\bar{u}(t),t)\delta x(t) \tag{1.9}$$

Adding and subtracting like terms

$$\delta\dot{x}(t) = f(x(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t) + f_x(\bar{x}(t),\bar{u}(t),t)\delta x(t) - f_x(\bar{x}(t),\bar{u}(t),t)\delta x(t) + f(\bar{x}(t),u(t),t) - f(\bar{x}(t),u(t),t) + f(\bar{x}(t),\bar{u}(t),t) - f(\bar{x}(t),\bar{u}(t),t)$$

and rearranging

$$\delta\dot{x}(t) = f_x(\bar{x}(t),\bar{u}(t),t)\delta x(t) + f(\bar{x}(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t) + f(x(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t) - f(\bar{x}(t),u(t),t) + f(\bar{x}(t),\bar{u}(t),t) - f_x(\bar{x}(t),\bar{u}(t),t)\delta x(t)$$

we can write

$$\delta\dot{x}(t) = f_x(\bar{x}(t),\bar{u}(t),t)\delta x(t) + \Delta f(t) + k(t,u,\bar{u}) . \quad (1.10)$$

From the theory of differential equations, if $\Phi(t,t_o)$ is the transition matrix for $\delta\dot{x}(t) = f_x(\bar{x},\bar{u},t)\delta x(t)$, then

$$\delta x(t) = \Phi(t,t_o)\delta x(t_o) + \int_{t_o}^{t} \Phi(t,\tau)(\Delta f(\tau) + k(\tau,u,\bar{u}))d\tau \quad (1.11)$$

for all $t \in T$

where $\delta x(t_o) = 0$ when $x(t_o) = \bar{x}(t_o) = x_o$.

Let the metric $d:G \times G \to R$ be defined by

$$d(u,\bar{u}) \triangleq \int_{t_o}^{t_f} ||u(t) - \bar{u}(t)||dt . \quad (1.12)$$

Then, we also have

$$\left[\int_{t_o}^{t_f} ||u(t) - \bar{u}(t)||^2 dt/(t_f-t_o)\right]^{\frac{1}{2}} \geq d(u,\bar{u})/(t_f-t_o). \quad (1.13)$$

The following definition of distance is employed by Halkin [22]. For $E \subset T$, let $\mu(E)$ be the length of set E. If $E = \{t, t \in T, u(t) \neq \bar{u}(t)\}$, then

$$d_1(u,\bar{u}) = \mu(E) . \quad (1.14)$$

We now prove uniform boundedness of variational trajectories.

PROPOSITION 1.4 Let H1, H2 be satisfied, $s = 1$. Let $u,\bar{u} \in G$. If either

(i) $f_u(x,u,t)$ exists and is continuous in (x,u,t) for all $(x,u,t) \in S$, and $d(u,\bar{u}) \leq \varepsilon$

or (ii) $d_1(u,\bar{u}) \leq \varepsilon$

then

$$||\delta x(t)|| \leq c\varepsilon, \; c < \infty, \text{ for all } t \in T .$$

Proof (i) From Proposition 1.3 we have that $x(t;x_o,t_o,u)$ is uniformly bounded over T, for all $u \in G$. From H1, H2 and the fact that $f_u(x,u,t)$ exists, we know that there are constants $c_1, c_2 < \infty$ s.t.

$$||f(x(t),u(t),t) - f(\bar{x}(t),u(t),t)|| \leq c_1||\delta x(t)||$$
$$||f(\bar{x}(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t)|| \leq c_2||u(t) - \bar{u}(t)|| .$$

From Proposition 1.1, the scalar function $||x(t) - \bar{x}(t)||$ is continuous over $\theta(\bar{u}) \cap \theta(u)$. Thus

$$\begin{aligned}\frac{d}{dt}||\delta x(t)|| &\leq ||\dot{x}(t) - \dot{\bar{x}}(t)|| \\ &= ||f(x(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t)|| \\ &\leq ||f(x(t),u(t),t) - f(\bar{x}(t),u(t),t)|| \\ &\quad + ||f(\bar{x}(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t)|| \\ &\leq c_1||\delta x(t)|| + c_2||\bar{u}(t) - u(t)|| .\end{aligned}$$

Applying the Gronwall inequality gives the result.

(ii) We have $||f(\bar{x}(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t)|| \leq L_3$. Then, for all $t \in \theta(u) \cap \theta(\bar{u})$

$$\frac{d}{dt}||x(t) - \bar{x}(t)|| \leq c_1||x(t) - \bar{x}(t)|| + c_2\chi(E)$$

where E is defined as before, and $\chi(E)$ is the characteristic function of the set E. i.e. a function equal to one when $t \in E$, and zero when $t \in T\backslash E$.

By applying Gronwall's inequality, we obtain

$$||\delta x(t)|| \leq c_3 \int_{t_o}^{t} \chi(E)d\tau \quad \text{for all } t \in T,\ c_3 < \infty.$$

i.e. $||\delta x(t)|| \leq c\varepsilon$, $c < \infty$.

Approximation Trajectory

For every $u \in G$, let $\hat{\delta x}(t)$ be a vector function of t defined and continuous over T, differentiable over $\theta(u) \cap \theta(\bar{u})$ and s.t.

$$\dot{\hat{\delta x}}(t) = f_x(\bar{x}(t),\bar{u}(t),t)\hat{\delta x}(t) + \Delta f(t) \tag{1.15}$$

$$\hat{\delta x}(t_o) = 0 \tag{1.16}$$

where $\Delta f(t)$ is defined in equation (1.8).

$\hat{\delta x}(t)$ is called the approximation trajectory for the variational trajectory. Then, letting $\Phi(t,t_o)$ be the transition matrix associated with $f_x(\bar{x},\bar{u},t)$, gives

$$\hat{\delta x}(t) = \int_{t_o}^{t} \Phi(t,\tau)\Delta f(\tau)d\tau \ . \tag{1.17}$$

Thus

$$\delta x(t) - \hat{\delta x}(t) = \int_{t_o}^{t} \Phi(t,\tau)k(\tau,u,\bar{u})d\tau \quad \text{for all } t \in T. \tag{1.18}$$

PROPOSITION 1.5. Let H1, H2 be satisfied, $s = 2$. Let $u,\bar{u} \in G$. Then

(i) there exists a $K_1 < \infty$ s.t.

$$||k(\tau,u,\bar{u})|| \leq K_1||\delta x(t)||^2 \quad \text{for all } t \in T\backslash E$$

(ii) there exists a $K_2 < \infty$ s.t.

$$||k(\tau,u,\bar{u})|| \leq K_2||\delta x(t)|| \quad \text{for all} \quad t \in T.$$

Proof. (i) By definition of $\delta x(t)$, and from Proposition (1.3), there exists $K < \infty$ s.t.

$$||\delta x(t)|| \leq K \qquad \text{for all} \quad t \in T$$

and from (1.9)

$$k(t,u,\bar{u}) = f(x(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t) - \Delta f(t) - f_x(\bar{x}(t),\bar{u}(t),t)\delta x(t) .$$

Now, for all $t \in T$ s.t. $u(t) = \bar{u}(t)$,

$$k(t,u,\bar{u}) = f(x(t),\bar{u}(t),t) - f(\bar{x}(t),\bar{u}(t),t) - f_x(\bar{x}(t),\bar{u}(t),t)\delta x(t)$$

Thus

$$k(t,u,\bar{u}) = f(\bar{x}(t) + \delta x(t),\bar{u}(t),t) - f(\bar{x}(t),\bar{u}(t),t) - f_x(\bar{x}(t),\bar{u}(t),t)\delta x(t) .$$

Denote the RHS by the vector function $\psi(t;\delta x(t),\bar{u})$. Clearly $\psi(t;0,\bar{u}) = 0$, and because of our assumptions $\psi(t;\delta x,\bar{u})$ has second partial derivatives w.r.t. $\delta x(t)$, which are continuous w.r.t. $\delta x(t)$ and piecewise continuous w.r.t. t. Then

$$\psi(t,\delta x(t),\bar{u}) = \left[\frac{\partial\psi(t;\delta x(t),\bar{u})}{\partial\delta x(t)}\right]\Bigg|_{\delta x(t)=0} \delta x(t) + \rho(t;\delta x(t),\bar{u})$$

and there exists $K_1 < \infty$ s.t.

$$||p(t,\delta x(t),\bar{u})|| \leq K_1||\delta x(t)||^2 \qquad \forall t \in T$$

and $\forall \delta x(t)$ with $||\delta x(t)|| \leq K$.

Now

$$\left(\frac{\partial \psi(t;\delta x(t),\bar{u})}{\partial \delta x(t)}\right)\Big|_{\delta x(t) = 0} = 0$$

and so

$$||\psi(t,\delta x(t),\bar{u})|| \leq K_1||\delta x(t)||^2 \qquad \forall t \in T, \forall\ \delta x \text{ with } ||\delta x(t)|| \leq K$$

i.e. $||k(t,u,\bar{u})|| \leq K_1||\delta x(t)||^2 \quad \forall u,u \in G, \forall t \in T\backslash E.$

(ii) $$k(t,u,\bar{u}) = f(\bar{x}(t) + \delta x(t),u(t),t) - f(\bar{x}(t),\bar{u}(t),t) - f_x(\bar{x}(t),\bar{u}(t),t)\delta x(t) \ .$$

Denote the RHS by the vector function $\psi(t,\delta x(t),u)$. Clearly $\psi(t,0,u) = 0$, and because of our assumptions, $\psi(t,\delta x(t),u)$ has a derivative w.r.t. $\delta x(t)$ which is bounded $\forall t \in T$, $\forall u \in G$, and $\forall \delta x(t)$. s.t.

$$||\delta x(t)|| \leq K.$$

Hence, there exists $K_2 < \infty$ s.t.

$||\psi(t;\delta x(t),u)|| \leq K_2||\delta x||$ $\forall t \in T$, $u \in G$ and $\delta x(t)$ s.t. $||\delta x(t)|| \leq K$,

i.e. $||k(t,u,\bar{u})|| \leq K_2||\delta x(t)||$ $\forall t \in T$, $u,\bar{u} \in G$.

It can now be seen that $\delta x(t) - \delta\hat{x}(t)$ is small whenever $k(t,u,\bar{u})$ is small; i.e. from above result whenever $\delta x(t)$ is small; i.e. whenever the two trajectories $x(t)$ and $\bar{x}(t)$ are close together. Intuitively we can see that if $\mu(E)$ is "made" smaller, i.e. if $u(t) = \bar{u}(t)$ over a larger subset of T, then $\delta x(t)$ is made smaller (Proposition 1.4), and then by Proposition 1.5, $\delta\hat{x}(t)$ approximates $\delta x(t)$.

PROPOSITION 1.6. Let H1, H2 be satisfied, $s = 2$. Let $u,\bar{u} \in G$. If either

(i) $f_u(x,u,t), f_{xu}(x,u,t)$ exist and are continuous in (x,u,t) for all $(x,u,t) \in S$, and $d(u,\bar{u}) \leq \varepsilon$

or (ii) $d_1(u,\bar{u}) \leq \varepsilon$

then

$$||\delta x(t) - \delta\hat{x}(t)|| \leq c\varepsilon^2 \ , \ c < \infty \ , \ \forall t \in T.$$

Proof. (i) Define $z(t) \triangleq \delta x(t) - \delta\hat{x}(t)$.

Then $z(t_o) = 0$.

So, from (1.18)

$$\dot{z}(t) = \frac{d}{dt}\left[\int_{t_o}^{t} \Phi(t,\tau)k(\tau,u,\bar{u})d\tau\right]$$

$$= \int_{t_o}^{t} f_x(\bar{x}(t),\bar{u}(t),t)\Phi(t,\tau)k(\tau,u,\bar{u})d\tau + k(t,u,\bar{u})$$

$$= f_x(\bar{x}(t),\bar{u}(t),t)(\delta x(t) - \delta\hat{x}(t)) + k(t,u,\bar{u})$$

$$= f(x,u(t),t) - f(\bar{x}(t),u(t),t) - f_x(\bar{x}(t),\bar{u}(t),t)\delta\hat{x}(t)$$

$$= f_x(\bar{x}(t),u(t),t)\delta x(t) - f_x(\bar{x}(t),\bar{u}(t),t)\delta\hat{x}(t) + r(t)$$

where $\quad ||r(t)|| \leq K_1||\delta x(t)||^2 \qquad K_1 < \infty$

$$\leq K_2\varepsilon^2 \qquad \text{from Proposition 1.4, } K_2 < \infty.$$

Hence

$$\dot{z}(t) = f_x(\bar{x}(t),\bar{u}(t),t)z(t) + r(t)$$
$$+ [f_x(\bar{x}(t),u(t),t) - f_x(\bar{x}(t),\bar{u}(t),t)]\delta x(t) .$$

Now

$$\frac{d}{dt}||z(t)|| \leq ||\dot{z}(t)|| \qquad \text{for all } t \in \theta(u) \cap \theta(\bar{u})$$
$$\leq ||f_x(\bar{x}(t),\bar{u}(t),t)||\,||z(t)|| + ||r(t)||$$
$$+ K_3||u(t) - \bar{u}(t)||\,||\delta x(t)|| .$$

Applying the Gronwall inequality gives the result.

(ii) We have that

$$\dot{z}(t) = f_x(\bar{x}(t),\bar{u}(t),t)z(t) + k(t,u,\bar{u})$$

$$z(t_o) = 0$$

i.e. $$||z(t)|| \leq \int_{t_o}^{t_f} K_1||z(\tau)||d\tau + \int_{t_o}^{t_f} ||k(\tau,u,\bar{u})||d\tau \quad . \tag{1.19}$$

Now

$$\int_{t_o}^{t_f} ||k(\tau,u,\bar{u})||d\tau = \int_E ||k(\tau,u,\bar{u})||d\tau + \int_{T/E} ||k(\tau,u,\bar{u})d\tau$$

from Proposition 1.5 (i), there exists $K_2 < \infty$ s.t.

$$||k(\tau,u,\bar{u})|| \leq K_2||\delta x(t)||^2 \qquad \text{for all } t \in T\backslash E$$

and 1.5 (ii), there exists $K_3 < \infty$ s.t.

$$||k(\tau,u,\bar{u})|| \leq K_3||\delta x(t)|| \qquad \text{for all } t \in E.$$

From Proposition 1.4 (ii), we have, remembering $d(\bar{u},u) = \mu(E)$),

$$\int_{t_o}^{t_f} ||k(\tau,u,\bar{u})||d\tau \leq K_4\varepsilon^2 \quad .$$

Applying an integral version of Proposition 1.2 to equation (1.19) gives the result.

The hypothesis in Proposition 1.4 (i), (1.6 (i)) can be replaced by

(i) H1A, H2A are satisfied, $s = 1$, $(s=2)$ and $d(u,\bar{u}) \leq \varepsilon$.

PROPOSITION 1.7 Let $u,\bar{u} \in G$. If either

(i) H1, H2 are satisfied, $s = 2$, f is linear in u, and $d(u,\bar{u}) \leq \varepsilon$ or $d_1(u,\bar{u}) \leq \varepsilon$, or

(ii) H1A, H2A are satisfied, $s = 2$ and $\|u(t)-\bar{u}(t)\| \leq \varepsilon$ for all $t \in T$. Then

$$\|\delta x(t)-\delta\hat{x}(t)\| \leq c\varepsilon^2, \quad c < \infty, \ \forall t \in T$$

where $\delta\hat{x}(t)$ is the solution of

$$\delta\dot{\hat{x}}(t) = f_x(\bar{x}(t),\bar{u}(t),t)\delta\hat{x}(t)+f_u(\bar{x}(t),\bar{u}(t),t)\delta u \qquad 1.20$$

$$\delta x(t_o) = 0 \qquad 1.21$$

and

$$\delta u(t) \triangleq u(t) - \bar{u}(t)$$

Proof

(i) $\Delta f(t) = f_u(\bar{x}(t),\bar{u}(t),t)\delta u(t)$ and the result follows immediately from Proposition 1.6

(ii) $\Delta f(t) = f_u(\bar{x}(t),\bar{u}(t).t)\delta u(t) + r(t)$

where $\|r(t)\| \leq K_1\varepsilon^2$, $K_1 < \infty$

Therefore

$$\delta x(t)-\delta\hat{x}(t) = \int_{t_o}^{t} \Phi(t,\tau)[k(\tau,u,\bar{u}) + r(t)]d\tau$$

so

$$\dot{z}(t) = \int_{t_o}^{t} f_x(\bar{x}(t),\bar{u}(t),t)\Phi(t,\tau)[k(\tau,u,\bar{u})+r(\tau)]d\tau + k(t,u,\bar{u})+r(t)$$

$$= f_x(\bar{x}(t),\bar{u}(t),t)z(t) + k(t,u,\bar{u}) + r(t)$$

Using the proof of Proposition 1.6 (i) we have that an extra term of norm $K_2\varepsilon^2$ is introduced. □

CHAPTER 2

PROPERTIES OF THE COST FUNCTION

1. DIFFERENTIABILITY OF THE COST FUNCTION : Mayne [35].

PROPOSITION 2.1 Let H1, H2 be satisfied, $u \in G$. Then

(i) $V^u(x(t),t)$ and its partial derivatives w.r.t. x up to order s exist and are continuous in (x,t),

(ii) $V^u_t(x(t),t)$ and its partial derivatives w.r.t. x up to order s-1 exist and are continuous in (x,t), except where u is discontinuous.

Let H1A, H2A be satisfied. Then

(iii) $V^k(x(t),t)$ and its partial derivatives w.r.t. x up to order s exist and are continuous in (x,t),

(iv) $V^k_t(x(t).t)$ and its partial derivatives w.r.t. x up to order s-1 exist and are continuous in (x,t), except where k is discontinuous.

Proof: Consider $V^u(x(t_1),t_1)$, and the following system:

$$\dot{z}(t) = g(z(t).t)) \tag{2.1}$$

where

$$z(t) \triangleq (x_o(t),x(t))$$

$$g(t) \triangleq (L(x(t),u(t),t),\ f(x(t),u(t)t))$$

i.e. $$\dot{x}_o(t) = L(x(t),u(t),t). \tag{2.2}$$

Then, if

$$z(t_1) \triangleq (0,x(t_1)) \tag{2.3}$$

$$H(x_o(t),x(t)) \triangleq x_o(t) + F(x)$$

we have

$$V^u(x(t_1),t_1) = H(z(t_f;z(t_1),t_1)) \tag{2.4}$$

and H is s times continuously differentiable.

(i) (McShane [42], Theorem 69.4).

Let $z(t;z_1,t_1)$ denote the solution of equ (2.1) with initial condition (z_1,t_1), z_1 given in equ (2.3). (The dependance on u is omitted for convenience).

$z(t,z_1,t_1)$ and its partial derivitives w.r.t. z_1 of all orders less than or equal to s exist, and are continuous in (t,z_1,t_1) for all $t,t_1 \in T$, all $z_1 \in R^{n+1}$. Setting $t = t_f$ and using equation 2.4, gives (i).

(ii) Let $t,t_1 \in \theta(u)$. Then there exists $\varepsilon > 0$ s.t. if $|\delta t_1| < \varepsilon$, u is continuous on $[t_1,t_1+\delta t_1]$. Let z^i denote the i^{th} component of z, g^i the i^{th} component of g etc.

Define

$$\phi^i(\delta t_i) \triangleq [z^i(t;z_1,t_1+\delta t_1) - z^i(t;z_1,t_1)]/\delta t_1 \qquad (2.5)$$

Now

$$z^i(t;z_1,t_1+\delta t_1) = z^i(t;y,t_1)$$

where

$$y^i \triangleq z^i(t_1;z_1;t_1+\delta t_1)$$
$$= z_1^i - g^i(z(t_1^*;z_1,t_1+\delta t_1),t_1^*)\delta t_1 \qquad (2.6)$$

where

$$t_1^* \in [t_1,t_1+\delta t_1].$$

Now

$$z^i(t;y,t_1) - z^i(t;z_1,t_1) = \sum_{j=1}^{n} (\partial z^i/\partial z_1^j)(t;z_1^*,t_1)(y^j - z_1^j) \qquad (2.7)$$

where

$$z_1^* = \theta z_1 + (1-\theta)(y-z_1) \text{ for some } 0 \le \theta \le 1.$$

Hence, substituting (2.6) into (2.7) and using (2.5) gives:

$$\phi^i(\delta t_1) = \sum_{j=1}^{n} (\partial z^i/\partial z_1^j)(t;z_1^*,t_1)[-g^j(z(t_j^*;z_1,t_1+\delta t_1),t_j^*)]$$

As $\delta t_1 \to 0$, $t_i^* \to t_1$, $y \to z_1$, $z_i^* \to z_1$.

Hence from the continuity of z_{z_1} and g, z_{t_1} exists for all $t,t_1 \in \theta(u)$ and is given by

$$z_{t_1}(t;z_1,t_1) = -z_{z_1}(t;z_1,t_1)g(z_1,t_1) \qquad (2.8)$$

The RHS of (2.8) and its partial derivatives w.r.t. z up to order s-1 exist and are continuous in (z_1,t_1) for all $t,t_1 \in \theta(u)$, possessing left and/or right limits at the remaining points. Allowing $t \to t_f$ from the left, and using equation (2.4) yields (ii).

(iii) & (iv) With the extra assumptions on k,

$$\bar{f}(x,t) \triangleq f(x,k(x.t),t)$$

satisfies the original assumptions on f. Hence, the previous discussion is applicable to V^k.

□

II PARTIAL DIFFERENTIAL EQUATION FOR THE COST FUNCTION

PROPOSITION 2.2: Let H1, H2 be satisfied, $u \in G$. Then

$$-V_t^u(x(t),t) = H(x(t),u(t),V_x^u(x(t),t),t) \tag{2.9}$$

for all $t \in \theta(u)$, where

$$H(x(t),u(t),\lambda(t),t) \triangleq L(x(t),u(t),t) + \lambda^T(t) f(x(t),u(t),t) \tag{2.10}$$

For all $t \in T - \theta(u)$, the left and/or right limits are given by the same expression with u(t) having the corresponding left and/or right limits.

Proof: From (1.4)

$$V^u(x(t),t) = \int_t^{t_f} L(x(\tau),u(\tau),\tau)\,d\tau + F(x(t_f),t_f) \tag{2.11}$$

Now we apply the Dynamic Programming Technique to equation (2.11).

$$V^u(x(t),t) = \int_t^{t+\Delta t} L(x(\tau),u(\tau),\tau)\,d\tau + \int_{t+\Delta t}^{t_f} L(x(\tau),u(\tau),\tau)\,d\tau + F(x(t_f),t_f)$$

$$= \int_t^{t+\Delta t} L(x(\tau),u(\tau),\tau)\,d\tau + V^u(x(t+\Delta t),t+\Delta t)$$

$$= \int_t^{t+\Delta t} L(x(\tau),u(\tau),\tau)\,d\tau + V^u(x+\Delta x,t+\Delta t).$$

Now, make Δt small, and approximate $u(\tau)$ over $[t,t+\Delta t]$ by the constant value $u(t)$, where for any $t_1 \in T - \theta(u)$, $t_1 \notin [t,t+\Delta t]$. Then

$$V^u(x(t),t) = L(x(t),u(t),t)\Delta t + V^u(x+\Delta x,t+\Delta t)$$

Now, because of our assumptions, expand $V^u(x+\Delta x,t+\Delta t)$ in a power series about the point (x,t) for all $t \in \theta(u)$:

$$V^u(x(t),t) = L(x(t),u(t),t)\Delta t + V^u(x(t),t) + \frac{\partial V^u}{\partial t}(x(t),t)\Delta t$$

$$+ \langle V_x^u(x(t),t), f(x(t),u(t),t)\rangle . \Delta t + O(\Delta t^2)$$

i.e.

$$-\frac{\partial V^u}{\partial t}(x(t),t)\Delta t = L(x(t),u(t),t)\Delta t$$

$$+ \langle V_x^u(x(t),t), f(x(t),u(t),t)\rangle \Delta t + O(\Delta t^2)$$

V^u, $\partial V^u/\partial t$ are independent of $u(t)$. Dividing by Δt and allowing Δt to tend to zero, using equ (2.10) gives the result. □

Note: We cannot expand $V^u(x+\Delta x, t+\Delta t)$ in a power series about the point $(x(t_1),t_1)$, say, where $t_1 \in T - \theta(u)$, because V_t^u is not continuous in (x,t) at t_1 - from Proposition 2.1 (ii).

If H1A, H2A are satisfied for some policy $k(x,t)$, then we have

$$-\frac{\partial V^k}{\partial t}(x(t),t) = H(x(t),k(x(t),t),V_x^k(x(t),t),t) \qquad (2.12)$$

for $t \in \theta(k)$, where

$$H(x(t),k(x(t),t),\lambda(t),t) \triangleq L(x(t),k(x(t),t),t)$$

$$+ \lambda^T(t) f(x(t),k(x(t),t),t). \qquad (2.13)$$

III THE ADJOINT DIFFERENTIAL EQUATION

Usually in the control literature, the solution $\bar{\lambda}(t)$ of the adjoint differential equation is given. Define

$$\bar{\lambda}(t) \overset{\Delta}{=} V_x^{\bar{u}}(\bar{x}(t),t) \tag{2.14}$$

$$\bar{P}(t) \overset{\Delta}{=} V_{xx}^{\bar{u}}(\bar{x}(t),t). \tag{2.15}$$

PROPOSITION 2.3 : Let $\bar{u} \in G$

(i) Let H1, H2 be satisfied, $s = 2$. Then $\bar{\lambda}(t)$ is the solution of

$$-\dot{\bar{\lambda}}(t) = H_x(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t), \quad \forall \ t \in T \tag{2.16}$$

$$\bar{\lambda}(t_f) = F_x(\bar{x}(t_f),t_f) \tag{2.17}$$

(ii) Let H1A, H2A be satisfied, $s = 3$. Then $\bar{P}(t)$ is the solution of

$$-\dot{\bar{P}}(t) = H_{xx}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) + f_x^T(\bar{x}(t),\bar{u}(t),t)\bar{P}(t) + \bar{P}(t)f_x(\bar{x}(t),\bar{u}(t),t), \ \forall \ t \in T \tag{2.18}$$

$$\bar{P}(t_f) = F_{xx}(\bar{x}(t_f),t_f) \tag{2.19}$$

Proof

(i) $$\frac{d}{dt}V_x^{\bar{u}}(\bar{x}(t),t) = V_{xt}^{\bar{u}}(\bar{x}(t),t) + V_{xx}^{\bar{u}}(\bar{x}(t),t)f(\bar{x}(t),\bar{u}(t),t) \tag{2.20}$$

and by Proposition 2.2

$$-V_t^{\bar{u}}(\bar{x}(t),t) = H(\bar{x}(t),\bar{u}(t),V_x^{\bar{u}}(\bar{x}(t),t),t)$$

$$= L(\bar{x}(t),\bar{u}(t),t) + V_x^{\bar{u}}(\bar{x}(t),t)f(\bar{x}(t),\bar{u}(t),t)$$

for all $t \in \theta(\bar{u})$, and so

$$-V_{xt}^{\bar{u}}(\bar{x}(t),t) = H_x(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) + V_{xx}^{\bar{u}}(\bar{x}(t),t)f(\bar{x}(t),\bar{u}(t),t) \tag{2.21}$$

Substituting equation (2.21) into equation (2.20) gives equation (2.16), for all $t \in \theta(\bar{u})$. The result follows from Proposition 2.1(i).

(ii) $$\frac{d}{dt}V^{\bar{u}}_{xx}(\bar{x}(t),t) = V^{\bar{u}}_{xxt}(\bar{x}(t),t) + \sum_{i=1}^{n} V^{\bar{u}}_{xxx_i}(\bar{x}(t),t) f_i(\bar{x}(t),\bar{u}(t),t) \qquad (2.22)$$

and from equation (2.21)

$$-V^{\bar{u}}_{xxt}(\bar{x}(t),t) = H_{xx}(\bar{x}(t),\bar{u}(t),t) + f_x^T(\bar{x}(t),\bar{u}(t),t)\bar{P}(t)$$

$$+ \bar{P}(t) f_x(\bar{x}(t),\bar{u}(t),t)$$

$$+ \sum_{i=1}^{n} V^{\bar{u}}_{xxx_i}(\bar{x}(t),t) f_i(\bar{x}(t),\bar{u}(t),t) \qquad (2.23)$$

Substituting (2.23) into equation (2.22) gives the result.

□

IV EXACT DIFFERENTIAL EQUATIONS FOR $V_x(\bar{x}(t),t)$, $V_{xx}(\bar{x}(t),t)$

It is the purpose of this section to obtain the general differential equations obtained by Mayne [36] for $V_x(\bar{x}(t),t)$ and $V_{xx}(\bar{x}(t),t)$, where V denotes V^u or V^k according to the context. However, we shall use a dynamic programming approach, similar to that used in [13], [15], [18], except that we will compare two arbitrary nonoptimal controls $u, \bar{u} \in G$.

We will need the following results:

$$\frac{d}{dt}V(\bar{x}(t),t) = V_t(\bar{x}(t).t) + V_x(\bar{x}(t),t)f(\bar{x}(t),\bar{u}(t),t) \tag{2.24}$$

$$\frac{d}{dt}V_x(\bar{x}(t),t) = V_{xt}(\bar{x}(t),t) + V_{xx}(\bar{x}(t),t)f(\bar{x}(t),\bar{u}(t),t) \tag{2.25}$$

$$\frac{d}{dt}V_{xx}(\bar{x}(t),t) = V_{xxt}(\bar{x}(t),t) + \langle V_{xxx}(\bar{x}(t),t),f(\bar{x}(t),\bar{u}(t),t)\rangle \tag{2.26}$$

Define

$$a(t) = V(\bar{x}(t),t) - V^{\bar{u}}(\bar{x}(t),t) \tag{2.27}$$

where V denotes V^u or V^k according to the context. We have, from equation (2.27) that

$$\Delta V = a(t_o).$$

Define

$$\lambda(t) = V_x(\bar{x}(t),t) \tag{2.28}$$

$$P(t) = V_{xx}(\bar{x}(t),t) \tag{2.29}$$

Assume that the cost $V^u(x,t)$ is 'smooth enough' to allow a power series expansion in δx about $\bar{x}(t)$, the trajectory associated with the control $\bar{u}(t)$. Expanding to

first order

$$V^u(\bar{x}+\delta x,t) = V^u(\bar{x}(t),t)+V^u_x(\bar{x}(t),t)\delta x + \text{higher order terms in } \delta x.$$

Substituting into equation (2.9), we have, in terms of the trajectory $\bar{x}(t)$:

$$-\frac{\partial}{\partial t}V^u(\bar{x}(t),t) - \frac{\partial}{\partial t}V^u_x(\bar{x}(t),t)\delta x + H(0) \tag{2.30}$$

$$= L(\bar{x}+\delta x,u,t) + \langle V^u_x(\bar{x},t) + V^u_{xx}(\bar{x},t)\delta x + H(0), f(\bar{x}+\delta x,u,t)\rangle$$

If δx is sufficiently small, then equation (2.30) can be expanded to first order in δx with error $O(\delta x^2)$.

At this stage we introduce the concept of weak and strong variations. Define a weak variation in the following way:

If we have a trajectory $\bar{x}(t), t \in T$, then $\delta x(t)$ is a weak variation of $\bar{x}(t)$ iff

$$\|\delta x(t)\| \leq \varepsilon \qquad \forall t \in T \tag{2.31}$$

$$\|\delta \dot{x}(t)\| \leq \varepsilon \qquad \varepsilon > 0 \qquad \forall t \in T \tag{2.32}$$

A strong variation is when only condition (2.32) holds. How to restrict the size of the δx's artificially for strong variations will be discussed later in Chapter 5-II-4.

Expanding the cost $V^u(x,t)$ to second order about $\bar{x}(t)$ and substituting the result into equation (2.9)

gives, in terms of $\bar{x}(t)$:

$$-\frac{\partial}{\partial t}V^{\mu}(\bar{x},t)-\frac{\partial}{\partial \bar{x}}V^{\mu}_{x}(\bar{x},t)\,\delta x-\tfrac{1}{2}<\delta x,V^{\mu}_{xx}(\bar{x},t)\,\delta x>+H(0) \qquad (2.33)$$

$$=L(\bar{x}+\delta x,u,t)+<V^{\mu}_{x}(\bar{x},t)+V^{\mu}_{xx}(\bar{x},t)\,\delta x+\tfrac{1}{2}V^{\mu}_{xxx}(\bar{x},t)\,\delta x\delta x+H(0),f(\bar{x}+\delta x,u,t)>$$

We use these equations to obtain differential equations for $a(t)$, $\lambda(t)$ and $P(t)$. Throughout the next result we assume that the δx's remain small enough for our expansions to be true.

PROPOSITION 2.4: Let $u,\bar{u} \in G$.

(i) Let H1, H2 be satisfied, $s = 2$. Then, if the above assumption is satisfied, $a(t)$, $\lambda(t)$ are the solutions of

$$-\dot{a}(t) = H(\bar{x}(t),u(t),\lambda(t),t) - H(\bar{x}(t),\bar{u}(t),\lambda(t),t) \qquad (2.34)$$

$$-\dot{\lambda}(t) = H_{x}(\bar{x}(t),u(t),\lambda(t),t) + P(t)\Delta f(t) \qquad (2.35)$$

with boundary conditions

$$a(t_f) = 0 \qquad (2.36)$$

$$\lambda(t_f) = F_{x}(\bar{x}(t_f),t_f) \qquad (2.37)$$

(ii) Let H1, H2 be satisfied, $s = 3$. Then if the above assumption is satisfied, $P(t)$ is the solution of

$$-\dot{P}(t) = H_{xx}(\bar{x}(t),u(t),\lambda(t),t)+f_x^T(\bar{x}(t),u(t),t)P(t) \qquad (2.38)$$

$$+ P(t)f_x(\bar{x}(t),u(t),t)+\sum_{i=1}^{n} V^{\mu}_{xxx_i}(\bar{x}(t),t)\Delta f_i(t)$$

with boundary conditions

$$P(t_f) = F_{xx}(\bar{x}(t_f),t_f) \qquad (2.39)$$

$\Delta f(t)$ is defined in equation (1.8)

Proof:

(i) Expand the RHS of equ (2.30) to first order in δx. Then, for δx sufficiently small, we have, using equ (2.27):

$$-\frac{\partial}{\partial t}V^{\bar{\mu}}(\bar{x}(t),t) - \frac{\partial a}{\partial t}(t) - \frac{\partial}{\partial t}V^{\mu}_{x}(\bar{x}(t),t)\delta x$$

$$= H(\bar{x}(t),u(t),\lambda(t),t)+H_x(\bar{x}(t),u(t),\lambda(t),t)\delta x+\langle P(t)\delta x, f(\bar{x}(t),u(t),t)\rangle$$

Since equality holds for all δx sufficiently small, we may equate like powers of δx. This gives

$$-\frac{\partial}{\partial t}V^{\bar{\mu}}(\bar{x}(t),t) - \frac{\partial a}{\partial t}(t) = H(\bar{x}(t),u(t),\lambda(t),t)$$

$$-\frac{\partial}{\partial t}V^{\mu}_{xx}(\bar{x},t) = H_x(\bar{x}(t),u(t),\lambda(t),t)+P(t)f(\bar{x}(t),u(t),t)$$

Now, because

$$\frac{d}{dt}V^{\bar{\mu}}(\bar{x}(t),t) = -L(\bar{x}(t),\bar{u}(t),t)$$

we have from equations (2.24) - (2.25), equations (2.34), (2.35).

To prove the terminal conditions

$$a(t_f) = V^u(\bar{x}(t_f),t_f) - V^{\bar{u}}(\bar{x}(t_f),t_f)$$

$$= F(\bar{x}(t_f),t_f) - F(\bar{x}(t_f),t_f) = 0$$

$$\lambda(t_f) = V^u_x(\bar{x}(t_f),t_f)$$

$$= F_x(\bar{x}(t_f),t_f)$$

(ii) Expand the RHS of equ (2.33) to second order in δx. Then, for δx sufficiently small, we have, using equ (2.27):

$$-\frac{\partial}{\partial t}V^{\bar{u}}(\bar{x}(t),t) - \frac{\partial a}{\partial t}(t) - \frac{\partial}{\partial t}V^u(\bar{x}(t),t) - \tfrac{1}{2}\langle\delta x, \frac{\partial}{\partial t}V^u_{xx}(\bar{x},t)\delta x\rangle$$

$$= H(\bar{x}(t),u(t),\lambda(t),t) + H_x(\bar{x}(t),u(t),\lambda(t),t)\delta x + \langle P(t)\ \delta x, f(\bar{x}(t),u(t),t)\rangle$$

$$+\langle\delta x, (\tfrac{1}{2}H_{xx}(\bar{x}(t),u(t),\lambda(t),t) + P(t)f_x(\bar{x}(t),u(t),t)$$

$$+\tfrac{1}{2}V^u_{xxx}(\bar{x}(t),t)f(\bar{x}(t),u(t),t))\delta x\rangle$$

Using equs (2.34), (2.35), we are left with the coefficients of (δx^2). Finally, using equ (2.26) gives (2.38).

The terminal condition:

$$P(t_f) \triangleq V^u_{xx}(\bar{x}(t_f),t_f) = F_{xx}(\bar{x}(t_f),t_f)$$

□

The differential equation for $V_x^u(\bar{x}(t),t)$ has a term involving $V_{xx}^u(\bar{x}(t),t)$. This term arises because we are interested in the rate of change of V_x^u not along x, the trajectory generated by u(t), but along $\bar{x}(t)$. The d.e. for $V_{xx}^u(\bar{x}(t),t)$ is given by (2.38).

In the rest of this section we concern ourselves with obtaining differential equations for a(t), $V_x^k(x(t),t)$, $V_{xx}^k(x(t),t)$ where k(x,t) is a policy which generates $(x(t),u(t))$.

Suppose that we have a control $\bar{u}(t) \in G$ with associated trajectory $\bar{x}(t)$. Then assume

$$k(x(t),t) = k(\bar{x}+\delta x,t)$$
$$= k(\bar{x},t) + K_x(\bar{x},t)\delta x + \tfrac{1}{2}\delta x^T k_{xx}(\bar{x},t)\delta x \qquad (2.40)$$

to second order in δx, assuming that δx is small enough.

Now, in the same way as before, assume that $V^k(x(t),t)$ is 'smooth enough' to allow a power series expansion in δx about $\bar{x}(t)$. Expanding to second order

$$V^k(\bar{x}+\delta x,t) = V^k(\bar{x},t) + V_x^k(\bar{x},t)\delta x + \tfrac{1}{2}\delta x^T V_{xx}^k(\bar{x},t)\delta x$$
$$+ \text{higher order terms in } \delta x$$

Substituting into equation (2.12), using equ (2.27):

$$-\frac{\partial}{\partial t}V^{\bar{u}}(\bar{x}(t),t)-\frac{\partial a}{\partial t}(t)-\frac{\partial}{\partial t}V_x^k(\bar{x}(t),t)\,\delta x-\tfrac{1}{2}<\delta x,\frac{\partial}{\partial t}V_{xx}^k(\bar{x}(t),t)\,\delta x>+H(0)$$

$$=L(\bar{x}+\delta x,k(\bar{x}+\delta x,t),t)+<V_x^k(\bar{x},t)+V_{xx}^k(\bar{x},t)\,\delta x$$

$$+\tfrac{1}{2}V_{xxx}^k(\bar{x},t)\,\delta x\delta x+H(0),\ f(\bar{x}+\delta x,k(\bar{x}+\delta x,t),t)>. \qquad (2.41)$$

Since we have not introduced any approximations this equation is still valid globally.

(2.41) is generally impossible to solve numerically because of the possibility of infinite storage space required for the parameters of the power series expansion. However, if δx is sufficiently small, then $V^k(\bar{x}+\delta x,t)$ can be expanded to second order in δx with error (δx^3). Equ (2.41) becomes:

$$-\frac{\partial}{\partial t}V^{\bar{u}}(\bar{x}(t),t)-\frac{\partial a}{\partial t}(t)-\frac{\partial}{\partial t}V_x^k(\bar{x}(t),t)\,\delta x-\tfrac{1}{2}<\delta x,\frac{\partial}{\partial t}V_{xx}^k(\bar{x}(t),t)\,\delta x>$$

$$=L(\bar{x}+\delta x,k(\bar{x}+\delta x,t),t)+<V_x^k(\bar{x},t)+V_{xx}^k(\bar{x},t)\,\delta x$$

$$+\tfrac{1}{2}V_{xxx}^k(\bar{x}t)\,\delta x\delta x,f(\bar{x}+\delta x,k(\bar{x}+\delta x,t),t)> \qquad (2.42)$$

We are now in a position to obtain differential equations for $a(t)$, $V_x^k(\bar{x},t)$, $V_{xx}^k(\bar{x},t)$. Again, we assume that the δx's remain small enough so that the expansions we make w.r.t. δx are valid. (For the case of strong variations, we restrict the size of δx artificially using a procedure explained in detail in Ch 5-II, and intuitively based on the result of Proposition 1.4).

Define

$$K(t) \triangleq k_x(\bar{x}(t),t) \tag{2.43}$$

$$\gamma(t) \triangleq k_{xx}(\bar{x}(t),t) \tag{2.44}$$

PROPOSITION 2.5 Let H1A, H2A be satisfied, s = 3. Let $\bar{u} \in G$. Then, providing the above assumption holds, $a(t)$, $\lambda(t)$ and $P(t)$ satisfy the following differential equations:

$$-\dot{a}(t) = H(\bar{x}(t),k(\bar{x}(t),t),\lambda(t),t) - H(\bar{x}(t),\bar{u}(t),\lambda(t),t) \tag{2.45}$$

$$-\dot{\lambda}(t) = H_x(\bar{x}(t),k(\bar{x}(t),t),\lambda(t),t) + P(t)\Delta f(t) + K^T(t)H_u(\bar{x}(t),k(\bar{x}(t),t),\lambda(t),t) \tag{2.46}$$

$$\begin{aligned}-\dot{P}(t) = {} & H_{xx}(\bar{x}(t),k(\bar{x}(t),t),\lambda(t),t)+f_x^T(\bar{x}(t),k(\bar{x}(t),t),t)P(t) \\ & + P(t)f_x(\bar{x}(t),k(\bar{x}(t),t),t) \\ & + K^T(t)[H_{ux}(\bar{x}(t),k(\bar{x}(t),t),\lambda(t),t)+f_u^T(\bar{x}(t),k(\bar{x}(t),t),t)P(t)] \\ & + [H_{ux}(\bar{x}(t),k(\bar{x}(t),t)\lambda(t),t)+f_u^T(\bar{x}(t),k(\bar{x}(t),t),t)P(t)]^TK(t) \\ & + K^T(t)H_{uu}(\bar{x}(t),k(\bar{x}(t),t),\lambda(t),t)K(t) \\ & + \sum_{i=1}^{n} H_u^i(\bar{x}(t),k(\bar{x}(t),t),\lambda(t),t)\gamma_i(t) \\ & + \sum_{i=1}^{n} V^k{}_{xxx_i}(\bar{x}(t),t)\Delta f_i(t)\end{aligned} \tag{2.47}$$

with boundary conditions

$$a(t_f) = 0 \tag{2.48}$$

$$\lambda(t_f) = F_x(\bar{x}(t_f), t_f) \tag{2.49}$$

$$P(t_f) = F_{xx}(\bar{x}(t_f), t_f) \tag{2.50}$$

where

$$\Delta f(t) = f(\bar{x}(t), k(\bar{x}(t), t), t) - f(\bar{x}(t), \bar{u}(t), t) \tag{2.51}$$

and H_u^i is the i^{th} component of H_u, γ_i is the i^{th} component of γ, Δf_i is the i^{th} component of Δf etc.

Proof: From equ (2.42) we have

$$-\frac{\partial}{\partial t}V^{ij}(\bar{x},t) - \frac{\partial a}{\partial t}(t) - \frac{\partial}{\partial t}V_x^k(\bar{x},t)\delta x - <\delta x, \frac{\partial}{\partial t}V_{xx}^k(\bar{x},t)\delta x>$$

$$= H(\bar{x}+\delta x, k(\bar{x}+\delta x,t), \lambda(t), t) + <P(t)\delta x$$

$$+ \tfrac{1}{2}V_{xxx}^k(\bar{x},t)\delta x\delta x, f(\bar{x}+\delta x, k(\bar{x}+\delta x,t),t)> \tag{2.52}$$

Expanding the RHS of (2.52) about $\bar{x}, k(\bar{x}(t),t)$, we obtain

$$H(\bar{x}, k(\bar{x},t), \lambda, t)$$

$$+<H_x(\bar{x}, k(\bar{x},t), \lambda, t) + P(t)f(\bar{x}, k(\bar{x},t),t) + K^T(t)H_u(\bar{x}, k(\bar{x},t), \lambda, t), \delta x>$$

$$+ \tfrac{1}{2}<H_u(\bar{x}, k(\bar{x},t), \lambda, t), k_{xx}(\bar{x},t)\delta x\delta x>$$

$$+ \tfrac{1}{2}<K(t)\delta x, (H_{ux}(\bar{x}, k(\bar{x},t), \lambda, t) + f_u^T(\bar{x}, k(\bar{x},t),t)P(t))\delta x>$$

$$+ \tfrac{1}{2}\langle \delta x(H_{ux}(\bar{x},k(\bar{x},t),\lambda,t) + f_u^T(\bar{x},k(\bar{x},t),t)P(t)),K(t)\delta x\rangle$$

$$+ \tfrac{1}{2}\langle K(t)\delta x, H_{uu}(\bar{x},k(\bar{x},t)\lambda,t)K(t)\delta x\rangle$$

$$+ \tfrac{1}{2}\langle \delta x, (H_{xx}(\bar{x},k(\bar{x},t),\lambda,t)+f_x^T(\bar{x},k(\bar{x},t),t)P(t)+P(t)f(\bar{x},k(\bar{x},t),t))\delta x\rangle$$

$$+ \tfrac{1}{2}\langle V_{xxx}^k(\bar{x},t)\delta x\delta x, f(\bar{x},k(\bar{x},t),t)\rangle + \text{higher order terms.}$$

Now, for all δx sufficiently small, the higher order terms can be neglected and the coefficients of like powers of δx on the LHS and RHS of equ (2.52) may be equated:

$$-\frac{\partial}{\partial t}V^{\bar{u}}(\bar{x},t) - \frac{\partial a}{\partial t} = H(\bar{x},k(\bar{x},t),\lambda,t)$$

$$-\frac{\partial}{\partial t}V_x^k(\bar{x},t) = H_x(\bar{x},k(\bar{x},t),\lambda,t) + P(t)f(\bar{x},k(\bar{x},t),t) + K^T(t)H_u(\bar{x},k(\bar{x},t),\lambda,t)$$

$$-\frac{\partial}{\partial t}V_{xx}^k(\bar{x},t) = H_{xx}(\bar{x},k(\bar{x},t),\lambda,t) + f_x^T(\bar{x},k(\bar{x},t),t)P(t) + P(t)f_x(\bar{x},k(\bar{x},t),t) + K^T(t)[H_{ux}(\bar{x},k(\bar{x},t),\lambda,t)+f_u^T(\bar{x},k(\bar{x},t),t)P(t)] + [H_{ux}(\bar{x},k(\bar{x},t),\lambda,t)+f_u^T(\bar{x},k(\bar{x},t),t)P(t)]^TK(t) + K^T(t)H_{uu}(\bar{x},k(\bar{x},t),\lambda,t)K(t) + \sum_{i=1}^{n} H_u^i(\bar{x},k(\bar{x},t),\lambda,t)\gamma_i(t) + \sum_{i=1}^{n} V_{xxx_i}^k(\bar{x},t)f(\bar{x},k(\bar{x},t),t)$$

Using (2.24) - (2.26) in the above equations, gives equations (2.45) - (2.47).
The boundary conditions : see Proposition 2.4. □

The equations obtained above are the same as those obtained by Mayne [36], which are the nonoptimal version of those derived by Jacobson [13], [15], [18].

V APPROXIMATE DIFFERENTIAL EQUATIONS FOR $V_x(\bar{x}(t),t)$, $V_{xx}(\bar{x}(t),t)$: Mayne [35].

Proposition 2.4 is not useful computationally for a first order algorithm because of the appearance of the $V^u_{xx}(\bar{x}(t),t)$ term which is found by solving equations (2.38) and (2.39). In the following result we estimate the error arising in $a(t)$, $\lambda(t)$ from the ommission of the $V^u_{xx}(\bar{x},t)\Delta f(t)$ term.

PROPOSITION 2.6: Let H1, H2 be satisfied, $s = 2$. If either

(i) $f_u(x,u,t)$ exists and is continuous in (x,u,t) for all $(x,u,t) \in S$, and $d(u,\bar{u}) \leq \varepsilon$, or

(ii) $d_1(u,\bar{u}) \leq \varepsilon$,

then, $\hat{a}(t)$, $\hat{\lambda}(t)$, where:

$$-\dot{\hat{a}}(t) = H(\bar{x}(t),u(t),\hat{\lambda}(t),t) - H(\bar{x}(t),\bar{u}(t),\hat{\lambda}(t),t) \tag{2.53}$$

$$-\dot{\hat{\lambda}}(t) = H_x(\bar{x}(t),u(t),\hat{\lambda}(t),t) \tag{2.54}$$

with the usual boundary conditions

$$a(t_f) = 0 \tag{2.55}$$

$$\hat{\lambda}(t_f) = F_x(\bar{x}(t_f),t_f) \tag{2.56}$$

are the estimates for $a(t)$, $\lambda(t)$ defined by (2.34), (2.35) such that

$$\|a(t)-\hat{a}(t)\| \leq c_1\varepsilon^2 \tag{2.57}$$

$$\|\lambda(t)-\hat{\lambda}(t)\| \leq c_2\varepsilon \qquad c_1, c_2 < \infty \tag{2.58}$$

Proof

$$-\frac{d}{dt}(a(t)-\hat{a}(t)) = H(\bar{x},u,\lambda,t)-H(\bar{x},\bar{u},\lambda,t)-H(\bar{x},u,\hat{\lambda},t)+H(\bar{x},\bar{u},\hat{\lambda},t)$$

$$= L(\bar{x},u,t)+\lambda^T f(\bar{x},u,t)-L(\bar{x},\bar{u},t)-\lambda^T f(\bar{x},\bar{u},t)$$
$$-L(\bar{x},u,t)-\hat{\lambda}^T f(\bar{x},u,t)+L(\bar{x},\bar{u},t)+\hat{\lambda} f(\bar{x},\bar{u},t)$$
$$= [\lambda(t)-\hat{\lambda}(t)]^T \Delta f(t).$$

$$-\frac{d}{dt}(\lambda(t)-\hat{\lambda}(t)) = H_x(\bar{x},u,\lambda,t)-H(\bar{x},u,\hat{\lambda},t)+P(t)\Delta f(t)$$
$$= f_x^T(\bar{x}(t),u(t),t)[\lambda(t)-\hat{\lambda}(t)]+P(t)\Delta f(t)$$

(i) $\|\Delta f(t)\| \leq c_3 \|u(t)-\bar{u}(t)\|$ $\qquad c_3 < \infty$

(ii) $\|\Delta f(t)\| \leq c_4$, $c_4 < \infty$ $\quad t \in E$, $\quad \Delta f(t) = 0$, $\quad t \notin E$

Applying Gronwalls inequality gives (2.58). Using (2.58) and applying Gronwalls inequality again gives (2.57) □

Proposition 2.5 is not useful computationally since the third order partial derivative of $V(\bar{x},t)$ w.r.t. x is not known. In the following theorem we estimate the error arising from the omission of the unknown quantities $V^k_{xxx}(\bar{x},t)\Delta f(t)$.

Define

$$u^*(t) \triangleq k(\bar{x}(t),t). \tag{2.59}$$

PROPOSITION 2.7: Let H1A, H2A be satisfied, s = 3. If either

(i) $f_u(x,u,t)$ exists and is continuous in (x,u,t) for all $(x,u,t) \in S$, and $d(u^*,\bar{u}) \leq \varepsilon$, or

(ii) $d_1(u^*,\bar{u}) \leq \varepsilon$

then $\hat{a}(t)$, $\hat{\lambda}(t)$, $\hat{P}(t)$, where:

$$-\dot{\hat{a}}(t) = H(\bar{x}(t),u^*(t),\hat{\lambda}(t),t) - H(\bar{x}(t),\bar{u}(t),\hat{\lambda}(t),t) \tag{2.60}$$

$$-\dot{\hat{\lambda}}(t) = H_x(\bar{x}(t),u^*(t),\hat{\lambda}(t),t) + K^T(t)H_u(\bar{x}(t),u^*(t),\hat{\lambda}(t),t) + \hat{P}(t)\Delta f(t) \tag{2.61}$$

$$\begin{aligned}-\dot{\hat{P}}(t) = {} & H_{xx}(\bar{x}(t),u^*(t),\hat{\lambda}(t),t) + f_x^T(\bar{x}(t),u^*(t),t)\hat{P}(t) \\ & + \hat{P}(t)f_x(\bar{x}(t),u^*(t),t) \\ & + K^T(t)[H_{ux}(\bar{x}(t),u^*(t),\hat{\lambda}(t),t) + f_u^T(\bar{x}(t),u^*(t),t)\hat{P}(t)] \\ & + [H_{ux}(\bar{x}(t),u^*(t),\hat{\lambda}(t),t) + f_u^T(\bar{x}(t),u^*(t),t)\hat{P}(t)]^T K(t) \\ & + K^T(t)H_{uu}(\bar{x}(t),u^*(t),\hat{\lambda}(t),t)K(t) \\ & + \sum_{i=1}^{n} H_u^i(\bar{x}(t),u^*(t),\hat{\lambda}(t),t)\gamma_i(t)\end{aligned} \tag{2.62}$$

with the usual boundary conditions

$$\hat{a}(t_f) = 0 \tag{2.63}$$

$$\hat{\lambda}(t_f) = F_x(\bar{x}(t_f),t_f) \tag{2.64}$$

$$\hat{P}(t_f) = F_{xx}(\bar{x}(t_f),t_f) \tag{2.65}$$

are the estimates for $a(t),\lambda(t),P(t)$ given by equations (2.45) - (2.47) *such that*

$$||a(t) - \hat{a}(t)|| \leq c_1\varepsilon^3 \tag{2.66}$$

$$||\lambda(t) - \hat{\lambda}(t)|| \leq c_2\varepsilon^2 \tag{2.67}$$

$$||P(t) - \hat{P}(t)|| \leq c_3\varepsilon \qquad c_1,c_2,c_3 < \infty \tag{2.68}$$

<u>Proof</u> $-\frac{d}{dt}(a(t)-\hat{a}(t)) = [\lambda(t)-\hat{\lambda}(t)]^T\Delta f(t)$ (2.69)

$$\begin{aligned}-\frac{d}{dt}(\lambda(t)-\hat{\lambda}(t)) = {} & H_x(\bar{x}(t),u^*(t),\lambda(t),t)-H_x(\bar{x}(t),u^*(t),\hat{\lambda}(t),t) \\ & + K^T(t)[H_u(\bar{x}(t),u^*(t),\lambda(t),t)-H_u(\bar{x}(t),u^*(t),\hat{\lambda}(t),t)] \\ & + [P(t) - \hat{P}(t)]\Delta f(t) \\ = {} & [f_x(\bar{x}(t)),u^*(t),t)+f_u(\bar{x}(t),u^*(t),t)K(t)]^T(\lambda(t)-\hat{\lambda}(t)) \\ & + [P(t) - \hat{P}(t)]\Delta f(t)\end{aligned} \tag{2.70}$$

$$
\begin{aligned}
-\frac{d}{dt}(P(t)-\hat{P}(t)) &= H_{xx}(\bar{x}(t),u^*(t),\lambda(t),t)-H_{xx}(\bar{x}(t),u^*(t),\hat{\lambda}(t),t) \\
&+ f_x^T(\bar{x}(t),u^*(t),t)(P(t)-\hat{P}(t)) \\
&+ (P(t)-\hat{P}(t))f_x(\bar{x}(t),u^*(t),t) \\
&+ K^T(t)[H_{ux}(\bar{x}(t),u^*(t),\lambda(t),t)-H_{ux}(\bar{x}(t),u^*(t),\hat{\lambda}(t),t \\
&\qquad + f_u^T(\bar{x}(t),u^*(t),t)(P(t)-\hat{P}(t))] \\
&+ [H_{ux}(\bar{x}(t),u^*(t),\lambda(t),t)-H_{ux}(\bar{x}(t),u^*(t),\hat{\lambda}(t),t \\
&\qquad + f_u^T(\bar{x}(t),u^*(t),t)(P(t)-\hat{P}(t))]^T K(t) \\
&+ K^T(t)[H_{uu}(\bar{x}(t),u^*(t),\lambda(t),t)-H_{uu}(\bar{x}(t),u^*(t),\hat{\lambda}(t),t)]K(t) \\
&+ \sum_{i=1}^{n}[H_u^i(\bar{x}(t),u^*(t),\lambda(t),t) - H_u^i(\bar{x}(t),u^*(t),\hat{\lambda}(t),t)]\gamma_i(t) \\
&+ \sum_{i=1}^{n} V_{xxx_i}^k(\bar{x}(t),t)\Delta f_i(t) \\
&= \sum_{i=1}^{n}(\lambda_i-\hat{\lambda}_i)(f_{xx_i}^T(\bar{x},u^*,t)) \\
&+[f_x(\bar{x},u^*,t)+f_u(\bar{x},u^*,t)K(t)]^T(P(t)-\hat{P}(t)) \\
&\quad (P(t)-\hat{P}(t))[f_x(\bar{x},u^*,t)+f_u(\bar{x},u^*,t)K(t)] \\
&+ \sum_{i=1}^{n}(\lambda_i-\hat{\lambda}_i)[K^T(t)f_{ux_i}^T(\bar{x},u^*,t)+f_{ux_i}(\bar{x},u^*,t)K(t) \\
&\qquad + f_u(\bar{x},u^*,t)\gamma_i(t)] \\
&+ \sum_{i=1}^{n} V_{xxx_i}^k(\bar{x},t)\Delta f_i(t) \\
&= \sum_{i=1}^{n}(\lambda_i-\hat{\lambda}_i)G_i(t)+\sum_{i=1}^{n} V_{xxx_i}^k(\bar{x}(t),t)\Delta f_i(t) \\
&+ [f_x(\bar{x},u^*,t)+f_u(\bar{x},u^*,t)K(t)]^T(P(t)-\hat{P}(t)) \\
&+ (P(t)-\hat{P}(t))[f_x(\bar{x},u^*,t)+f_u(\bar{x},u^*,t)K(t)] \qquad (2.71)
\end{aligned}
$$

where, because of our assumptions, the $G_i(t)$, i=1, .. n are piecewise continuous functions.

Equations (2.69) - (2.71) have terminal conditions of zero at time t_f.

Now, from the assumptions on f, for all $t \in T$

$$\|\Delta f(t)\| \leq L_1 \|u^*(t) - \bar{u}(t)\| \qquad L_1 < \infty$$

From the continuity of $\partial^3 V(x(t),t)/\partial x_i \partial x_j \partial x_k$ w.r.t. $(x(t),t)$ (See Proposition 2.1), and the continuity of $\bar{x}(t)$ w.r.t. t, $\partial V_{xx}(\bar{x}(t),t)/\partial x_i$ is continuous w.r.t. $t \in T$ and hence bounded. Therefore

$$\| \sum_{i=1}^{n} (\partial V_{xx}(\bar{x}(t),t)/\partial x_i) \Delta f_i(t) \| \leq L_2 \|u^*(t) - \bar{u}(t)\| , \; L_2 < \infty$$

Define z(t) to be the n^2+n vector with components $\lambda_i(t)$, $P_{ij}(t)$, $i,j = 1, .. n$ and $\hat{z}(t)$ the corresponding vector constituted from $\hat{\lambda}(t)$, $\hat{P}(t)$.

Then, for all $t \in \theta(u^*) \cap \theta(u)$

$$-\frac{d}{dt}(z(t)-\hat{z}(t)) = H(t)[z(t)-\hat{z}(t)] + \sum_{k=1}^{n} \Delta f_k(t) h_k(t)$$

where the matrix H(t) is piecewise continuous and the vector $h_k(t)$, $k = 1, .. n$ is continuous in T. Hence z(t) is continuous and piecewise differentiable in T. Applying Gronwalls inequality to equation (2.71) where

(i) $\|\Delta f(t)\| \leq L_1 \|u^*(t) - \bar{u}(t)\| \qquad L_1 < \infty$

(ii) $\|\Delta f(t)\| \leq L_3, \; t \in E, \; L_3 < \infty; \quad \Delta f(t) = 0, \; t \notin E.$

we obtain

$$\|z(t)-\hat{z}(t)\| \leq L_4 \varepsilon, \qquad L_4 < \infty.$$

Hence $\|P(t)-\hat{P}(t)\| \leq c_3 \varepsilon \qquad c_3 < \infty.$

Using the above result in equ (2.70) and applying Gronwalls inequality yields equ (2.67). Then, using (2.67) in (2.69) and applying Gronwalls inequality again, yields (2.66)

□

VI EXACT EXPRESSIONS FOR ΔV Mayne [36].

Because u, $\bar{u}$ are arbitrary members of G, they can be interchanged in the previous section. However, we now have for $a(t)$

$$a(t) \triangleq V^{u}(x(t),t) - V(x(t),t) \tag{2.72}$$

where V denotes $V^{\bar{u}}$ or $V^{\bar{k}}$ according to the context. Clearly, we still have

$$a(t_o) = \Delta V.$$

PROPOSITION 2.8: Let H1, H2 be satisfied, $s = 2$. Let $u, \bar{u} \in G$. Then $a(t)$, $V_x^{\bar{u}}(x(t),t)$ are the solutions of

$$-\dot{a}(t) = H(x(t),u(t),V_x^{\bar{u}}(x(t),t),t) - H(x(t),\bar{u}(t),V_x^{\bar{u}}(x(t),t),t) \tag{2.73}$$

$$-\dot{V}_x^{\bar{u}}(x(t),t) = H_x(x(t),\bar{u}(t),V_x^{\bar{u}}(x(t),t),t) + V_{xx}^{\bar{u}}(x(t),t)\Delta f_1(t) \tag{2.74}$$

with boundary conditions

$$a(t_f) = 0 \tag{2.75}$$

$$V_x^{\bar{u}}(x(t_f),t_f) = F_x(x(t_f),t_f) \tag{2.76}$$

where

$$\Delta f_1(t) \triangleq f(x(t),\bar{u}(t),t) - f(x(t),u(t),t) \tag{2.77}$$

Proof Interchange (x,u) with $(\bar{x},\bar{u})$ in the proof of Proposition 2.4. □

Suppose we use policy $\bar{k}(\bar{x}(t),t)$ to generate $(\bar{x},\bar{u})$

Let $$\bar{k}(\bar{x}(t),t) = \bar{k}(x+\delta x,t)$$

$$= \bar{k}(x,t) + \bar{k}_x(x,t)\delta x + \tfrac{1}{2}\delta x^T \bar{k}_{xx}(x,t)\delta x \tag{2.78}$$

to second order in δx, assuming δx is small enough. Define

$$\bar{k}(t) \triangleq \bar{k}_x(x(t),t) \tag{2.79}$$

$$\bar{\gamma}(t) \triangleq \bar{k}_{xx}(x(t),t) \tag{2.80}$$

PROPOSITION 2.9: Let H1A, H2A be satisfied, $s = 3$. Then $a(t)$ satisfies the following equation

$$-\dot{a}(t) = H(x(t),u(t),V_x^{\bar{k}}(x(t),t),t) - H(x(t),\bar{k}(x(t),t),V_x^{\bar{k}}(x(t),t),t) \quad (2.81)$$

$$a(t_f) = 0 \quad (2.82)$$

and $V_x^{\bar{k}}(x,t)$, $V_{xx}^{\bar{k}}(x,t)$ satisfy equations (2.46) and (2.47) with $(x,\bar{k}(x,t),\bar{K},\bar{\gamma})$ replacing $(\bar{x},k(\bar{x},t),K,\gamma)$.

Proof See the proof of Proposition 2.5. □

Let H1, H2 be satisfied $s = 1$. Then we have

(i) $$\Delta V = \int_{t_o}^{t_f} [H(\bar{x}(t),u(t),\lambda(t),t) - H(\bar{x}(t),\bar{u}(t),\lambda(t),t)]dt \quad (2.83)$$

from Proposition 2.4.

(ii) $$\Delta V = \int_{t_o}^{t_f} [H(x(t),u(t),V_x^{\bar{u}}(x(t),t),t) - H(x(t),\bar{u}(t),V_x^{\bar{u}}(x(t),t),t)]dt \quad (2.84)$$

from Proposition 2.8.

Let H1A, H2A be satisfied, $s = 1$. Then we have

(iii) $$\Delta V = \int_{t_o}^{t_f} [H(\bar{x}(t),k(\bar{x}(t),t),\lambda(t),t) - H(\bar{x}(t),\bar{u}(t),\lambda(t),t)]dt \quad (2.85)$$

from Proposition 2.5.

(iv) $$\Delta V = \int_{t_o}^{t_f} [H(x(t),u(t),V_x^{\bar{k}}(x(t),t),t) - H(x(t),\bar{k}(x(t),t),V_x^{\bar{k}}(x(t),t),t)]dt \quad (2.86)$$

from Proposition 2.9.

These expressions for ΔV are the same as those given by Mayne [36]. They are used in further chapters to establish conditions of optimality.

CHAPTER 3

SUFFICIENT CONDITIONS OF OPTIMALITY

Our approach contributes nothing to the discussion of obtaining necessary conditions. The major difficulty is proving [Halkin] the existence of a separating hyperplane between the reachable set of the linearized system (equs (1.15), (1,16)) and the set which is the intersection of the set of decreased costs and the set of permissible final states.

However, the expressions we have obtained for ΔV are far more useful for proving the sufficiency of the minimum principle, i.e. that if there does not exist $u_1 \in b$, $t_1 \in 0(\bar{u})$ such that $H(\bar{x}(t_1),u_1,\bar{\lambda}(t_1),t_1) < (H(\bar{x}(t_1),u(t_1),\bar{\lambda}(t_1),t_1)$ then $\bar{u}$ is optimal. Satisfaction of the minimum principle is, of course, sufficient for optimality only for problems of special structure.

I GLOBAL SUFFICIENCY RESULTS

We give the results from Mayne [36].

1. f, L, F linear in x

Consider the system defined by:

$$f(x,u,t) = A(t)x + \phi(u,t)$$
$$L(x,u,t) = m^T(t)x + 0(u,t)$$
$$F(x) = n^T x$$

where A, ϕ, m, θ are continuous.

$$H_x(x,u,\lambda,t) = m(t) + A^T(t)\lambda(t)$$

which is independant of x. Therefore, $\forall t \in T$, $\forall x \in R^n$, we have

$$V_x^{\bar{u}}(x(t),t) = V_x^{\bar{u}}(\bar{x}(t),t) \triangleq \bar{\lambda}(t)$$

If we now use equation (2.84), with $\bar{\lambda}(t)$ replacing $V_x^{\bar{u}}(x(t),t)$, we see that the satisfaction of the minimum principle by $(\bar{x},\bar{u})$ implies

$$\Delta V \geq 0 \quad \text{for all} \quad u \in G.$$

2. f linear in x; L, F convex in x

Let f be defined as above, and L, F by

$$L(x,u,t) = \eta(x,t) + \theta(u,t)$$

$$F(x) = \xi(x)$$

where η, ξ are convex in x and L, F satisfy H1, H2, s = 2. Then

$$\hat{L}(x,u,t) \triangleq \eta(\bar{x}(t),t) + \eta_x^T(\bar{x}(t),t)(x-\bar{x}(t)) + \theta(u,t)$$

$$L(x,u,t) \geq \hat{L}(x,u,t)$$

$$\hat{F}(x) \triangleq F(\bar{x}(t_f)) + F_x(\bar{x}(t_f))(x-\bar{x}(t))$$

$$F(x) \geq \hat{F}(x).$$

Then, if $\hat{V}$ denotes the cost with L replaced by $\hat{L}$, F by $\hat{F}$, we have

$$V^{\bar{u}}(x_o,t_o) = \hat{V}^{\bar{u}}(x_o,t_o)$$

$$V^{u}(x_o,t_o) \geq \hat{V}^{u}(x_o,t_o) \qquad \forall u \in G$$

Therefore

$$V^{u}(x_o,t_o) - V^{\bar{u}}(x_o,t_o) \geq \hat{V}^{u}(x_o,t_o) - \hat{V}^{\bar{u}}(x_o,t_o) \qquad \forall u \in G.$$

$$\text{i.e.} \quad \Delta V \geq \Delta \hat{V} \qquad \forall\, u \in G.$$

But, from the previous example, satisfaction of the minimum principle by $(\bar{x},\bar{u})$ implies $\Delta\hat{V} \geq 0$, and hence $\Delta V \geq 0$ for all $u \in G$.

3. Terminal Constraints

Consider the problem immediately above, with the added terminal constraints

$$x(t_f) = x_f.$$

Assume $(\bar{x},\bar{u})$ satisfy the minimum principle for the modified problem with terminal cost

$$\hat{F}(x) = F(x) + c^{T}(x-x_f)$$

and no constraint, i.e. with

$$\bar{\lambda}(t_f) = F_x(x(t_f)) + c$$

Since, for all $u \in G$ such that $x(t_f) = x_f$ the total cost for the original and modified problems are the same, satisfaction of the minimum principle implies, as before, that $\Delta\hat{V} \geq 0$ and $\Delta V \geq 0$ for the problem with terminal cost.

Assume now that we have a terminal inequality constraint:

$$a^{T}x(t_f) - b \leq 0.$$

Assume that $(\bar{x},\bar{u})$ where

$$\alpha^T\bar{x}(t_f) = b$$

satisfy the minimum principle for the modified problem with terminal cost

$$\hat{F}(x) = F(x) + c[\alpha^T x - b]$$

for some scalar $c > 0$, and no constraint, i.e. with

$$\bar{\lambda}(t_f) = F_x(\bar{x}(t_f)) + c\alpha.$$

Since, for all $u \in G$ satisfying the terminal constraint,

$$c[\alpha^T x(t_f) - b] = c[\alpha^T(x(t_f) - \bar{x}(t_f))] \leq 0$$

we have

$$\hat{F}(x) \leq F(x).$$

Hence, optimality of $(\bar{x},\bar{u})$ for the original problem follows.

4. f linear; L, F Quadratic

Consider the system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \qquad x(t_o) = 0$$

with

$$L(x,u,t) \triangleq \tfrac{1}{2}x^T Q(t)x + \tfrac{1}{2}u^T u$$

$$F(x) \triangleq \tfrac{1}{2}x^T Q_f x$$

where A, B, Q are piecewise continuous, $Q, Q_f \geq 0$

Consider the policy:

$$\bar{u}(x,t) = K(t)x$$

$$K(t) = -B^T(t)P(t)$$

where P(t) is the solution of the normal Riccati equation,

assumed to exist in (t_o,t_f). Then it is easily shown that $V^{\bar{k}}$ is quadratic in x.

$$V^{\bar{k}}(x,t) = \tfrac{1}{2}x^T P(t)x$$

so that

$$V_x^{\bar{k}}(x,t) = P(t)x$$

$$V_{xx}^{\bar{k}}(x,t) = P(t).$$

In fact P(t) is the solution of equ (2.47), along any $(\bar{x},\bar{u})$, since V_{xxx} is zero.

From equation (2.86) the cost $V^u(x_o,t_o) - V^{\bar{k}}(x_o,t_o)$ for any $u \in G$ is given by

$$\begin{aligned} V &= \int_{t_o}^{t_f} [H(x,u,V_x^{\bar{k}}(x,t),t) - H(x,\bar{k}(x,t),V_x^{\bar{k}}(x,t),t)]dt \\ &= \int_{t_o}^{t_f} [\tfrac{1}{2}x^T Q(t)x + \tfrac{1}{2}u^T u + x^T P(t)\{Ax+Bu\} \\ &\qquad - \tfrac{1}{2}x^T Q(t)x - \tfrac{1}{2}x^T K^T(t)K(t)x \\ &\qquad\qquad - x^T P(t)\{A+BK(t)\}x]dt \\ &= \int_{t_o}^{t_f} [\tfrac{1}{2}u^T u + x^T P(t)B(t)u - x^T P(t)B(t)K(t)x \\ &\qquad\qquad\qquad - \tfrac{1}{2}x^T K^T(t)K(t)x]dt \\ &= \int_{t_o}^{t_f} [\tfrac{1}{2}u^T u - x^T K^T(t)u + \tfrac{1}{2}x^T K^T(t)K(t)x]dt \\ &= \int_{t_o}^{t_f} \tfrac{1}{2}[u(t)-K(t)x(t)]^T[u(t)-K(t)x(t)]dt \geq 0. \end{aligned}$$

Consider now the addition of a terminal constraint

$$x(t_f) = x_f.$$

We deal with this by considering a modified problem with terminal cost

$$\hat{F}(x) \triangleq F(x) + c^T(x-x_f).$$

Consider the policy

$$\bar{k}(x,t) = u^*(t)+K(t)x(t)$$

$$u^*(t) = -B^T(t)\lambda(t)$$

$$K(t) = -B^T(t)P(t)$$

where

$$V_x^{\bar{k}}(x,t) = P(t)x(t) + \lambda(t)$$

$$V_{xx}^{\bar{k}}(x,t) = P(t)$$

Then, we have from equ (2.35), with $\bar{x}(t) = 0$ $\bar{u}(t) = 0$

$$-\dot{\lambda}(t) = H_x(\bar{x}(t),u^*(t),\lambda(t),t)+P(t)(f(\bar{x},u^*,t)-f(\bar{x},\bar{u},t))$$

$$= [A^T(t) + K^T(t)B^T(t)]\lambda(t)$$

with boundary condition

$$\lambda(t_f) = c.$$

and $P(t)$ is the solution of the normal Riccati equation assumed to exist on T. Hence:

$$\dot{x}(t) = A_1(t)x(t) + B(t)u^*(t), \qquad x(t_o) = 0$$

$$u^*(t) = -B^T(t)\lambda(t) \quad \text{and} \quad A_1(t) = A(t) + B(t):$$

$$-\dot{\lambda}(t) = A_1^T(t)\lambda(t), \ \lambda(t_f) = c.$$

If $\phi(t,\tau)$ is the transition matrix associated with $A_1(t)$, then

$$x(t_f) = \int_{t_o}^{t_f} \phi(t_f,\tau)B(\tau)u^*(\tau)d\tau$$

$$= -\int_{t_o}^{t_f} \Phi(t_f,\tau)B(\tau)B^T(\tau)\lambda(\tau)d\tau$$

$$= -\int_{t_o}^{t_f} \Phi(t_f,\tau)B(\tau)B^T(\tau)\Phi^T(t_f,\tau)d\tau \quad c$$

$$= -W(t_f,t_o)c$$

where W is the controllability matrix corresponding to $(A_1(t),B(t))$.

As before, we have from equ (2.86), for any $u \in G$

$$V = \int_{t_o}^{t_f} [\tfrac{1}{2}x^TQx+\tfrac{1}{2}u^Tu+[V_x^{\bar{k}}(x,t)]^T(A_1x+Bu)$$

$$-\tfrac{1}{2}x^TQx-\tfrac{1}{2}\bar{k}^T\bar{k}-[V_x^{\bar{k}}(x,t)]^T(A_1x+B\bar{k})]dt$$

$$= \int_{t_o}^{t_f} [\tfrac{1}{2}u^Tu+(x^TP+\lambda^T)Bu-\tfrac{1}{2}\bar{k}^T\bar{k}-(x^TP+\lambda^T)B\bar{k}]dt$$

$$= \int_{t_o}^{t_f} [\tfrac{1}{2}u^Tu-x^T\bar{k}^Tu+\tfrac{1}{2}\bar{k}^T\bar{k}]dt$$

$$= \int_{t_o}^{t_f} \tfrac{1}{2}[u(t)-\bar{k}(x,t)]^T[u(t)-\bar{k}(x,t)]dt \geq 0$$

Therefore, $\bar{k}$ is optimal for the modified problem (i.e. $\Delta V \geq 0$ for all $u \in G$). Thus $\bar{k}$ is optimal for the original problem with terminal constraint:

$$x(t_f) = -W(t_f,t_o)c = x_f.$$

Obviously, if the system is completely controllable, x_f may be arbitrary. For $c = c_1$, say, the resultant policy

$\bar{k}$ satisfies $V^u(x_o,t_o) \geq V^{\bar{k}}(x_o,t_o)$ for all $x(t_f;x_o,t_o,u) = -W(t_f,t_o)c_1$.

5. The Weierstrass Excess function

For the calculus of variation problem of choosing an optimal curve, the system dynamics are:

$$\dot{x}(t) = u(t)$$

i.e. $H(x,u,\lambda,t) = L(x,u,t) + \lambda^T u$

Let $\bar{k}$ be a policy which satisfies our assumptions, and also

$$H_u(x,\bar{k}(x,t),V_x^{\bar{k}}(x,t),t) = 0 \quad \text{for all } x \in R^n,\ t \in T.$$

Hence

$$V_x^{\bar{k}}(x,t) = -L_u(x,\bar{k}(x,t),t).$$

From equ (2.86)

$$\Delta V = \int_{t_o}^{t_f} E(x(t),u(t),\bar{k}(x(t),t),t)dt$$

where

$$E(x,u,w,t) \triangleq L(x,u,t) - L(x,w,t) - L_u(x,w,t)(u-w)$$

is the Weierstrass Excess function. Clearly the non-negativity of E for $w = \bar{k}(x,t)$ for all $x \in R^n$, $t \in T$ is a sufficient condition for the global optimality of $\bar{k}$.

6. Hamilton - Jacobi - Bellman Result - Bellman and Dreyfus [1].

From (2.84), if there exists a $\bar{u}$, satisfying H1, H2 and

$$H(x,u,V_x^{\bar{u}}(x,t),t) \geq H(x,\bar{u},V_x^{\bar{u}}(x,t),t)$$

for all $(x,u,t) \in S$, then $V^u(x,t) \geq V^{\bar{u}}(x,t)$ for all $u \in G$, $x \in R^n$, $t \in T$.

We have from Proposition 2.2

$$-V_t^{\bar{u}}(x,t) = H(x,\bar{u},V_x^{\bar{u}}(x,t),t) \qquad \forall t \in \theta(\bar{u}).$$

Now suppose that the optimal control for state $x(t)$ is continuous, and denote the optimal cost function for the state $x(t)$, by $V^o(x(t),t)$. Then we have

$$-V_t^o(x(t),t) = \min_{u \in G} H(x(t),u(t),V_x^o(x(t),t),t) \qquad \forall t \in T.$$

- the H-J-B equation.

Using the above equation in place of (2.9) Jacobson obtains optimal versions of the equations given in Propositions 2.4, 2.5 in the sense of a maximum reduction in cost.

It can be shown (see Chapter 8) that if the optimal control is discontinuous, then Proposition 2.1(i) does not necessarily hold for $V_{xx}^o(x(t),t)$.

II A LOCAL SUFFICIENCY THEOREM FOR STRONG VARIATIONS IN CONTROL

LEMMA 3.1: Let Z, W denote closed, bounded subsets of R^n and R^m respectively. Let $\phi : Z \times W \times T \to R$, and its partial derivatives w.r.t. (z,w) be continuous, except at a finite number of points in T, and satisfy, for some $c_1 < \infty$:

(i) $\phi(z,o,t) = 0$ for all $(z,t) \in Z \times T$

(ii) $\|\phi_w(z,o,t)\| \leq c_1 \|z\|^2$ for all $(z,t) \in Z \times T$

(iii) $\phi_{ww}(o,o,t) > 0$ for all $t \in T$

(iv) $\phi(o,w,t) > 0$ for all $(w,t) \in W \times T$ s.t. $w \neq 0$.

Then, there exists an $\epsilon_3 > 0$ and an $\alpha > 0$ s.t.

$$\phi(z,w,t) \geq \tfrac{1}{2}\alpha w^T w + r$$

for all $(z,w,t) \in Z \times W \times T$ s.t. $\|z\| \leq \epsilon_3$ where $|r| \leq c_1 \|z\|^2 \|w\|^2$.

Proof: From (iii) and the piecewise continuity of $\phi_{ww}(o,o,\cdot)$ there exists $\epsilon_1 > 0$, $\alpha_1 > 0$ s.t.

$$\phi_{ww}(z,w,t) - \alpha_1 I \geq 0 \tag{3.1}$$

for all $(z,w,t) \in Z \times W \times T$ s.t. $\|z\| \leq \epsilon_1$, $\|w\| \leq \epsilon_1$. Hence, for all $\|w\| \leq \epsilon_1$, we have from (3.1) and (i)

$$\phi(z,w,t) \geq \tfrac{1}{2}\alpha_1 w^T w + \phi_w^T(z,o,t)w.$$

From (iv) and the continuity of $\phi(\cdot)$, there exists an $\varepsilon_2 > 0$, $\alpha_2 > 0$ such that

$$\phi(z,w,t) \geq \alpha_2$$

for all $(z,w,t) \in Z \times W \times T$ s.t. $\|z\| \leq \varepsilon_2$, $\|w\| > \varepsilon_1$.

Let d be defined as

$$d \triangleq \max_{w \in W} w^T w$$

The lemma follows from (ii) $\{\|\phi_w^T(z,o,t)w\| \leq c_1\|z\|^2\|w\|\}$ with $\varepsilon_1 = \min[\varepsilon_1, \varepsilon_2]$, $\alpha = \min[\alpha_1, \alpha_2/d]$.

□

LEMMA 3.2: Let H1A, H2A be satisfied, s = 3. Let $u, \bar{u} \in G$. Define $\phi(\cdot)$ by

$$\phi(\delta x,w,t) \triangleq H(\bar{x}(t)+\delta x, w+\bar{u}(t)+\bar{K}(t)\delta x, V_x^{\bar{K}}(\bar{x}(t)+\delta x,t),t)$$

$$-H(\bar{x}(t)+\delta x,\bar{u}(t)+\bar{K}(t)\delta x, V_x^{\bar{K}}(\bar{x}(t)+\delta x,t),t)$$

where

$$w \triangleq u(t)-\bar{u}(t)-\bar{K}(t)\delta x.$$

Then, if there exists an $\varepsilon > 0$ s.t. $d(u,\bar{u}) \leq \varepsilon$,

$$\phi(\delta x(t),w(t),t) \geq \tfrac{1}{2}\alpha w^T(t)w(t)+r(t) \tag{3.2}$$

where

$$|r(t)| \leq c[d(u,\bar{u})]^3. \tag{3.3}$$

Also, there exists a $\bar{c} < \infty$ s.t.

$$\int_{t_o}^{t_f} w^T(t)w(t)dt \geq \bar{c}[d(u,\bar{u})]^2.$$

Proof: Consider the system

$$\dot{x}(t) = g(x(t),w(t),t), \qquad x(t_o) = x_o$$

where

$$g(x,w,t) \triangleq f(x,w+\bar{u}(t)+K(t)(x-\bar{x}(t)),t)$$

If $w(t) \equiv 0$, the corresponding trajectory is $\bar{x}(t)$; if $w(t) \not\equiv 0$, the corresponding trajectory is $x(t)$.

Since $g(\cdot)$ satisfies the hypothesis of Proposition 1.4(ii),

$$\|\delta x(t)\| \leq c_2 d(w,o) \qquad c_2 < \infty$$

where $\delta x(t) \triangleq x(t) - \bar{x}(t)$.
Hence, for some $c_3 < \infty$

$$d(u,\bar{u}) \leq c_3 d(w,o). \tag{3.4}$$

Since $w(t) \equiv u(t)-\bar{u}(t)-\bar{K}(t)\delta x(t)$ and for some $c_4 < \infty$, $\|\delta x(t)\| \leq c_4 d(u,\bar{u})$ for all $t \in T$, it follows that for some $c_5 < \infty$, for all $u \in G$

$$d(w,o) \leq c_5 d(u,\bar{u}).$$

i.e. $$d(w,o) < \min[d(u,\bar{u})/c_3, c_5 d(u,\bar{u})].$$

Then, if $d(u,\bar{u}) < \varepsilon$, there exists $\varepsilon_2 > 0$ s.t.

$$\|\delta x(t)\| < \varepsilon_2 \qquad \forall t \in T.$$

Now it can be easily shown that $\phi(\cdot)$ as defined above satisfies the hypothesis of Lemma 3.1. Hence, if $d(u,\bar{u}) < \varepsilon$ (so that $\|\delta x(t)\| < \varepsilon_2 \ \forall t \in T$) we have that $\phi(\delta x(t),w(t),t)$ satisfies equ (3.2) with $|r(t)| \leq c_1\|\delta x(t)\|^2\|w(t)\|$. So there exists a $c < \infty$ s.t. for

all $t \in T$, $u \in G$ s.t. $d(u,\bar{u}) \leq \varepsilon$, $|r(t)|$ satisfies equ (3.3)

Also

$$\int_{t_o}^{t_f} w^T(t)w(t)dt = \int_{t_o}^{t_f} \|w(t)\|^2 dt$$

$$\geq c_o[d(w,o)]^2 \quad \text{from (1.13)}$$

$$\geq \bar{c}[d(u,\bar{u})]^2 \quad \text{from (3.4)}$$

for some $\bar{c} \in (o,\infty)$.

□

PROPOSITION 3.1 Let H1A, H2A be satisfied $s = 3$, $\bar{u}(t) \in$ interior of Ω for all $t \in T$. If for all $t \in T$:

(i) $H_u(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) = 0$

(ii) $H_{uu}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) > 0$

(iii) $H(\bar{x}(t),u,\bar{\lambda}(t),t) > H(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)$ for all $u \in \Omega$ s.t. $u \neq \bar{u}(t)$

(iv) The matrix Riccati differential equation

$$-\dot{\bar{P}}(t) = H_{xx}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)+f_x^T(\bar{x}(t),\bar{u}(t),t)\bar{P}(t)$$
$$+ \bar{P}(t)f_x(\bar{x}(t),\bar{u}(t),t)$$
$$- \bar{K}^T(t)H_{uu}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)\bar{K}(t)$$

with boundary condition

$$\bar{P}(t_f) = F_{xx}(\bar{x}(t_f),t_f)$$

is bounded (has no conjugate points) in T,

then, $\bar{u}$ is locally optimal in the sense that $V^{u}(x_o,t_o) > V^{\bar{u}}(x_o,t_o)$ for all $u \in G$ s.t. $u \neq \bar{u}$ and $d(u,\bar{u}) \leq \varepsilon$ for some $\varepsilon > 0$.

$$\bar{K}(t) \triangleq -\bar{H}_{uu}^{-1}(\bar{x},\bar{u},\bar{\lambda},t)[H_{ux}(\bar{x},\bar{u},\bar{\lambda},t)+f_{u}^{m}(\bar{x},\bar{u},t)\bar{P}(t)] \tag{3.5}$$

and $\bar{\lambda}(t)$ is the solution of Equs (2.16) and (2.17).

Proof: Consider the policy defined by

$$\bar{k}(x,t) = \bar{u}(t) + \bar{K}(t)(x(t)-\bar{x}(t)).$$

From (iv), $\bar{P}(t)$ exists for all $t \in T$, so that $\bar{k}$ is well defined and satisfies our assumptions. From Proposition 2.5 we have, for $u^{*} = \bar{u}$ and $H_{u}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) = 0$

$$V_{x}^{\bar{k}}(\bar{x}(t),t) = \bar{\lambda}(t)$$

$$V_{xx}^{\bar{k}}(\bar{x}(t),t) = \bar{P}(t).$$

Also, the control u that minimizes the second order expansion w.r.t. $(\bar{x},\bar{u})$ of $H(x,u,V_{x}^{\bar{k}}(x,t),t)$ about $(\bar{x},\bar{u})$ is $\bar{u}(t) + \bar{K}(t)[x-\bar{x}(t)]$ with $\bar{K}(t)$ defined in equ (3.5). From equ (2.86) we have

$$\begin{aligned}\delta V &= \int_{t_o}^{t_f}[H(\bar{x}+\delta x,u,V_{x}^{\bar{k}}(\bar{x}+\delta x,t),t) \\ &\quad - H(\bar{x}+\delta x,\bar{u}+\bar{K}(t)\delta x,V_{x}^{\bar{k}}(\bar{x}+\delta x,t),t)]dt \\ &\geq \tfrac{1}{2}\alpha\int_{t_o}^{t_f} w^{T}(t)w(t)dt + \int_{t_o}^{t_f} r(t)dt \quad \text{from Lemma 3.2}\end{aligned}$$

where,

$$|r(t)| \leq c[d(u,\bar{u})]^3 \qquad \forall t \in T.$$

so that

$$\Delta V \geq \tfrac{1}{4}\alpha\bar{c}[d(u,\bar{u})]^2 + \bar{r} \qquad \text{from Lemma 3.2}$$

where

$$|\bar{r}| \leq c(t_f - t_o)[d(u,\bar{u})]^3$$

Hence, there exists an $\bar{\varepsilon} \in (o,\varepsilon)$ s.t. $d(u,\bar{u}) \leq \bar{\varepsilon}$ implies

$$\Delta V \geq \tfrac{1}{4}\alpha\bar{c}[d(u,\bar{u})]^2.$$

□

COROLLARY 3.2 There exists $\varepsilon > 0$ s.t. $V^u(x_o+\delta x,t_o) > V^{\bar{u}}(x_o+\delta x,t_o)$ for all $u \in G$ s.t. $u \neq \bar{u}$, $d(u,\bar{u}) \leq \varepsilon$ for all δx s.t. $\|\delta x\| \leq \varepsilon$.

CHAPTER 4

FURTHER ESTIMATES OF ΔV Mayne [36].

We first derive a first order estimate of ΔV which will be used later in Chapter 5-I-3.

PROPOSITION 4.1: Let $u, \bar{u} \in G$. If either

(i) H1A, H2A are satisfied, $s = 2$ and $d(u,\bar{u}) \leq \varepsilon$, or

(ii) H1, H2 are satisfied, $s = 2$ and $d_1(u,\bar{u}) \leq \varepsilon$, then:

$$\Delta V = \int_{t_o}^{t_f} [H(\bar{x}(t),u(t),\bar{\lambda}(t),t) - H(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)]dt + e_1 \qquad (4.1)$$

where $|e_1| \leq c\varepsilon^2$, $\quad c < \infty$

and $\bar{\lambda}(t)$, is the solution of equs (2.16) (2.17).

Proof: From (2.84):

$$\Delta V = \int_{t_o}^{t_f} [H(\bar{x}+\delta x,u,V_x^{\bar{u}}(\bar{x}+\delta x,t),t) - H(\bar{x}+\delta x,\bar{u},V_x^{\bar{u}}(\bar{x}+\delta x,t),t)]dt$$

$$\simeq \int_{t_o}^{t_f} [H(\bar{x},u,\bar{\lambda}(t),t) - H(\bar{x},\bar{u},\bar{\lambda}(t),t)]dt$$

$$+ \int_{t_o}^{t_f} \{[H_x(x^*,u,V_x^{\bar{u}}(x^*(t),t),t) - H_x(x^*,\bar{u},V_x^{\bar{u}}(x^*(t),t),t)$$

$$+ V_{xx}^{\bar{u}}(x^*(t),t)(f(x,u,t) - f(x,\bar{u},t))]\delta x\}dt$$

where $x^*(t) = \bar{x}(t) + \theta\delta x(t), \quad 0 \leq \theta \leq 1.$

We have from Proposition 1.4 that $\|\delta x(t)\| \leq c_1\varepsilon$, $c_1 < \infty$, for all $t \in T$. The terms inside the second integrand are of the form $\theta(u(t))-\theta(\bar{u}(t))$. Hence the result follows.

□

We now give a second order estimate of ΔV. The corollary following the result will lead to a necessary condition of optimality for singular control problems (Chapter 10), Mayne [36]. A further development of the corollary leads to a strong version of the usual second variation formula.

PROPOSITION 4.2: Let $u,\bar{u} \in G$. If either

(i) H1A,H2A are satisfied, $s = 3$ and $d(u,\bar{u}) \leq \varepsilon$, or

(ii) H1, H2 are satisfied, $s = 3$ and $d(u,\bar{u}) \leq \varepsilon$, then

$$\Delta V = \int_{t_o}^{t_f} [\Delta H(t)+\{\Delta H_x(t)+\bar{P}(t)\Delta f(t)\}^T\delta x(t)]dt + e_1 \qquad (4.2)$$

where $|e_1| \leq c\varepsilon^3$, $c < \infty$

$$\delta x(t) \triangleq x(t)-\bar{x}(t)$$

$$\Delta H(t) \triangleq H(\bar{x},u,\bar{\lambda},t)-H(\bar{x},\bar{u},\bar{\lambda},t)$$

$$\Delta H_x(t) \triangleq H_x(\bar{x},u,\bar{\lambda},t)-H_x(\bar{x},\bar{u},\bar{\lambda},t)$$

$$\Delta f(t) = f(\bar{x},u,t)-f(\bar{x},\bar{u},t) \qquad (4.3)$$

and $\bar{\lambda}(t)$, $\bar{P}(t)$ are the solutions of (2.16) - (2.19).

Proof: Expanding the integrand of equ (2.84) about $\bar{x}(t)$ s.r.t. x up to second order and neglecting the second order terms yields the integrand in equ (4.2). The neglected terms are:

$$\tfrac{1}{2}\delta x^T(t)\big[H_{xx}(x^*(t),u(t),V_x^{\bar{u}}(x^*(t),t),t) - H_{xx}(x^*,\bar{u}(t),V_x^{\bar{u}}(x^*(t),t),t)$$

$$+ \{f_x(x^*(t),u(t),t) - f_x(x^*(t),\bar{u}(t),t)\}^T V_{xx}^{\bar{u}}(x^*(t),t)$$

$$+ V_{xx}^{\bar{u}}(x^*(t),t)\{f_x(x^*(t),u(t),t) - f_x(x^*(t),\bar{u}(t),t)\}$$

$$+ \sum_{i=1}^{n} V_{xxx_i}^{\bar{u}}(x^*(t),t)(f(x,u,t) - f(x,\bar{u},t))\big]\delta x(t)$$

where $x^*(t) = \bar{x}(t) + \theta\delta x(t)$; $0 \le \theta \le 1$.

Since, from Proposition (1.4) $\|\delta x(t)\| \le c_1\varepsilon$, $c_1 < \infty$ for all $t \in T$, and the terms inside the bracket are all of the form $\phi(u(t)) - \phi(\bar{u}(t))$, the result follows. □

COROLLARY 4.3: Proposition 4.2 holds if $\delta x(t)$ is replaced by the approximation $\delta\hat{x}(t)$ given by equs (1.15), (1.16).

Proof: If $\delta x(t)$ is replaced by $\delta\hat{x}(t)$ in the integrand of Proposition 4.2, then a further error of the form

$$\int_{t_o}^{t_f} [\theta(u(t)) - \theta(\bar{u}(t))]^T [\delta x(t) - \delta\hat{x}(t)]\,dt$$

is introduced. Since, from Proposition 1.6, $\| \delta x(t) - \delta\hat{x}(t) \| \leq c_2\varepsilon^2$, $c_2 < \infty$, the error is not greater than $c_3\varepsilon^3$.

□

PROPOSITION 4.4: Let the hypothesis of Proposition 4.2 be satisfied. Then

$$\Delta V = \int_{t_o}^{t_f} [\Delta H(t) + \Delta H_x^T(t)\delta\hat{x}(t) + \tfrac{1}{2}\delta x^T(t)Q(t)\delta\hat{x}(t)]dt$$

$$+ \tfrac{1}{2}\delta\hat{x}(t_f)F_{xx}(x(t_f),t_f)\delta\hat{x}(t_f) + e_1$$

where

$|e_1| \leq c\varepsilon^3, \quad c < \infty$

$Q(t) \triangleq H_{xx}(\bar{x},\bar{u},\bar{\lambda},t)$

and the remaining terms are as defined in Proposition 4.2.

Proof: We have from equ (1.15) that

$$\delta\dot{\hat{x}}(t) = f_x(\bar{x}(t),\bar{u}(t),t)\delta\hat{x}(t) + \Delta f(t)$$

where $\Delta f(t)$ is given in equ (4.3).

Therefore, replacing $\Delta f(t)$ by $\delta\dot{\hat{x}}(t) - f_x(\bar{x}(t),\bar{u}(t),t)\delta\hat{x}(t)$ in equ (4.2), we have, from corollary 4.3 that

$$\Delta V = \int_{t_o}^{t_f} [\Delta H(t) + \Delta H_x(t)\delta\hat{x}(t) + (\delta\dot{\hat{x}}(t) - f_x(\bar{x}(t),\bar{u}(t),t)\delta\hat{x}(t))^T\bar{p}(t)\delta\hat{x}(t)]dt + e_1$$

where $|e_1| \leq c_1\varepsilon^3, \quad c_1 < \infty$

$$= \int_{t_o}^{t_f} [\Delta H(t) + \Delta H_x(t)\delta\underline{x}(t) - \delta\underline{x}^T(t) f_x^T(\bar{\underline{x}}(t), \bar{\underline{u}}(t), t)\bar{P}(t)\delta\underline{x}(t)$$

$$+ \tfrac{1}{2}\delta\dot{\underline{x}}^T(t)\bar{P}(t)\delta\underline{x}(t) + \tfrac{1}{2}\delta\underline{x}^T(t)\bar{P}(t)\delta\dot{\underline{x}}(t)]dt + e_1$$

$$= \int_{t_o}^{t_f} [\Delta H(t) + \Delta H_x(t)\delta\underline{x}(t) - \delta\underline{x}^T(t) f_x^T(\bar{\underline{x}}(t), \bar{\underline{u}}(t), t)\bar{P}(t)\delta\underline{x}(t)$$

$$- \tfrac{1}{2}\delta\underline{x}^T(t)\dot{\bar{P}}(t)\delta\underline{x}(t) - \tfrac{1}{2}\delta\underline{x}^T(t)\bar{P}(t)\delta\dot{\underline{x}}(t) + \tfrac{1}{2}\delta\underline{x}(t)\bar{P}(t)\delta\underline{x}(t)]dt$$

$$+ \tfrac{1}{2}\delta\underline{x}^T(t_f) F_{xx}(\bar{\underline{x}}(t_f), t_f)\delta\underline{x}(t_f) + e_1$$

$$= \int_{t_o}^{t_f} [\Delta H(t) + \Delta H_x(t)\delta\underline{x}(t) + \tfrac{1}{2}\delta\underline{x}(t)\,(-f_x^T(\bar{\underline{x}}(t), \bar{\underline{u}}(t), t)\bar{P}(t)$$

$$- \bar{P}(t) f_x^T(\bar{\underline{x}}(t), \bar{\underline{u}}(t), t) - \dot{\bar{P}}(t))\delta\underline{x}(t)]dt$$

$$+ \tfrac{1}{2}\delta\underline{x}^T(t_f) F_{xx}(\bar{\underline{x}}(t_f), t_f)\delta\underline{x}(t_f) + e_1$$

$$= \int_{t_o}^{t_f} [\Delta H(t) + \Delta H_x(t)\delta\underline{x}(t) + \tfrac{1}{2}\delta\underline{x}(t) Q(t)\delta\underline{x}(t)]dt$$

$$+ \tfrac{1}{2}\delta\underline{x}(t_f) F_{xx}(\bar{\underline{x}}(t_f), t_f)\delta\underline{x}(t_f) + e_1 \quad \text{from equ (2.18).}$$

Hence result □

The weak version of the last result is the usual first and second variation result.

PROPOSITION 4.5: Let $u, \bar{u} \in G$. If either:

(i) H1, H2 are satisfied $s = 3$, f and L are linear in u and $d(u,\bar{u}) \leq \varepsilon$ or $d_1(u,\bar{u}) \leq \varepsilon$, or

(ii) H1A, H2A are satisfied $s = 3$ and $\|u(t)-\bar{u}(t)\| \leq \varepsilon$ for all $t \in T$. Then

$$\Delta V = \int_{t_o}^{t_f} H_u^T(\bar{x}(t),\bar{u}(t),\lambda(t),t)\delta u(t)\,dt$$

$$+ \int_{t_o}^{t_f} [\tfrac{1}{2}\delta u^T(t)R(t)\delta u(t) + \delta u^T(t)C(t)\delta\hat{x}(t) + \tfrac{1}{2}\delta\hat{x}(t)Q(t)\delta\hat{x}(t)]dt$$

$$+ \tfrac{1}{2}\delta\hat{x}(t_f)F_{xx}(\bar{x}(t_f),t_f)\delta\hat{x}(t_f) + e_1$$

where $|e_1| \leq c_1\varepsilon^3$, $c_1 < \infty$ and

$$\delta u(t) \triangleq u(t) - \bar{u}(t)$$

$$R(t) \triangleq H_{uu}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)$$

$$C(t) \triangleq H_{ux}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t).$$

$\delta\hat{x}(t)$ is the solution of equs (1.15), (1.16) and the remaining terms are as defined in Proposition (4.2) and (4.3).

Proof: Expand the integrand of Proposition 4.4 w.r.t u up to terms of second order. The resultant error is not greater than $c_2\varepsilon^3$, $c_2 < \infty$.

□

So far we have been concerned with comparing V^u and $V^{\bar{u}}$. In the sequel, we will compare V^u with $V^{\bar{k}}$, where $\bar{k}$ is the local linear policy defined by

$$\bar{k}(x,t) = \bar{u}(t) + \bar{K}(t)[x-\bar{x}(t)]$$

where $\bar{u} \in G$, and $\bar{K}$ is piecewise continuous. Clearly $\bar{k}$ generates $(\bar{x},\bar{u})$ so that $V^{\bar{k}}(x_o,t_o) = V^{\bar{u}}(x_o,t_o)$.

PROPOSITION 4.6: Let H1A, H2A be satisfied, $s = 3$. Let $u,\bar{u} \in G$. Then, if either $d(u,\bar{u}) \leq \varepsilon$ or $d_1(u,\bar{u}) \leq \varepsilon$,

$$\Delta V \triangleq \int_{t_o}^{t_f} [\Delta H(t)+\{\Delta H_x+\bar{P}(t)\Delta f(t)\}\delta\hat{x}(t)]dt + e_1$$

$$|e_1| \leq c\varepsilon^3, \quad c < \infty$$

where

$$\Delta H(t) \triangleq H(\bar{x}(t),u(t),\bar{\lambda}(t),t)-H(\bar{x}(t),\bar{u}(t)+\bar{K}(t)\delta\hat{x}(t),\bar{\lambda}(t),t)$$

$$\Delta H_x(t) \triangleq H_x(\bar{x}(t),u(t),\bar{\lambda}(t),t)-H_x(\bar{x}(t),\bar{u}(t)+\bar{K}(t)\delta\hat{x}(t),\bar{\lambda}(t),t)$$

$$\Delta f(t) \triangleq f(\bar{x}(t),u(t),t)-f(\bar{x}(t),\bar{u}(t)+\bar{K}(t)\delta\hat{x}(t),t)$$

and $\bar{\lambda}(t)$, $\bar{P}(t)$ are the solutions of

$$-\dot{\bar{\lambda}}(t) = H_x(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)+\bar{K}^T(t)H_u(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) \tag{4.4}$$

$$-\dot{\bar{P}}(t) = H_{xx}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)+\bar{P}(t)f_x(\bar{x}(t),\bar{u}(t),t)+f_x^T(\bar{x}(t),\bar{u}(t),t)\bar{P}(t)$$

$$+ \bar{K}^T(t)[H_{ux}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) + f_u^T(\bar{x}(t),\bar{u}(t),t)\bar{P}(t)]$$

$$+ [H_{ux}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) + f_u^T(\bar{x}(t),\bar{u}(t),t)\bar{P}(t)]^T\bar{K}(t)$$

$$+ \bar{K}^T(t)H_{uu}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)\bar{K}(t) \qquad (4.5)$$

with the usual terminal conditions.

Proof: $\bar{\lambda}(t) = V_x^{\bar{K}}(\bar{x}(t),t)$

$\bar{P}(t) = V_{xx}^{\bar{K}}(\bar{x}(t),t)$

The proof follows in much the same way as for Proposition 4.2 and Corollary 4.3 using equation (2.86) instead of (2.84).

□

PROPOSITION 4.7: Let the hypothesis of Proposition 4.6 be satisfied. Then

$$\Delta V = \int_{t_o}^{t_f} [\Delta H(t) + \Delta H_x^T \delta\hat{x}(t) + \tfrac{1}{2}\delta\hat{x}^T(t)Q(t)\delta\hat{x}(t)]dt$$

$$+ \tfrac{1}{2}\delta\hat{x}^T(t_f)F_{xx}(\bar{x}(t_f),t_f)\delta\hat{x}(t_f) + e_1$$

$|e_1| \leq cc^3, \quad c < \infty$

where

$$Q(t) \triangleq H_{xx}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)$$

and the remaining terms are as defined in Proposition 4.6.

Proof: See Proposition 4.4. □

PROPOSITTION 4.8: Let H1A, H2A be satisfied, s = 3. Let $u,\bar{u} \in G$. If either

(i) f, L are linear in u and $d(u,\bar{u}) \leq \varepsilon$ or

$d_1(u,\bar{u}) \leq \varepsilon$, or

(ii) $\|u(t)-\bar{u}(t)\| \leq \varepsilon$ for all $t \in T$,

then

$$\Delta V = \int_{t_o}^{t_f} [\tfrac{1}{2} w^T(t) R(t) w(t) + w^T(t)\{R(t)\bar{x}(t) + c(t) + \ell_u^T(\bar{x}(t),\bar{u}(t),t)\bar{p}(t)\}\delta\hat{x}(t)]dt$$

$$+ \int_{t_o}^{t_f} H_u^T(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) w(t) dt + e_1$$

$|e_1| \leq \varepsilon\sigma^3$, $\sigma < \infty$

where

$$R(t) \triangleq H_{uu}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)$$

$$C(t) \triangleq H_{ux}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)$$

$$w(t) \triangleq u(t)-\bar{u}(t)-\bar{R}(t)\delta\hat{x}(t)$$

and $\delta\hat{x}(t)$ is the solution to equations (1.15), (1.16), $\bar{\lambda}(t)$, $\bar{p}(t)$ are the solutions to equations (4.4), (4.5) with the usual boundary conditions.

Proof: Expand the integrand of Proposition 4.6 to second order w.r.t. u about $\bar{u}$. □

COROLLARY 4.9: If, in addition

$$H_u(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) = 0 \qquad \forall t \in T$$

$$R(t)\bar{R}(t)+C(t)+f_u^T(\bar{x}(t),\bar{u}(t),t)\bar{P}(t) = 0 \qquad \text{a.e. } t \in T \qquad (4.6)$$

then

$$\Delta V = \int_{t_o}^{t_f} \tfrac{1}{2} w^T(t) R(t) w(t)\,dt + e_1$$

$$|e_1| \leq c\varepsilon^3, \quad c < \infty.$$

CHAPTER 5

OPTIMIZATION ALGORITHMS

We attempt to find a control which satisfies the H-J-B equation, Chapter 3-I-6. The dynamic programming approach may be summarized as follows:

1. Set $V^o(x(t_f),t_f) = F(x(t_f),t_f)$ (5.1)

2. Solve the partial differential equation

$$V_t^o(x(t),t)+V_x^o(x(t),t)f(x(t),u^o(t),t)+L(x(t),u^o(t),t) = 0 \quad (5.2)$$

where

$$u^o(t) = \arg\min_u [L(x(t),u(t),t)+V_x^o(x(t),t)f(x(t),u(t),t)] \quad (5.3)$$

The difficulty of the numerical solution of (5.2) is enormous, primarily because of the high dimensionality of the equation, which results in tremendous storage requirements. The purpose of the algorithms to be presented is to obtain an approximate solution to (5.2) using a less direct approach. Making use of the equations obtained in Propositions (2.4) - (2.7), we attempt to obtain a solution using successive approximations to the optimal control function.

I THE FIRST ORDER ALGORITHMS

Propositions (2.4) or (2.6) are used.

1. Gradient Method : Kelley [28], Bryson and Denham [2].

Suppose we have a nominal control $\bar{u}(t)$, $t \in T$, with associated trajectory $\bar{x}(t)$.

Choose $u(t)$ to minimize $H(\bar{x},w,\lambda,t)$ w.r.t. w

Write

$$u(t) = \bar{u}(t) + \delta u \tag{5.4}$$

i.e. $\delta u(t) = \arg \min_{\delta w} H(\bar{x},\bar{u}+\delta w,\lambda,t)$

A first order estimate of the minimizing δu is

$$\delta u = -\varepsilon H_u(\bar{x},\bar{u},\lambda,t) \qquad \varepsilon > 0 \tag{5.5}$$

where λ is the solution to (2.35).

Thus, we have for δu sufficiently small

$$\begin{aligned} -\dot{\lambda}(t) &= H(\bar{x},\bar{u}+\delta u,\lambda,t) - H(\bar{x},\bar{u},\lambda,t) \\ &= H_u(\bar{x},\bar{u},\lambda(t),t)\delta u + O(\varepsilon^2) \\ &= -\varepsilon H_u^T(\bar{x},\bar{u},\lambda(t),t)H_u(\bar{x},\bar{u},\lambda(t),t) + O(\varepsilon^2) \end{aligned}$$

$$\begin{aligned} -\dot{\lambda}(t) &= H_x(\bar{x},\bar{u}+\delta u,\lambda,t) + P(t)\Delta f(t) \\ &= H_x(\bar{x},\bar{u},\lambda,t) + O(\varepsilon) \end{aligned}$$

Neglecting the $O(\varepsilon^2)$ and $O(\varepsilon)$ terms gives

$$-\dot{\hat{a}}(t) = -\varepsilon H_u(\bar{x},\bar{u},\hat{\lambda}(t),t)H_u(\bar{x},\bar{u},\hat{\lambda}(t),t) \tag{5.6}$$

$$-\dot{\hat{\lambda}}(t) = H_x(\bar{x},\bar{u},\hat{\lambda}(t),t) \tag{5.7}$$

with boundary conditions

$$\hat{a}(t_f) = 0 \tag{5.8}$$

$$\hat{\lambda}(t_f) = F_x(\bar{x}(t_f),t_f) \tag{5.9}$$

as estimates for $a(t)$, $\lambda(t)$ with

$$\|a(t) - \hat{a}(t)\| \leq c_1\varepsilon^2$$

$$\|\lambda(t) - \hat{\lambda}(t)\| \leq c_2\varepsilon \qquad c_1, c_2 < \infty$$

Note that $\hat{\lambda}(t)$ given by (5.7) is the same as $\bar{\lambda}(t)$ in (2.16). From (5.4), (5.5) the new control is given by

$$u(t) = \bar{u}(t) - \varepsilon H_u(\bar{x},\bar{u},\hat{\lambda}(t),t) \tag{5.10}$$

Computational Procedure

Step 0. Choose nominal control $\bar{u}(t)$, $t \in T$.
Run the $\bar{x}(t)$ trajectory. Store $\bar{u}(t)$, $\bar{x}(t)$.
Calculate the cost $V^{\bar{u}}(x_o,t_o)$, and store.

Step 1. Calculate $\hat{\lambda}(t)$ using equations (5.7), (5.9).
At the same time compute and store the gradient
$H_u(\bar{x},\bar{u},\hat{\lambda}(t),t)$.

Step 2. Choose $\varepsilon > 0$

Using new control (5.10), integrate the state equation forward to obtain $x(t)$.

Calculate $V^{u}(x_o,t_o)$

Step 3. If $V^{u}(x_o,t_o) \geq V^{\bar{u}}(x_o,t_o)$, set $\varepsilon = \varepsilon/2$ and repeat 2.

Otherwise, set $\bar{u}(t) = u(t)$ and go to 1.

Steps 1, 2, 3 are repeated until optimal solution is found or until no further improvement can be found.

Remarks

(i) A stopping rule can be incorporated into the algorithm by replacing Step 1 with:

Step 1* Calculate $\hat{\lambda}(t)$, $\hat{a}(t)$ using equations (5.6) - (5.9). At the same time compute and store the gradient $H_u(\bar{x},\bar{u},\hat{\lambda}(t),t)$.

Check the value of $|a(x_o,t_o)|$ and stop if it is less than a predetermined small positive quantity η, η obtained from numerical stability considerations.

(ii) It is found that the rate of convergence becomes very slow as optimality is approached. This is because in the neighbourhood of the minimum of H w.r.t. w second order terms in the Taylor series expansion dominate, since the first order coefficient

H_u is tending to zero. This means a first order approximation to H in the neighbourhood of its minimum is poor in the sense that it is valid for only very small changes δu.

(iii) We might choose another estimate of the minimizing δu. e.g. setting

$$\delta u = -\varepsilon \operatorname{sgn}(H_u)$$ (see Dyer and McReynolds [8])

where $\operatorname{sgn}(x) = 1 \quad x \geq 0$

$= -1 \quad x < 0$

Then, for δu small enough (i.e. ε small enough)

$$-\dot{a}(t) = H(\bar{x},\bar{u}+\delta u,\lambda,t) - H(\bar{x},\bar{u},\lambda,t)$$

$$= -\varepsilon H_u(\bar{x},\bar{u},\lambda,t) \operatorname{sgn}(H_u(\bar{x},\bar{u},\lambda,t))$$

$$\leq 0$$

i.e. $\dot{a}(t) \geq 0$ and $a(t_f) = 0$

This implies

$\Delta V = a(t_o) \leq 0$, so we have a reduction in cost for $H_u(\bar{x},\bar{u},\lambda,t) \neq 0$ (if $H_u(\bar{x},\bar{u},\lambda,t) = 0$, then $\bar{u}$ is a local minimum).

(iv) Control constraints of the form $g(u(t),t) \leq 0$ can be handled, provided $g(\bar{u}+\delta u,t) \leq 0$.

2. Jacobson's First Order Algorithm [13], [15], [18].

... equations (2.53) and (2.54).

Let $u^* = \arg\min_w H(\bar{x},w,\hat{\lambda}(t),t)$

We let the new control be given by

$$u(t) = \bar{u}(t) \qquad t \in [t_o, t_1]$$
$$= u^*(t) \qquad t \in [t_1, t_f]$$

where t_1 is chosen according to the Step Size Adjustment method given in Chapter 5-II-4.

The computational procedure is similar to that given in Chapter 5-II-4.

Jacobson notes that this algorithm could be expected to have a better convergence rate than the gradient methods in view of the fact that here global changes are made in u.

Control constraints are easily handled by finding the nominal control $\bar{u}(t) \in \Omega$ and

$$\hat{u} = \arg \min_{w \in \Omega} H(\bar{x}, w, \bar{\lambda}(t), t)$$

The new control is then given by

$$u(t) = \bar{u}(t) \qquad t \in [t_o, t_1]$$
$$= \hat{u}(t) \qquad t \in [t_1, t_f].$$

3. Polak and Mayne's First Order Algorithm [37].

This algorithm makes use of the estimate $\Delta\hat{V}$ we obtained for ΔV in Proposition 4.1. It handles control constraints in much the same way as the previous algorithm.

Consider the case where Ω is a closed bounded subset of R^m such that $\max\{\|u\| : u \in \Omega\} \leq r$. The Step size

rule of the last algorithm differs from that of this last algorithm and enables the authors to prove convergence.

Without any loss of generality, let $T \triangleq [0,1]$. Let $\tilde{G}$ be the space of equivalence classes of functions in G which are equal a.e.

The following hypothesis must be satisfied: H1, H2 are satisfied, s = 2. In addition

(i) f_u, f_{ux}, L_u, L_{ux} exist and are continuous in S.

(ii) for any $\bar{u} \in \tilde{G}$, there exists a $\hat{u} \in G$ s.t. for almost all $t \in T$

$$\hat{u}(t) \in \hat{U}(\bar{u},t) \triangleq \arg\min_{w \in \Omega} H(\bar{x}(t),w,\bar{\lambda}(t),t)$$

where $\bar{x}(t) \triangleq x(t;x_o,t_o,\bar{u})$ and $\bar{\lambda}(t)$ is the solution to equations (2.16), (2.17).

We have from Proposition 4.1 that for $\bar{u},u \in \tilde{G}$

$$\Delta\hat{V} = \int_0^1 [H(\bar{x},u,\bar{\lambda},t) - H(\bar{x},\bar{u},\bar{\lambda},t)]dt$$

is an estimate for $\Delta V \triangleq V^u(x_o,t_o) - V^{\bar{u}}(x_o,t_o)$ such that

$$|\Delta V - \Delta\hat{V}| \leq c[d(u,\bar{u})]^2 \qquad c < \infty.$$

For some $\bar{u} \in \tilde{G}$, define

$$\hat{U}(u) \triangleq \hat{U}(\bar{u},\cdot) \cap \tilde{G}$$
$$\bar{H}(\bar{u},t) \triangleq \min_{w \in \Omega} H(\bar{x},w,\bar{\lambda},t), \quad t \in T \qquad (5.11)$$

and

$$\theta(\bar{u}) \triangleq \Delta V(\hat{u},\bar{u})$$

$$= \int_0^1 [H(\bar{x},\hat{u},\lambda,t)-H(\bar{x},\bar{u},\bar{\lambda},t)]dt$$

$$= \int_0^1 [\bar{H}(\bar{u},t)-H(\bar{x},\bar{u},\bar{\lambda},t)]dt \qquad (5.12)$$

where $\hat{u} \in \hat{U}(\bar{u})$.

The purpose of the algorithm is to determine a $\bar{u} \in G$ such that $\theta(\bar{u}) = 0$, i.e. $\bar{u}$ is desirable.

Step Size Rule

Define $I^H_{\bar{u}} \subset T$ by

$$I^H_{\bar{u}} \triangleq \{t \in T \mid \bar{H}(\bar{u},t)-H(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) \leq 0(\bar{u})\}$$

Note the $I^H_{\bar{u}}$ consists of the union of at most a countable number of disjoint intervals.

Let $m(\bar{u}) \triangleq \mu(I^H_{\bar{u}})$.

Now for $\alpha \in [0,1]$, let $I_{\alpha\bar{u}}$ be any subset of T having the following properties (the last 2 of which are designed to make $I_{\alpha\bar{u}}$ the union of a small number of disjoint intervals).

(i) $\mu(I_{\alpha\bar{u}}) = \alpha$

(ii) If $\alpha \in [0,m(\bar{u})]$, $I_{\alpha\bar{u}} \subset I^H_{\bar{u}}$

(iii) If $\alpha \in (m(\bar{u}),1]$, $I_{\alpha\bar{u}} \supset I^H_{\bar{u}}$

(iv) For all $\alpha \in [0,m(\bar{u})]$, $\{t \in I^H_{\bar{u}}, t' \in I_{\alpha\bar{u}}, t < t'\} \Rightarrow \{t \in I_{\alpha\bar{u}}\}$

(v) For all $\alpha \in (m(\bar{u}),1]$, $\{t \in T, t' \in I_{\alpha\bar{u}} \setminus I^H_{\bar{u}}, t < t'\} \Rightarrow \{t \in I_{\alpha\bar{u}}\}$

Define, for $\bar{u} \in \tilde{G}$, $\alpha \in [0,1]$

$$u_\alpha(t) \in \hat{U}(\bar{u},t) \qquad \forall t \in I_{\alpha u} \tag{5.13}$$

$$u_\alpha(t) = \bar{u}(t) \qquad \forall t \in T \setminus I_{\alpha u}.$$

We calculate the α which gives us the 'step length' from:

$$\alpha(\bar{u}) = \{\alpha \mid \alpha = \max\{\alpha \in [0,1] \mid \Delta V(u_{\alpha'},\bar{u}) \leq \alpha'\theta(\bar{u})/2\} \quad \forall \alpha' \in [0,\alpha]\} \tag{5.14}$$

where u_α is any function satisfying (5.13).

The Computational Procedure

Step 0. Select a nominal control $\bar{u}(t) \in \tilde{G}$.

Step 1. Run the nominal trajectory $\bar{x}(t)$. Compute the nominal cost $V^{\bar{u}}(\bar{x}_o,t_o)$. Store $\bar{x}$, $V^{\bar{u}}$.

Step 2. Using boundary condition (2.17) integrate equation (2.16) backwards in time (to obtain $\bar{\lambda}(t)$), all the while minimizing $H(\bar{x},w,\bar{\lambda}(t),t)$ w.r.t. w to obtain $\hat{u}(t)$. Store $\bar{\lambda}(t)$, $\hat{u}(t)$.

Step 3. Compute $\theta(\bar{u})$ using (5.12).
If $\theta(\bar{u}) = 0$ Stop.
Else go to step 4.

Step 4. Compute an $\alpha \in \alpha(\bar{u})$ from (5.14) to obtain the step length.

Step 5. Set $\bar{u}(t) = u_{\alpha}(t)$
Go to 1.

II THE SECOND ORDER ALGORITHMS

Propositions (2.5) and (2.7) are used. The $\gamma_i(t)$ are set identically equal to zero for all $t \in T$, $i=1, \ldots n$.

SMALL VARIATIONS IN CONTROL

For a detailed discussion and comparison of the next three algorithms the reader should consult Jacobson, [13], [14].

1. Jacobson's Second Order Algorithm [13], [14].

Let $u^*(t) = k(\bar{x}(t),t)$ where

$$u^*(t) = \arg\min_u H(\bar{x},u,\hat{\lambda}(t),t)$$

so $H_u(\bar{x},u^*,\hat{\lambda}(t),t) = 0$

Write $u^*(t) = \bar{u}(t) + \delta u^*(t)$

i.e. $$\delta u^*(t) = \arg\min_{\delta u} H(\bar{x},\bar{u},+\delta u,\hat{\lambda}(t),t) \qquad (5.15)$$

where $\hat{\lambda}(t)$ is the solution to equations (2.61),(2.64).

We may obtain a second order estimate of the *minimizing* δu^* by *expanding* H to second order in δu about $\bar{u}(t)$, differentiating w.r.t. δu and equating to zero, i.e.

$$H(\bar{x},\bar{u}+\delta u,\hat{\lambda}(t),t) = H + \langle H_u,\delta u\rangle + \tfrac{1}{2}\langle \delta u\, H_{uu}\delta u\rangle$$

Thus

$$H_u(\bar{x},\bar{u},\hat{\lambda}(t),t) + H_{uu}(\bar{x},\bar{u},\hat{\lambda}(t),t)\delta u^* = 0$$

so

$$\delta u^*(t) = -H_{uu}^{-1}(\bar{x},\bar{u},\hat{\lambda}(t),t)H_u(\bar{x},\bar{u},\hat{\lambda}(t),t) \tag{5.16}$$

This is a Newton-Raphson minimization method which produces a good estimate of the minimizing δu^* if;

(i) δu^* is small

(ii) $H_{uu}^{-1}(\bar{x},\bar{u},\hat{\lambda}(t),t) > 0 \qquad \forall t \in T.$ (5.17)

To ensure that δu^* is small we introduce $\varepsilon : 0<\varepsilon\leq 1$ into (5.16), i.e.

$$\delta u^*(t) = -\varepsilon H_{uu}^{-1}(\bar{x},\bar{u},\hat{\lambda}(t),t)H_u(\bar{x},\bar{u},\hat{\lambda}(t),t). \tag{5.18}$$

Now $K(t)$ is chosen to minimize $\hat{P}(t)$. Therefore, from equation (2.62) we have

$$K(t) = -H_{uu}^{-1}(\bar{x},u^*,\hat{\lambda}(t),t)[H_{xu}(\bar{x},u^*,\hat{\lambda}(t),t)+f_u^T(\bar{x},u^*,t)\hat{P}(t)] \tag{5.19}$$

and our new control is given by:

$$\begin{aligned} u(t) &= k(x(t),t) \\ &= u^*(t) + K(t)\delta x \\ &= \bar{u}(t) + \varepsilon\delta u^*(t) + K(t)\delta x \end{aligned} \tag{5.20}$$

We have from the remarks before Proposition 2.5 that our equations are only valid for δx small enough. The δx's produced in T are due only to the effect of $\delta u^*(t)+\delta u(t)$ acting through

$$(\bar{x}+\delta x)^{\cdot} = f(\bar{x}+\delta x, \bar{u}+\delta u^*+\delta u, t)$$

since $\delta x(t_o) = 0$.

It is shown in [13], [14] that an ε can be found that will limit the size of $\delta u^*, \delta x, \delta u$.

We solve the following equations:

$$-\dot{\hat{a}}(t) = H(\bar{x}(t), u^*(t), \hat{\lambda}(t), t) - H(\bar{x}(t), \bar{u}(t), \hat{\lambda}(t), t) \tag{5.21}$$

$$-\dot{\hat{\lambda}}(t) = H_x(\bar{x}(t), u^*(t), \hat{\lambda}(t), t) + \hat{P}(t)\Delta f(t) \tag{5.22}$$

$$\begin{aligned}-\dot{\hat{P}}(t) = {} & H_{xx}(\bar{x}(t), u^*(t), \hat{\lambda}(t), t) + f_x^T(\bar{x}(t), u^*(t), t)\hat{P}(t) \\ & + \hat{P}(t) f_x(\bar{x}(t), u^*(t), t) \\ & + K^T(t)[H_{ux}(\bar{x}(t), u^*(t), \hat{\lambda}(t), t) + f_u^T(\bar{x}(t), u^*(t), t), \hat{P}(t)] \\ & + [H_{ux}(\bar{x}(t), u^*(t), \hat{\lambda}(t), t) + f_u^T(\bar{x}(t), u^*(t), t)\hat{P}(t)]^T K(t) \\ & + K^T(t) H_{uu}(\bar{x}(t), u^*(t), \hat{\lambda}(t), t) K(t)\end{aligned} \tag{5.23}$$

with boundary conditions:

$$\hat{a}(t_f) = 0 \tag{5.24}$$

$$\hat{\lambda}(t_f) = F_x(\bar{x}(t_f), t_f) \tag{5.25}$$

$$\hat{P}(t_f) = F_{xx}(\bar{x}(t_f), t_f) \tag{5.26}$$

where $u^*(t) = \bar{u}(t) + \delta u^*(t)$,

$\delta u^*(t)$ is given in equ (5.18), and $K(t)$ is given in (5.19).

Computational Procedure

Step 0. Choose a nominal control $\bar{u}(t)$, $t \in T$
Run the $\bar{x}(t)$ trajectory.
Calculate the nominal cost $V^{\bar{u}}(x_o, t_o)$.
Store, $\bar{u}(t)$, $\bar{x}(t)$, $V^{\bar{u}}$.

Step 1. Guess a value for ε, $0 < \varepsilon \leq 1$.
Using boundary conditions (5.24), (5.25), (5.26) integrate equations (5.21), (5.22), (5.23) backwards in time from t_f to t_o, all the while storing $K(t)$, $H_{uu}^{-1} H_u$.
Check the value of $|a(x_o, t_o)|$ and stop if it is less than a predetermined η. η obtained from numerical stability considerations.

Step 2. Using new control (5.20) integrate the state equations forward to obtain $x(t)$.
Calculate $V^u(x_o, t_o)$.

Step 3. If $V^u(x_o, t_o) \geq V^{\bar{u}}(x_o, t_o)$, set $\varepsilon = \varepsilon/2$ and go to 1 otherwise, set $\bar{u}(t) = u(t)$ and go to 1.

Drawbacks

(i) We need to assume that $H_{uu}^{-1}(\bar{x},\bar{u},\hat{\lambda},(t),t) > 0$ and $H_{uu}^{-1}(\bar{x},\bar{u}+\delta u^*,\hat{\lambda}(t),t) > 0$. In non-linear problems along non-optimal trajectories these requirements are often violated with resultant failure of the method.

(ii) The procedure is complicated by the fact that if ε is guessed too large, the backwards differential equations have to be reintegrated along the same nominal trajectory with smaller ε.
This last drawback is overcome by the next algorithm.

2. Mayne's Second Order Algorithm [33].

Use equations (2.45) - (2.50)
Let $u^*(t) = k(\bar{x}(t),t)$
$= \bar{u}(t) + \delta u^*(t)$

We have from (5.16) that

$$\delta u^*(t) = -\varepsilon H_{uu}^{-1}(\bar{x},\bar{u},\lambda(t),t)H_u(\bar{x},\bar{u},\lambda(t),t) \qquad (5.27)$$

Assume δu^* is small, i.e. either ε is small and/or H_u is small. Then:

$$\begin{aligned}-\dot{\lambda}(t) &= H(\bar{x},\bar{u}+\delta u^*,\lambda(t),t)-H(\bar{x},\bar{u},\lambda(t),t)\\ &= H_u(\bar{x},\bar{u},\lambda(t),t)\delta u^*+\tfrac{1}{2}\delta u^{*T}H_{uu}(\bar{x},\bar{u},\lambda(t),t)\delta u^* + O(\varepsilon^3)\\ &= -\varepsilon(1-\varepsilon/2)H_u^T(\bar{x},\bar{u},\lambda(t),t)H_{uu}^{-1}(\bar{x},\bar{u},\lambda(t),t)H_u(\bar{x},\bar{u},\lambda(t),t) + O(\varepsilon^3) \qquad (5.28)\end{aligned}$$

$$-\dot{\lambda}(t) = H_x(\bar{x},\bar{u}+\delta u^*,\lambda(t),t)+K^T(t)H_u(\bar{x},\bar{u}+\delta u^*,\lambda(t),t) + P(t)[f(\bar{x},\bar{u}+\delta u^*,t)-f(\bar{x},\bar{u},t)]$$

$$= H_x(\bar{x},\bar{u},\lambda(t),t)+H_{ux}(\bar{x},\bar{u},\lambda(t),t)\delta u^*+P(t)f_u(\bar{x},\bar{u},t)\delta u^* +K^T(t)H_u(\bar{x},\bar{u},\lambda(t),t)+K^T(t)H_{uu}(\bar{x},\bar{u},\lambda(t),t)\delta u^* +O(\varepsilon^2)$$

$$= H_x(\bar{x},\bar{u},\lambda(t),t)+K^T(t)H_u(\bar{x},\bar{u},\lambda(t),t)-\varepsilon K^T(t)H_u(\bar{x},\bar{u},\lambda(t),t) -\varepsilon[H_{ux}(\bar{x},\bar{u},\lambda(t),t)+P(t)f_u(\bar{x},\bar{u},t)]^T H_{uu}^{-1}(\bar{x},\bar{u},\lambda(t),t)H_u(\bar{x},\bar{u},\lambda(t),t)+O(\varepsilon^2) \tag{5.29}$$

and as above, setting $\gamma_i(t) \equiv 0$, i=1, ..n

$$-\dot{P}(t) = H_{xx}(\bar{x},\bar{u},\lambda(t),t)+f_x^T(\bar{x},\bar{u},t)P(t)+P(t)f_x(\bar{x},\bar{u},t) + K^T(t)[H_{ux}(\bar{x},\bar{u},\lambda(t),t)+f_u^T(\bar{x},\bar{u},t)P(t)] + [H_{ux}(\bar{x},\bar{u},\lambda(t),t)+f_u^T(\bar{x},\bar{u},t)P(t)]^T K(t) + K^T(t)H_{uu}(\bar{x},\bar{u},\lambda(t),t)K(t) + O(\varepsilon) \tag{5.30}$$

Now, ignoring the terms $O(\varepsilon)$, $O(\varepsilon^2)$, $O(\varepsilon^3)$ in (5.30), (5.29), (5.28) respectively, we introduce an error of the order of ε in P(t), ε^2 in $\lambda(t)$, ε^3 in a(t).

We choose K(t) to minimize $\hat{P}(t)$ (ignoring the $O(\varepsilon)$ terms). We get

$$K(t) = -H_{uu}^{-1}(\bar{x},\bar{u},\hat{\lambda}(t),t)[H_{ux}(\bar{x},\bar{u},\hat{\lambda}(t),t)+f_u^T(\bar{x},\bar{u},t)\hat{P}(t)] \tag{5.31}$$

Our new control is given by

$$u(t) = \bar{u}(t) + \delta u^*(t) + K(t)\delta x(t) \tag{5.32}$$

We solve the following equations:

$$-\dot{\hat{a}}(t) = -\varepsilon(1-\varepsilon/2)H_u^T(\bar{x},\bar{u},\hat{\lambda}(t),t)H_{uu}^{-1}(\bar{x},\bar{u},\hat{\lambda}(t),t)H_u(\bar{x},\bar{u},\hat{\lambda}(t),t) \tag{5.33}$$

$$-\dot{\hat{\lambda}}(t) = H_x(\bar{x},\bar{u},\hat{\lambda}(t),t) + K^T(t)H_u(\bar{x},\bar{u},\hat{\lambda}(t),t) \tag{5.34}$$

$$\begin{aligned}-\dot{\hat{P}}(t) = {} & H_{xx}(\bar{x},\bar{u},\hat{\lambda}(t),t) + f_x^T(\bar{x},\bar{u},t)\hat{P}(t) + \hat{P}(t)f_x(\bar{x},\bar{u},t) \\ & + K^T(t)[H_{ux}(\bar{x},\bar{u},\hat{\lambda},t) + f_u^T(\bar{x},\bar{u},t)\hat{P}(t)] \\ & + [H_{ux}(\bar{x},\bar{u},\hat{\lambda},t) + f_u^T(\bar{x},\bar{u},t)\hat{P}(t)]^T K(t) \\ & + K^T(t)H_{uu}(\bar{x},\bar{u},\hat{\lambda}(t),t)K(t)\end{aligned} \tag{5.35}$$

with boundary conditions

$$\hat{a}(t_f) = 0 \tag{5.36}$$

$$\hat{\lambda}(t_f) = F_x(\bar{x}(t_f),t_f) \tag{5.37}$$

$$\hat{P}(t_f) = F_{xx}(\bar{x}(t_f),t_f) \tag{5.38}$$

where $\hat{a}(t)$, $\hat{\lambda}(t)$, $\hat{P}(t)$ are estimates for $a(t)$, $\lambda(t)$ and $P(t)$ respectively such that eqns (2.66) - (2.68) are satisfied.

The computational procedure of II - 1 is modified in the following way. If ε is guessed too large so that no improvement in cost results, the backwards equations do not have to be re-integrated along the same nominal trajectory for smaller ε since they are now independent of ε.

This is not true of the $\dot{a}$ equation, but $a(t) = -\varepsilon(1-\varepsilon/2)\int H_u^T H_{uu} H_u d\tau$, so $a(t)$ is calculated easily for different ε's.

Equations (5.34), (5.35) are the same as those obtained by Dyer and McReynolds [8] for their successive sweep algorithm, which is the same as the modified algorithm above. See also McReynolds [41].

3. The Second Variation Algorithm. McReynolds and Bryson [40], Mitter [44].

Consider the equations given in Proposition 2.5 where we recall that $V_x^u(\bar{x}(t),t) \triangleq \lambda(t)$.
Let us now set

$$V_x^u(\bar{x}(t),t) = \lambda(t) + h(t) \tag{5.39}$$

and $V_{xx}^u(\bar{x}(t),t) = P(t)$, as before.

Let $u^*(t) = k(\bar{x}(t),t)$ where

$$u^*(t) = \arg\min_u H(\bar{x},u,\lambda(t)+h(t),t).$$

Write $u^*(t) = \bar{u}(t)+\delta u^*$

i.e. $$\delta u^* = \arg\min_{\delta u} H(\bar{x},\bar{u}+\delta u,\lambda(t)+h(t),t). \tag{5.40}$$

Expanding H to second order in δu about $\bar{u}$ gives:

$$H(\bar{x},\bar{u}+\delta u,\lambda+h,t) = H(\bar{x},\bar{u}+\delta u,\lambda,t)+\langle h, f(\bar{x},\bar{u}+\delta u^*,t)\rangle$$

$$= H(\bar{x},\bar{u},\lambda,t)+\langle H_u(\bar{x},\bar{u},\lambda,t),\delta u\rangle+\tfrac{1}{2}\langle \delta u, H_{uu}(\bar{x},\bar{u},\lambda,t)\delta u\rangle$$

$$+\langle h, f(\bar{x},\bar{u},t)+f_u(\bar{x},\bar{u},t)\delta u+\tfrac{1}{2}f_{uu}(\bar{x},\bar{u},t)\delta u\delta u\rangle. \qquad (5.41)$$

On an optimal trajectory $-\dot{\lambda} = H_x$, so $V_x = \lambda$, so h is zero. Now, although we are not on an optimal trajectory, let us assume that h is small. We can then neglect the term $\langle h, \tfrac{1}{2}f_{uu}\delta u^2\rangle$ as it can be considered third-order.

Neglecting $\langle h, f_{uu}\delta u^2\rangle$ and differentiating the RHS of (5.41) w.r.t. δu and setting to zero (for a minimum) gives

$$H_u + H_{uu}\delta u^* + f_u^T h = 0$$

i.e. $$\delta u^*(t) = -H_{uu}^{-1}(\bar{x},\bar{u},\lambda,t)[H_u(\bar{x},\bar{u},\lambda,t) + f_u^T(\bar{x},\bar{u},t)h(t)]$$

Now, for the same reasons as those given in II-1 we introduce ε : $0 < \varepsilon \leq 1$ and set

$$\delta u^*(t) = -\varepsilon H_{uu}^{-1}(\bar{x},\bar{u},\lambda,t)[H_u(\bar{x},\bar{u},\lambda,t) + f_u^T(\bar{x},\bar{u},t)h(t)] \qquad (5.42)$$

Assume δu^* is small i.e. either ε is small and/or H_u is small (h is assumed small already). Then

$$-\dot{h}(t) = H(\bar{x},\bar{u}+\delta u^*,\lambda+h,t) - H(\bar{x},\bar{u},\lambda+h,t)$$

$$= H(\bar{x},\bar{u}+\delta u^*,\lambda,t) + \langle h, f(\bar{x},\bar{u}+\delta u^*,t)\rangle$$

$$- H(\bar{x},\bar{u},\lambda,t) - \langle h, f(\bar{x},\bar{u},t)\rangle$$

$$= H_u(\bar{x},\bar{u},\lambda,t)\delta u^* + \tfrac{1}{2}\delta u^{*T}H_{uu}(\bar{x},\bar{u},\lambda,t)\delta u^*$$

$$+ h^T f_u(\bar{x},\bar{u},t)\delta u^* + O(\varepsilon^3)$$

$$= -\varepsilon H_u^T(\bar{x},\bar{u},\lambda,t)H_{uu}^{-1}(\bar{x},\bar{u},\lambda,t)[H_u(\bar{x},\bar{u},\lambda,t)+f_u^T(\bar{x},\bar{u},t)h]$$

$$+\tfrac{1}{2}\varepsilon^2[H_u(\bar{x},\bar{u},\lambda,t)+f_u^T(\bar{x},\bar{u},t)h]^T H_{uu}^{-1}(\bar{x},\bar{u},\lambda,t)[H_u(\bar{x},\bar{u},\lambda,t)+f_u^T(\bar{x},\bar{u},t)h]$$

$$-\varepsilon h^T f_u(\bar{x},\bar{u},t)H_{uu}^{-1}(\bar{x},\bar{u},\lambda,t)[H_u(\bar{x},\bar{u},\lambda,t)+f_u^T(\bar{x},\bar{u},t)h] \quad + O(\varepsilon^3)$$

$$= -\varepsilon(1-\varepsilon/2)[(H_u(\bar{x},\bar{u},\lambda,t)+f_u^T(\bar{x},\bar{u},t)h]^T H_{uu}^{-1}(\bar{x},\bar{u},\lambda,t)[H_u(\bar{x},\bar{u},\lambda,t)+f_u^T(\bar{x},\bar{u},t)h]$$

$$+ O(\varepsilon^3) \qquad (5.43)$$

$$-\dot{\lambda}(t)-\dot{h}(t) = H_x(\bar{x},\bar{u}+\delta u^*,\lambda+h,t) + K^T(t)H_u(\bar{x},\bar{u}+\delta u^*,\lambda+h,t)$$

$$+P(t)[f(\bar{x},\bar{u}+\delta u^*,t) - f(\bar{x},\bar{u},t)]$$

$$= H_x(\bar{x},\bar{u}+\delta u^*,\lambda,t) + <h, f_x(\bar{x},\bar{u}+\delta u^*,t)>$$

$$+K^T(t)[H_u(\bar{x},\bar{u}+\delta u^*,\lambda,t)+f_u^T(\bar{x},\bar{u}+\delta u^*,t)h]$$

$$+P(t)[f(\bar{x},\bar{u}+\delta u^*,t) - f(\bar{x},\bar{u},t)]$$

$$= H_x(\bar{x},\bar{u},\lambda,t) + H_{ux}(\bar{x},\bar{u},\lambda,t)\delta u^* + f_x^T(\bar{x},\bar{u},t)h$$

$$+K^T(t)H_u(\bar{x},\bar{u},\lambda,t) + K^T(t)H_{uu}(\bar{x},\bar{u},\lambda,t)\delta u^*$$

$$+K^T(t)f_u^T(\bar{x},\bar{u},t)h + P(t)f_u(\bar{x},\bar{u},t)\delta u^* \qquad +O(\varepsilon^2)$$

$$= H_x(\bar{x},\bar{u},\lambda,t)+f_x^T(\bar{x},\bar{u},t)h+K^T(t)[H_u(\bar{x},\bar{u},\lambda,t)+f_u^T(\bar{x},\bar{u},t)h]$$

$$-\varepsilon[K^T(t)+H_{ux}(\bar{x},\bar{u},\lambda,t)+P(t)f_u(\bar{x},\bar{u},t))H_{uu}^{-1}(\bar{x},\bar{u},\lambda,t)][f^T(\bar{x},\bar{u},t)h$$

$$+H_u(\bar{x},\bar{u},\lambda,t)] \quad +O(\varepsilon^2) \qquad (5.44)$$

$$-\dot{P}(t) = H_{xx}(\bar{x},\bar{u},\lambda,t)+f_x^T(\bar{x},\bar{u},t)P(t)+P(t)f_x(\bar{x},\bar{u},t)$$

$$+ K^T(t)[H_{ux}(\bar{x},\bar{u},\lambda,t)+f_u^T(\bar{x},\bar{u},t)P(t)]$$

$$+ [H_{ux}(\bar{x},\bar{u},\lambda,t)+f_u^T(\bar{x},\bar{u},t)P(t)]^T K(t)$$

$$+ K^T(t)H_{uu}^{-1}(\bar{x},\bar{u},\lambda,t)K(t) \qquad +O(\varepsilon) \qquad (5.45)$$

Ignoring the terms $O(\varepsilon)$, $O(\varepsilon^2)$, $O(\varepsilon^3)$ in (5.45), (5.44), (5.43) introduces an error of the order ε in $P(t)$, ε^2 in $\lambda(t) + h(t)$ and ε^3 in $a(t)$.

We choose $K(t)$ to minimize $\dot{\hat{P}}(t)$ (ignoring the $O(\varepsilon)$ terms). We get

$$K(t) = -H_{uu}^{-1}(\bar{x},\bar{u},\hat{\lambda},t)[H_{ux}(\bar{x},\bar{u},\hat{\lambda},t)+f_u^T\hat{P}(t)]$$

Now, if we set $-\dot{\hat{\lambda}} = H_x$, we solve the following eqns:

$$-\dot{\hat{a}}(t) = -\varepsilon(1-\varepsilon/2)[H_u(\bar{x},\bar{u},\hat{\lambda},t)+f_u^T(\bar{x},\bar{u},t)h]^T H_{uu}^{-1}(\bar{x},\bar{u},\hat{\lambda},t)[H_u(\bar{x},\bar{u},\hat{\lambda},t) + f_u^T(\bar{x},\bar{u},t)h]$$

$$-\dot{\hat{\lambda}}(t) = H_x(\bar{x},\bar{u},\hat{\lambda},t)$$

$$-\dot{\hat{h}}(t) = f_x^T(\bar{x},\bar{u},t)+K^T(t)[H_u(\bar{x},\bar{u},\hat{\lambda},t)+f_u^T(\bar{x},\bar{u},t)h]$$

$$\begin{aligned}-\dot{\hat{P}}(t) = {} & H_{xx}(\bar{x},\bar{u},\hat{\lambda},t) + f_x^T(\bar{x},\bar{u},t)\hat{P} + \hat{P}(t)f_x(\bar{x},\bar{u},t) \\ & + K^T(t)[H_{ux}(\bar{x},\bar{u},\hat{\lambda},t)+f_u^T(\bar{x},\bar{u},t)\hat{P}(t)] \\ & + [H_{ux}(\bar{x},\bar{u},\hat{\lambda},t)+f_x^T(\bar{x},\bar{u},t)\hat{P}(t)]^T K(t) \\ & + K^T(t)H_{uu}(\bar{x},\bar{u},\hat{\lambda},t)K(t)\end{aligned}$$

with boundary conditions

$$\begin{aligned}\hat{a}(t_f) &= 0 \\ \hat{\lambda}(t_f) &= F_x(\bar{x}(t_f),t_f) \\ \hat{h}(t_f) &= 0 \\ \hat{P}(t_f) &= F_{xx}(\bar{x}(t_f),t_f)\end{aligned}$$

The computational procedure will be the same as that for the previous algorithm except that we now have an extra differential equation to solve. We have seen that in order to derive the Second Variation equations we had to assume h(t) small, $t \in T$, along non-optimal trajectories. As this is only true in the neighbourhood of an optimal trajectory Jacobson [13], [14], points out that the Second Variation method is not second order in the sense of D.D.P.

SUMMARY

The three second-order algorithms presented to far have the following drawbacks.

(i) $H_{uu}(\bar{x},\bar{u},\lambda,t)$ must be positive definite. This restriction is severe since it implies that H is strictly convex globally w.r.t. u. In many problems we find that H is non-convex globally w.r.t. u although it is strictly convex in the neighbourhood of its minimum w.r.t. u.

(ii) Inequality constraints on control variables cannot be handled directly. They have to be approximated by penalty functions.

(iii) Requirement (i) excludes the bang-bang type of control problem where $H_{uu} \equiv 0$.

We now dicuss an approach which can overcome difficulties (i) and (ii) using strong variations in δx.

GLOBAL VARIATIONS IN CONTROL

4. Jacobson's Second Order Algorithm [13], [15], [18].

We use equations (2.60), (2.61), (2.62).

Let $u^*(t) = k(\bar{x}(t),t)$ where

$$u^*(t) = \arg\min_u H(\bar{x},u,\hat{\lambda},t) \tag{5.46}$$

$K(t)$ is chosen to minimize $\hat{p}(t)$ for all $t \in T$ with $\gamma_i(t) \equiv 0$ $i=1,..n$. So

$$K(t) = -H_{uu}(\bar{x},u^*,\hat{\lambda},t)[H_{ux}(\bar{x},u^*,\hat{\lambda},t)+f_u^T(\bar{x},u^*,t)\hat{P}(t)] \tag{5.47}$$

The new control is

$$\begin{aligned} u(t) &= \bar{u}(t) \qquad t \in T \setminus E \\ &= u^*(t)+K(t)[x(t)-\bar{x}(t)] \qquad t \in E \end{aligned} \tag{5.48}$$

where $t_1 \in T \setminus E$, $t_2 \in E$ implies $t_1 < t_2$. $\mu(E)$ is chosen to ensure satisfactory cost reduction.

Step-Size Adjustment Method

This has been described in detail in [13], [15], [18]. Substituting (5.48) into the state equation gives

$$\begin{aligned} (\bar{x}(t)+\delta x(t))^{\cdot} &= f(\bar{x}+\delta x,u^*+K\delta x,t) \qquad t \in E \\ x(t_0)+\delta x(t_0) &= x_0 \end{aligned} \tag{5.49}$$

Because $\delta x(t_o) = 0$, the δx produced by equation is due to the driving action of $\delta u^* = u^* - \bar{u}$. We saw in the analysis of Chapter 2-4, that we need to restrict the size of δx.

Suppose we run along the trajectory $\bar{x}(t)$ from t_o to t_1, $t_o \leq t_1 < t_f$. At time $t = t_1$, $\delta x(t_1) = 0$. Integrating equ (5.49) over $[t_1, t_f]$, the δx produced in this interval will be small if $[t_1, t_f]$ is small, even for large δu^*, i.e. we use control u(t) defined by (5.48), with $t \in [t_1, t_f]$.

Notice that $\bar{u}(t_1)$ may be different from $u^*(t_1)$ and the new trajectory will sometimes have a corner at t_1. We are thus dealing with the notion of strong variations, because, although $\|\delta x\| < \varepsilon$, it is not necessarily true that $\|\delta \dot{x}\| < \varepsilon$.

The improvement in cost on application of a new control $u(t) = \bar{u}(t) + K(t)\delta x(t)$ is given by

$$\Delta V = V^{u}(x_o, t_o) - V^{\bar{u}}(x_o, t_o) \tag{5.50}$$

The predicted improvement in cost, using the new control on interval $[t_1, t_f]$ is given by

$$|\delta(\bar{x}, t_1)| = \left| \int_{t_f}^{t} [H(\bar{x}, u^*, \hat{\lambda}, t) - H(\bar{x}, \bar{u}, \hat{\lambda}, t)]dt \right| \tag{5.51}$$

The new trajectory is considered to be satisfactory if the following inequality is satisfied

$$-\Delta V > |a(\bar{x},t_1)|\ C \qquad C > 0 \qquad (5.52)$$

In practise C is set, say as, 0.5.

C must be greater than or equal to zero because positive ΔV is inadmissible, and less than unity because the improvements in cost should not be greater than the predicted improvement. Moreover C should be somewhat less than unity so that decisions based on (5.52) are not influenced by round off errors in the computation.

When using a digital computer we divide the interval into N-1 time steps (i.e. t from 1 to N). When integrating the backwards equations record the time N_{eff} when $|a(\bar{x},t)|$ becomes different from zero. We need to find $t_1 = N_1 \in [1,N_{eff}]$. If (5.52) is satisfied with $N_1 = 1$, then a reasonable reduction in cost has been made and the next iteration of the main algorithm may begin.

If (5.52) is not satisfied, set

$$N_1 = (N_{eff}-N_0)/2 + N_0 = N_{01}.$$

Repeat the above procedure. If (5.52) is satisfied, the next iteration is begun. Otherwise set

$$N_1 = (N_{eff}-N_{01})/2 + N_{01} = N_{02}.$$

Continuing to subdivide $[1,N_{eff}]$ in this way, we have

$$N_1 = (N_{eff}-N_{or})/2 + N_{or} = N_{or+1} \qquad r = 0,1, \ldots$$

where $N_{00} = 2N_1 - N_{eff} = 2 - N_{eff}$.

If $N_{eff} = 1$, then only $r = 0$ is used.
If r is increased until $N_1 = N_{eff} - 1$ and condition (5.52) is still not satisfied, then owing to round off errors, or N being too small, criterion (5.52) with $C = 0.5$ may be too severe. Then set $C = 0.0$ and repeat the procedure for determining N_1. $C = 0.0$ only asks that $\Delta V < 0$.

Summary of S.S.A.M.

Step 0. Obtain from main algorithm the time N_{eff} when $|a(\bar{x},t)|$ becomes greater than η_1. η_1 a small positive quantity obtained from numerical stability considerations.

Step 1. Set $C = 0,5$.

Step 2. Set $r = 0$.

Step 3. $N_1 = (N_{eff}-N_{or})/2+N_{or}$ $\quad N_{or+1}$, $N_{oc} = 2-N_{eff}$

Define $u = \bar{u}$ on $[1,N_1]$

$u = u^*+K(t)\delta x$ on $[N_1,N]$

Calculate the cost $V^u(x_o,1)$ and hence the improvement $\Delta V = V^u(x_o,1) - V^{\bar{u}}(\bar{x}_o,1)$

Step 4. If Criterion $\dfrac{\Delta V}{|a(\bar{x},N_1)|} > C$ is satisfied, N_1 is satisfactory. Return the improved control, trajectory and cost to the main algorithm. Otherwise go to 5.

Step 5. If $N_1 = N_{eff} - 1$ or $N_{eff} = 1$ go to 6
Otherwise set r = r + 1 and go to 3.

Step 6. If C = 0.0, stop. No improvement in Trajectory attainable.
Otherwise, set C = 0.0 and go to 2.

The Computational Procedure

Step 0. Choose a nominal control $\bar{u}(t)$, $t \in T$. Store.

Step 1. Run nominal $\bar{x}(t)$ trajectory. Calculate the cost $V^{\bar{u}}(x_o, t_o)$. Store $\bar{x}$, $V^{\bar{u}}$.

Step 2. Using boundary conditions (2.63), (2.64), (2.65) integrate eqns (2.60), (2.61), (2.62) backward in time from t_f to t_o, all the while minimizing H w.r.t. u to obtain $u^*(t)$ and K(t). Store $u^*(t)$, K(t).
Note the time N_{eff} when $|a(\bar{x},t)|$ becomes greater than η.

Step 3. If $N_{eff} < 1$. Stop. Optimal control found
Otherwise go to 4.

Step 4. Apply the S.S.A.M.
If an improved control cannot be found, then S.S.A.M. halts computation.
Otherwise S.S.A.M. returns an improved control u(t), trajectory and cost.

Step 5. Set $\bar{u} = u$, $\bar{x} = x$, $V^{\bar{u}}(\bar{x}_o,t_o) = V^u(x_o,t_o)$
Store the new $\bar{x}$, $\bar{u}$, $V^u(x_o,t_o)$
Go to 2.

If $(\bar{x},\bar{u})$ obtained from the above algorithm satisfies the condition of Proposition 3.1, then the policy k produced by the algorithm is the same as policy $\bar{k}$ of the Proposition. Hence, from Corollary (3.2), the policy k is locally optimal.

A Computational trick that Improves the Convergence Rate

In the algorithm the new control used is $u(t)=u^*+K\delta x$. It can happen that $K(t)\delta x$ becomes too large and invalidates the local expansion in δu. However, δx may still be small enough for

$$V_x(\bar{x}+\delta x,t) = V_x + V_{xx}\delta x.$$

Instead of storing $u^*(t)$ and $K(t)$ and computing $u(t)$, store V_x and V_{xx} - calculated from equations (2.61), (2.62).

Compute $u(t)$ directly by minimizing, w.r.t. u

$$H(\bar{x}+\delta x,u,V_x^u(\bar{x},t) + V_{xx}^u(\bar{x},t)\delta x,t) \qquad (5.53)$$

The radius of convergence of the algorithm may be increased.

Then Step 3 of S.S.A.M. must be modified in part

Define $u(t) = \bar{u}(t) \qquad t \in [1,N_1]$
$= $ the u(t) which minimizes (5.53), $t \in [N_1,N]$.

CHAPTER 6

TERMINAL CONSTRAINTS

I. TERMINAL EQUALITY CONSTRAINTS WITH FREE TERMINAL TIME

In this section we consider the class of problems where the endpoint of the trajectory is required to obey the following inequality

$$\psi(x(t_f),t_f) \equiv 0 \tag{6.1}$$

where ψ is an $s \leq n$ dimensional vector function and t_f is a free parameter that we may make use of to minimize the cost.

The cost function becomes

$$V^u(x(t_o),t_f:t_o) = \int_{t_o}^{t_f} L(x(\tau),u(\tau),\tau)d\tau + F(x(t_f),t_f) \tag{6.2}$$

where $$\dot{x}(t) = f(x(t),u(t),t); \qquad x(t_o) = x_o \tag{6.3}$$

and $\psi(x(t_f),t_f) = 0$.

L, F, f are defined as previously.

We treat the terminal constraints by a formal Lagrange Multipliers technique as follows: adjoin the constraints (6.1) to the cost function using a Lagrange multiplier $b \in R^s$, so

$$V^u(x(t),b,t_f;t) = \int_t^{t_f} L(x(\tau),u(\tau),\tau)d\tau + F(x(t_f),t_f)$$

$$+ \langle b,\psi(x(t_f),t_f)\rangle \qquad (6.4)$$

Assume we have some (nominal) control $\bar{u}(t), t \in \bar{T}$ with resultant trajectory $\bar{x}(t), t \in \bar{T}$; (nominal) multipliers $\bar{b}$ and (nominal) final time $\bar{t}_f$. $\bar{T} \triangleq [t_o, \bar{t}_f]$

Now suppose $k(x,b,t_f;t)$ is a policy which generates a new control $u(t)$, $t \in T$, with associated trajectory $x(t)$, $t \in T$, new multipliers b and new final time t_f. $T \triangleq [t_o, t_f]$. Let

$$k(x,b,t_f;t) = k(\bar{x}+\delta x, \bar{b}+\delta b, \bar{t}_f+\delta t_f;t)$$

$$= k(\bar{x},\bar{b},\bar{t}_f;t) + \langle k_x,\delta x\rangle + \langle k_b,\delta b\rangle + k_{t_f}\delta t_f$$

$$+ \langle \delta x, k_{xb}\delta b\rangle + \langle k_{xt_f},\delta x\rangle \delta t_f + \langle k_{bt_f},\delta b\rangle \delta t_f$$

$$+ \tfrac{1}{2}\langle \delta x, k_{xx}\delta x\rangle + \tfrac{1}{2}\langle \delta b, k_{bb}\delta b\rangle + \tfrac{1}{2}k_{t_f t_f}\delta t_f^2 \qquad (6.5)$$

to second order, assuming δx, δb, δt_f are small enough. All the above quantities are evaluated at $\bar{x},\bar{b},\bar{t}_f;t$. Define

$$a(\bar{x},\bar{b},\bar{t}_f;t) = V^k(\bar{x},\bar{b},\bar{t}_f;t) - V^{\bar{u}}(\bar{x},\bar{b},\bar{t}_f;t) \qquad (6.6)$$

$$(V^{\bar{u}+a})^{\cdot} = \frac{\partial}{\partial t}(V^{\bar{u}+a}) + \langle V^k_{\bar{x}}, f(\bar{x},\bar{u},t) \tag{6.7}$$

$$\dot{V}^k_x = \frac{\partial}{\partial t}V^k_x + \langle V^k_{xx}, f(\bar{x},\bar{u},t)\rangle \tag{6.8}$$

$$\dot{V}^k_b = \frac{\partial}{\partial t}V^k_b + \langle V^k_{xb}, f(\bar{x},\bar{u},t)\rangle \tag{6.9}$$

$$\dot{V}^k_{t_f} = \frac{\partial}{\partial t}V^k_{t_f} + \langle V^k_{xt_f}, f(\bar{x},\bar{u},t)\rangle \tag{6.10}$$

$$\dot{V}^k_{xb} = \frac{\partial}{\partial t}V^k_{xb} + \langle V^k_{xxb}, f(\bar{x},\bar{u},t)\rangle \tag{6.11}$$

$$\dot{V}^k_{xt_f} = \frac{\partial}{\partial t}V^k_{xt_f} + \langle V^k_{xxt_f}, f(\bar{x},\bar{u},t)\rangle \tag{6.12}$$

$$\dot{V}^k_{bt_f} = \frac{\partial}{\partial t}V^k_{bt_f} + \langle V^k_{xbt_f}, f(\bar{x},\bar{u},t)\rangle \tag{6.13}$$

$$\dot{V}^k_{xx} = \frac{\partial}{\partial t}V^k_{xx} + \langle V^k_{xxx}, f(\bar{x},\bar{u},t)\rangle \tag{6.14}$$

$$\dot{V}^k_{bb} = \frac{\partial}{\partial t}V^k_{bb} + \langle V^k_{xbb}, f(\bar{x},\bar{u},t)\rangle \tag{6.15}$$

$$\dot{V}^k_{t_f t_f} = \frac{\partial}{\partial t}V^k_{t_f t_f} + \langle V^k_{xt_f t_f}, f(\bar{x},\bar{u},t)\rangle \tag{6.16}$$

The V^k are all evaluated at $\bar{x},\bar{b},\bar{t}_f;t$.

PROPOSITION 6.1 Let H1A, H2A, H3 be satisfied, s=3. In addition let F, ψ and their derivatives up to order 3 be continuous in (x,t). Let $\bar{u},u \in G$. Then providing the δx, δb and δt_f are small enough for the expansions in the

proof to be valid, we have:

$$-\dot{a}(\bar{x},\bar{b},\bar{t}_f,t) = H(\bar{x},u^*,\lambda(t),t) - H(\bar{x},\bar{u},\lambda(t),t) \tag{6.17}$$

$$-\dot{\lambda}(t) = H_x(\bar{x},u^*,\lambda(t),t)+P(t)\Delta f(t)+k_x^T(\bar{x},\bar{b},\bar{t}_f,t)H_u(\bar{x},u^*,\lambda(t),t) \tag{6.18}$$

$$-\dot{v}_b^k(\bar{x},\bar{b},\bar{t}_f,t) = [V_{xb}^k(\bar{x},\bar{b},\bar{t}_f,t)]^T\Delta f(t)+k_b^T(\bar{x},\bar{b},\bar{t}_f,t)H_u(\bar{x},u^*,\lambda(t),t) \tag{6.19}$$

$$-\dot{v}_{t_f}^k(\bar{x},\bar{b},\bar{t}_f,t) = [V_{xt_f}^k(\bar{x},\bar{b},\bar{t}_f,t)]^T\Delta f(t)+k_{t_f}^T(\bar{x},\bar{b},\bar{t}_f,t)H_u(\bar{x},u^*,\lambda(t),t) \tag{6.20}$$

$$\begin{aligned}-\dot{V}_{xb}(\bar{x},\bar{b},\bar{t}_f,t) = {} & [f_x(\bar{x},u^*,t)+f_u(\bar{x},u^*,t)k_x(\bar{x},\bar{b},\bar{t}_f,t)]^T V_{xb}(\bar{x},\bar{b},\bar{t}_f,t)\\ & +[H_{ux}(\bar{x},u^*,\lambda(t),t)+H_{uu}(\bar{x},u^*,\lambda(t),t)k_x(\bar{x},\bar{b},\bar{t}_f,t)\\ & + f_u^T(\bar{x},u^*,t)P(t)]^T k_b(\bar{x},\bar{b},\bar{t}_f,t)\\ & + H_u(\bar{x},u^*,\lambda(t),t)k_{xb}(\bar{x},\bar{b},\bar{t}_f,t)\end{aligned} \tag{6.21}$$

$$\begin{aligned}-\dot{V}_{xt_f}^k(\bar{x},\bar{b},\bar{t}_f,t) = {} & [f_x(\bar{x},u^*,t)+f_u(\bar{x},u^*,t)k_x(\bar{x},\bar{b},\bar{t}_f,t)]^T V_{xt_f}^k(\bar{x},\bar{b},\bar{t}_f,t)\\ & +[H_{ux}(\bar{x},u^*,\lambda(t),t)+H_{uu}(\bar{x},u^*,\lambda(t),t)k_x(\bar{x},\bar{b},\bar{t}_f,t)\\ & +f_u^T(\bar{x},u^*,t)P(t)]^T k_{t_f}(\bar{x},\bar{b},\bar{t}_f,t)\\ & +H_u(\bar{x},u^*,\lambda(t),t)\ k_{xt_f}(\bar{x},\bar{b},\bar{t}_f,t)\end{aligned} \tag{6.22}$$

$$-\dot{V}^{k}_{bt_f}(\bar{x},\bar{b},\bar{t}_f;t) = [V^{k}_{xb}(\bar{x},\bar{b},\bar{t}_f;t)]^T f_u(\bar{x},u^*,t) k_{t_f}(\bar{x},\bar{b},\bar{t}_f;t)$$
$$+[H_{uu}(\bar{x},u^*,\lambda(t),t) k_b(\bar{x},\bar{b},\bar{t}_f;t)$$
$$+f^T_u(\bar{x},u^*,t) V^{k}_{xb}(\bar{x},\bar{b},\bar{t}_f;t)]^T k_{t_f}(\bar{x},\bar{b},\bar{t}_f;t)$$
$$+H_u(\bar{x},u^*,\lambda(t),t) k_{bt_f}(\bar{x},\bar{b},\bar{t}_f;t) \qquad (6.23)$$

$$-\dot{P}(t) = H_{xx}(\bar{x},u^*,\lambda(t),t)+f^T_x(\bar{x},u^*,t)P(t)+P(t)f(\bar{x},u^*,t)$$
$$+[H_{ux}(\bar{x},u^*,\lambda(t),t)+f^T_u(\bar{x},u^*,t)P(t)]^T k_x(\bar{x},\bar{b},\bar{t}_f;t)$$
$$+[k_x(\bar{x},\bar{b},\bar{t}_f;t)]^T[H_{ux}(\bar{x},u^*,\lambda(t),t)+f^T_u(\bar{x},u^*,t)P(t)]$$
$$+[k_x(\bar{x},\bar{b},\bar{t}_f;t)]^T H_{uu}(\bar{x},u^*,\lambda(t),t) k_x(\bar{x},\bar{b},\bar{t}_f;t)$$
$$+H_u(\bar{x},u^*,\lambda(t),t) k_{xx}(\bar{x},\bar{b},\bar{t}_f;t) \qquad (6.24)$$

$$-\dot{V}^{k}_{bb}(\bar{x},\bar{b},\bar{t}_f;t) = [k_b(\bar{x},\bar{b},\bar{t}_f;t)]^T f^T_u(\bar{x},u^*,t) V^{k}_{xb}(\bar{x},\bar{b},\bar{t}_f;t)$$
$$+[H_{uu}(\bar{x},u^*,\lambda(t),t) k_b(\bar{x},\bar{b},\bar{t}_f;t)$$
$$+f^T_u(\bar{x},u^*t) V^{k}_{xb}(\bar{x},\bar{b},\bar{t}_f;t)]^T k_b(\bar{x},\bar{b},\bar{t}_f;t)$$
$$+H_u(\bar{x},u^*(t),\lambda(t),t) k_{bb}(\bar{x},\bar{b},\bar{t}_f;t) \qquad (6.25)$$

$$-\dot{V}^k_{t_f t_f}(\bar{x},\bar{b},\bar{t}_f,t) = [k_{t_f}(\bar{x},\bar{b},\bar{t}_f,t)]^T f_u^T(\bar{x},u^*,t)V^k_{xt_f}(\bar{x},\bar{b},\bar{t}_f,t)$$
$$+[H_{u\lambda}(\bar{x},u^*,\lambda(t),t)k_{t_f}(\bar{x},\bar{b},\bar{t}_f,t)$$
$$+f_u^T(\bar{x},u^*,t)V^k_{xt_f}(\bar{x},\bar{b},\bar{t}_f,t)]^T k_{t_f}(\bar{x},\bar{b},\bar{t}_f,t)$$
$$+H_u(\bar{x},u^*,\lambda(t),t)k_{t_f t_f}(\bar{x},\bar{b},\bar{t}_f,t) \tag{6.26}$$

with boundary conditions:

$$a(\bar{x},\bar{b},\bar{t}_f,\bar{t}_f) = 0 \tag{6.27}$$

$$\lambda(\bar{t}_f) = F_x(\bar{x}(\bar{t}_f),\bar{t}_f) + \psi_x^T(\bar{x}(\bar{t}_f),\bar{t}_f)\bar{b} \tag{6.28}$$

$$V^k_b(\bar{x},\bar{b},\bar{t}_f,\bar{t}_f) = \psi(\bar{x}(\bar{t}_f),\bar{t}_f) \tag{6.29}$$

$$V^k_{t_f}(\bar{x},\bar{b},\bar{t}_f,\bar{t}_f) = H(\bar{x},u^*,\lambda(\bar{t}_f),\bar{t}_f)+F_t(\bar{x}(\bar{t}_f),\bar{t}_f)+\psi_t^T(\bar{x}(\bar{t}_f),\bar{t}_f)\bar{b} \tag{6.30}$$

$$V^k_{xb}(\bar{x},\bar{b},\bar{t}_f,\bar{t}_f) = \psi_x(\bar{x}(\bar{t}_f),\bar{t}_f) \tag{6.31}$$

$$V^k_{xt_f}(\bar{x},\bar{b},\bar{t}_f,\bar{t}_f) = F_{xt}(\bar{x}(\bar{t}_f),\bar{t}_f) + \psi_{xt}^T(\bar{x}(\bar{t}_f),\bar{t}_f)\bar{b} + H_x(\bar{x},u^*,\lambda(\bar{t}_f),\bar{t}_f)$$
$$+ P(\bar{t}_f)f(\bar{x},u^*,t_f)+k_x^T(\bar{x},\bar{b},\bar{t}_f,\bar{t}_f)H_u(\bar{x},u^*,\lambda(\bar{t}_f),\bar{t}_f) \tag{6.32}$$

$$V^k_{bt_f}(\bar{x},\bar{b},\bar{t}_f,\bar{t}_f) = \psi_t(\bar{x}(\bar{t}_f),\bar{t}_f)+\psi_x(\bar{x}(\bar{t}_f),\bar{t}_f)f(\bar{x},u^*,t_f)$$
$$+ k_b^T(\bar{x},\bar{b},\bar{t}_f,\bar{t}_f)H_u(\bar{x},u^*,\lambda(\bar{t}_f),\bar{t}_f) \tag{6.33}$$

$$P(\bar{t}_f) = F_{xx}(\bar{x}(\bar{t}_f),\bar{t}_f) + \bar{b}\psi_{xx}(\bar{x}(\bar{t}_f),\bar{t}_f) \tag{6.34}$$

$$V^k_{bb}(\bar{x},\bar{b},\bar{t}_f,\bar{t}_f) = 0 \tag{6.35}$$

$$\begin{aligned} V^k_{\bar{t}_f\bar{t}_f}(\bar{x},\bar{b},\bar{t}_f,\bar{t}_f) = {} & H_t(\bar{x},u^*,\lambda(\bar{t}_f),\bar{t}_f) + F_{\bar{t}\bar{t}}(\bar{x}(\bar{t}_f),\bar{t}_f) + \bar{b}\psi_{\bar{t}\bar{t}}(\bar{x}(\bar{t}_f),\bar{t}_f) \\ & + \langle H_u(\bar{x},u^*,\lambda(t),t),\dot{u}^*(t)\rangle + \langle f(\bar{x},u^*,\bar{t}_f), P(\bar{t}_f) f(\bar{x},u^*,\bar{t}_f)\rangle \\ & + 2\langle F_{x\bar{t}}(\bar{x}(\bar{t}_f)\bar{t}_f) + \psi^T_{x\bar{t}}(\bar{x}(\bar{t}_f),\bar{t}_f)\bar{b}, f(\bar{x},u^*,\bar{t}_f)\rangle \\ & + 2\langle k_{\bar{t}_f}(\bar{x},\bar{b},\bar{t}_f,\bar{t}_f), H_u(\bar{x},u^*,\lambda(\bar{t}_f),\bar{t}_f)\rangle \\ & + \langle H_x(\bar{x},u^*,\lambda(\bar{t}_f),\bar{t}_f), f(\bar{x},\bar{u},\bar{t}_f)\rangle \end{aligned} \tag{6.36}$$

where

$$\lambda(t) \triangleq V^k_x(\bar{x},\bar{b},\bar{t}_f,t) \tag{6.37}$$

$$P(t) \triangleq V^k_{xx}(\bar{x},\bar{b},t_f,t) \tag{6.38}$$

$$u^*(t) \triangleq k(\bar{x},\bar{b},\bar{t}_f,t) \tag{6.39}$$

and

$$\Delta f(t) \triangleq f(\bar{x},u^*,t) - f(\bar{x},\bar{u},t). \tag{6.40}$$

Proof: We have from (2.12)

$$-\frac{\partial}{\partial t}V^k(x,b,t_f,t) = H(x,k(x,b,t_f,t),V^k_x(x,b,t_f,t),t) \tag{6.41}$$

Expand $V^k(x,b,t_f,t)$ about $\bar{x},\bar{b},\bar{t}_f$ to second order in δx, δb and δt_f:

$$V^k(\bar{x}+\delta x,\bar{b}+\delta b,\bar{t}_f+\delta t_f,t)$$

$$= V^k + \langle V^k_x,\delta x\rangle + \langle V^k_b,\delta b\rangle + V^k_{t_f}\delta t_f$$

$$+ \langle \delta x, V^k_{xb}\delta b\rangle + \langle V^k_{xt_f},\delta x\rangle \delta t_f + \langle V^k_{bt_f},\delta b\rangle \delta t_f$$

$$+ \tfrac{1}{2}\langle \delta x, V^k_{xx}\delta x\rangle + \tfrac{1}{2}\langle \delta b, V^k_{bb}\delta b\rangle + \tfrac{1}{2}V^k_{t_f t_f}\delta t_f^2 \qquad (6.42)$$

Thus:

$$V^k_x(\bar{x}+\delta x,\bar{b}+\delta b,\bar{t}_f+\delta t_f;t) = V^k_x + V^k_{xx}\delta x + V^k_{xb}\delta b + V^k_{xt_f}\delta t_f \qquad (6.43)$$

All the above quantities are evaluated at $\bar{x},\bar{b},\bar{t}f,t$.

We ignore the terms $V_{xxb}\delta x\delta b$, $V_{xbb}\delta b\delta b$, $V_{xx}\delta x\delta t_f$, $V_{xbt_f}\delta b\delta t_f$, $V_{xxx}\delta x\delta x$, $V_{xt_f t_f}\delta t_f^2$ to avoid carrying cumbersome terms through the analysis and then neglecting them for the same reasons given in Chapter 2 Section V.

Using expansions (6.42) (6.43), and policy $k(x,b,t_f,t)$ defined in equ (6.5), in equation (6.41) gives:

$$-\frac{\partial}{\partial t}\hat{V}^k - \langle \frac{\partial}{\partial t}\hat{V}^k_x,\delta x\rangle - \langle \frac{\partial}{\partial t}\hat{V}^k_b,\delta b\rangle - \frac{\partial}{\partial t}\hat{V}^k_{t_f}\delta t_f - \langle \delta x,\frac{\partial}{\partial t}\hat{V}^k_{xb}\delta b\rangle$$

$$-\langle \frac{\partial}{\partial t}V^k_{xt_f},\delta x\rangle\delta t_f - \langle \frac{\partial}{\partial t}V^k_{bt_f},\delta b\rangle\delta t_f - \tfrac{1}{2}\langle \delta x,\frac{\partial}{\partial t}V^k_{xx}\delta x\rangle$$

$$-\tfrac{1}{2}\langle \delta b,\frac{\partial}{\partial t}V^k_{bb}\delta b\rangle - \tfrac{1}{2}\frac{\partial}{\partial t}V^k_{t_f t_f}\delta t_f^2 + \text{higher order terms}$$

$$= H(\bar{x}+\delta x,k(\bar{x}+\delta x,\bar{b}+\delta b,\bar{t}_f+\delta t_f,t),V^k_x,t)$$

$$+\langle V^k_{xx}\delta x + V^k_{xb}\delta b + V^k_{xt_f}\delta t_f + \text{H.O. terms},\ f(\bar{x}+\delta x,k(\bar{x}+\delta x,\bar{b}+\delta b,\bar{t}_f+\delta t_f,t),t)\rangle \qquad (6.44)$$

Now expand the R.H.S. of equ (6.44) to second order in δx, δb, δt_f (using the expansion for k given in equ (6.5)) and then, as in Proposition 2.5 equate the coefficients of δx, δb, δt_f. Then using equs (6.7) - (6.16) (ignoring the quantities mentioned earlier) gives equations (6.17) - (6.26).

We now determine the boundary conditions. We have:

$$V^k(x,b,t_f,t) = \int_t^{t_f} L(x,k,\tau)d\tau + F(x(t_f),t_f) + \langle b,\psi(x(t_f),t_f)\rangle$$

Writing this in terms of the nominal values:

$$V^k(\bar{x}+\delta x,\bar{b}+\delta b,\bar{t}_f+\delta t_f;t)$$

$$= \int_t^{\bar{t}_f+\delta t_f} L(x(\tau),u(\tau),\tau)d\tau + F(\bar{x}(\bar{t}_f+\delta t_f)+\delta x(\bar{t}_f+\delta t_f);\bar{t}_f+\delta t_f)$$

$$+\langle\bar{b}+\delta b,\psi(\bar{x}(\bar{t}_f+\delta t_f)+\delta x(\bar{t}_f+\delta t_f);\bar{t}_f+\delta t_f)\rangle$$

$$= \int_t^{\bar{t}_f+\delta t_f} L(x(\tau),u(\tau),\tau)d\tau + F(\bar{x}(t)+\delta x(t)+\Delta x(\bar{t}_f+\delta t_f);\bar{t}_f+\delta t_f)$$

$$+\langle\bar{b}+\delta b,\psi(\bar{x}(t)+\delta x(t)+\Delta x(\bar{t}_f+\delta t_f);\bar{t}_f+\delta t_f)\rangle \qquad (6.45)$$

where

$$\Delta x(\bar{t}_f+\delta t_f) = \int_t^{\bar{t}_f+\delta t_f} f(\bar{x}(\tau)+\delta x(\tau),k(\tau),\tau)d\tau \qquad (6.46)$$

Consider the integral

$$\int_{t}^{\bar{t}_f+\delta t_f} L(x(\tau),k(\tau),\tau)d\tau$$

$$= L(\bar{x}(\bar{t}_f)+\delta x(\bar{t}_f),k(\bar{x}+\delta x,\bar{b}+\delta b,\bar{t}_f+\delta t_f,\bar{t}_f),\bar{t}_f)\delta t_f$$

$$+ \tfrac{1}{2}\frac{d}{dt}L(\bar{x}+\delta x,k(\bar{x}+\delta x,\bar{b}+\delta b,\bar{t}_f+\delta t_f,\bar{t}_f),t_f)\delta t_f^2$$

$$= [L + \langle L_x,\delta x\rangle + \langle L_u,(k_x\delta x+k_b\delta b+k_{t_f}\delta t_f)\rangle]\delta t_f$$

$$+ \tfrac{1}{2}[L_t + \langle L_x,\dot{x}(\bar{x},\bar{u},\bar{t}_f)\rangle + \langle L_u,\dot{u}^*\rangle]\delta t_f^2 \qquad (6.47)$$

Consider the second term on the R.H.S. of (6.45)

$$F(\bar{x}(\bar{t}_f)+\delta x(\bar{t}_f)+\Delta x(\bar{t}_f+\delta t_f);\bar{t}_f+\delta t_f)$$

$$= F + \langle F_x,(\delta x(\bar{t}_f)+\Delta x(\bar{t}_f+\delta t_f))\rangle + F_t\delta t_f$$

$$+ \langle F_{xt},(\delta x(\bar{t}_f)+\Delta x(\bar{t}_f+\delta t_f))\rangle\delta t_f + \tfrac{1}{2}F_{tt}\delta t_f^2$$

$$+ \tfrac{1}{2}\langle\delta x(\bar{t}_f)+\Delta x(\bar{t}_f+\delta t_f),F_{xx}(\delta x(\bar{t}_f)+\Delta x(\bar{t}_f+\delta t_f))\rangle$$

$$= F + \langle F_x,\delta x + f\delta t_f + f_x\delta x\delta t_f + f_uk_x\delta x\delta t_f + f_uk_b\delta b\delta t_f$$

$$+ f_uk_{t_f}\delta t_f^2 + \tfrac{1}{2}f_{t_f}\delta t_f^2 + \tfrac{1}{2}f_x f(\bar{x},\bar{u},\bar{t}_f)\delta t_f^2 + \tfrac{1}{2}f_u\dot{u}^*\delta t_f^2\rangle$$

$$+ F_t\delta t_f + \langle F_{xt},\delta x + f\delta t_f\rangle\delta t_f + \tfrac{1}{2}\langle\delta x,F_{xx}\delta x\rangle$$

$$+ \langle\delta x,F_{xx}f\rangle\delta t_f + \tfrac{1}{2}\langle f,F_{xx}f\rangle\delta t_f^2 + \tfrac{1}{2}F_{tt}\delta t_f^2 \qquad (6.48)$$

because, from (6.46) using (6.47)

$$\Delta x(\bar{t}_f+\Delta t_f)$$

$$= [f+f_x\delta x+f_u(k_x\delta x+k_b\delta b+k_{t_f}\delta t_f)]\delta t_f$$

$$+ \tfrac{1}{2}[f_t+f_x f(\bar{x},\bar{u},\bar{t}_f)+f_u\dot{u}^*]\delta t_f^2.$$

Then, similarly, we have for the third term on the R.H.S. of equation (6.45)

$$\langle \bar{b}+\delta b,\psi(\bar{x}(\bar{t}_f)+\delta x(\bar{t}_f)+\Delta x(\bar{t}_f+\delta t_f);\bar{t}_f+\delta t_f)\rangle$$

$$= \langle \bar{b},\psi+\psi_x\,\delta x+f\delta t_f+f_x\delta x\delta t_f+f_u k_x\delta x\delta t_f$$

$$+ f_u k_b\delta b\delta t_f+f_u k_{t_f}\delta t_f^2+\tfrac{1}{2}f_t\delta t_f^2+\tfrac{1}{2}f_x f(\bar{x},\bar{u},\bar{t}_f)\delta t_f^2$$

$$+ \tfrac{1}{2}f_u\dot{u}^*\delta t_f^2]+\psi_t\delta t_f+\psi_{xt}[\delta x+f\delta t_f]\delta t_f+\tfrac{1}{2}\psi_{xx}\delta x\delta x$$

$$+ \psi_{xx}f\delta x\delta t_f+\tfrac{1}{2}\psi_{xx}ff\delta t_f^2+\tfrac{1}{2}\psi_{tt}\delta t_f^2\rangle$$

$$+ \langle \delta b,\psi+\psi_x[\delta x+f\delta t_f]+\psi_t\delta t_f\rangle \qquad (6.49)$$

to second order.

All the quantities above are evaluated at $\bar{x},\bar{b},\bar{t}_f;\bar{t}_f$ and u^* (defined in equ (6.39)), unless specified otherwise.

Now the L.H.S. of equ (6.46) at $t = t_f$, expanded to second order is given by equ (6.42) with $t = t_f$.

Equating co-efficients of like powers, gives at $t = \bar{t}_f$,

$$V_x^k = F_x + \psi_x^{\pi}\bar{b} \qquad (6.50)$$

$$V_b^k = \psi \tag{6.51}$$

$$V_{t_f}^k = L + \langle F_x + \psi_x^T \bar{b}, f \rangle + F_t + \langle \bar{b}, \psi_t \rangle \tag{6.52}$$

$$V_{xb}^k = \psi_x^T \tag{6.53}$$

$$V_{xt_f}^k = F_{xt} + \psi_{xt}^T \bar{b} + L_x + f_x^T (F_x + \psi_x^T \bar{b}) + k_x^T [L_u + f_u^T (F_x + \psi_x^T \bar{b})]$$
$$+ (F_{xx} + \bar{b}\psi_{xx}) f \tag{6.54}$$

$$V_{bt_f}^k = \psi_t + \psi_x f + k_b^T [L_u + f_u^T (F_x + \psi_x^T \bar{b})] \tag{6.55}$$

$$V_{xx}^k = F_{xx} + \bar{b}\psi_{xx} \tag{6.56}$$

$$V_{bb}^k = 0 \tag{6.57}$$

$$V_{t_f t_f}^k = L_t + \langle F_x + \psi_x^T \bar{b}, f_t \rangle + \langle L_u + f_u^T (F_x + \psi_x^T \bar{b}), \hat{u}^* \rangle$$
$$+ \langle L_x + f_x^T (F_x + \psi_x^T \bar{b}), f(\bar{x}, \bar{u}, \bar{t}_f) \rangle + 2\langle F_{xt} + \psi_{xt}^T \bar{b}, f \rangle$$
$$+ 2k_{t_f} [L_u + f_u^T (F_x + \psi_x^T \bar{b})] + F_{tt} + \langle \bar{b}, \psi_{tt} \rangle$$
$$+ \langle f, (F_{xx} + \bar{b}\psi_{xx}) f \rangle \tag{6.58}$$

Now since

$$V_x^k = F_x + \psi_x^T \bar{b}$$

$$V_{xx}^k = F_{xx} + \bar{b}\psi_{xx}$$

we have

$$H(\bar{x}, u^*, \lambda(\bar{t}_f), \bar{t}_f) = L + \langle F_x + \psi_x^T \bar{b}, f \rangle \tag{6.59}$$

All the above quantities are evaluated at $\bar{x}, \bar{b}, \bar{\tau}_f; \bar{v}_f$ and u^*, unless otherwise specified.

Substituting equ (6.59) into (6.50) - (6.58) gives boundary conditions (6.27) - (6.36). □

II SECOND ORDER ALGORITHM FOR TERMINAL EQUALITY CONSTRAINTS WITH FREE TERMINAL TIME

The problem studied by Jacobson is to find a $t_f, u(t), t \in T \triangleq [t_o, t_f]$ to minimize (6.4) (we ignore the possibility that the cost function defined in (6.4) does not have a minimum but only a stationary point w.r.t u) and to choose b s.t. equ (6.1) is satisfied. It turns out, see [3], that V must be maximized w.r.t. b.

Let $u^*(t) \triangleq k(\bar{x}, \bar{b}, \bar{t}_f; t)$ be the argument which minimizes $H(\bar{x}, u, \lambda(t), t)$ over $u \in G$. k_x, k_b, k_{t_f} are chosen to minimize $P(t)$, V^k_{bb}, $V^k_{t_f t_f}$ respectively, with

$$k_{xb} \equiv k_{bt_f} \equiv k_{xt_f} \equiv k_{xx} \equiv k_{bb} \equiv k_{t_f t_f} \equiv 0 \tag{6.60}$$

This gives:

$$k_x(\bar{x}, \bar{b}, \bar{t}_f; t) = -H_{uu}^{-1}[H_{ux} + f_u^T P(t)] \tag{6.61}$$

$$k_b(\bar{x}, \bar{b}, \bar{t}_f; t) = -H_{uu}^{-1} f_u^T V^k_{xb} \tag{6.62}$$

$$k_{t_f}(\bar{x}, \bar{b}, \bar{t}_f; t) = -H_{uu}^{-1} f_u^m V^k_{xt_f} \tag{6.63}$$

where it is assumed $H_{uu}(\bar{x}, u^*, \lambda(t), t) > 0 \quad \forall t \in [t_o, \bar{t}_f]$. Thus our new control is given by

$$u(t) = u^*(t) + k_x \delta x + k_b \delta b + k_{t_f} \delta t_f \tag{6.64}$$

All the above quantities are evaluated at $\bar{x},\bar{b},\bar{t}_f$ and u^*.

If equs (6.60) - (6.63) are substituted into (6.17) - (6.36) we obtain the same differential equations as Jacobson, however, the boundary conditions differ. Jacobxon's boundary conditions appear to be in error.

We now need to find δb and δt_f.

Set $b = \bar{b}$ and $t = \bar{t}_f$ and solve the free endpoint problem using the second-order algorithm of 5-II-4. This causes $\Delta f(t)$ to be zero, and because $u^*(t)$ is chosen to minimize $H(\bar{x},u,\lambda(t),t)$ w.r.t. u, we have

$$H_u(\bar{x},u^*,\lambda(t),t) = 0 \tag{6.65}$$

Hence, from equs (6.19), (6.20) and (6.65):

$$V^o_b(\bar{x},\bar{b},\bar{t}_f;t) = 0 \tag{6.66}$$

$$V^o_{t_f}(\bar{x},\bar{b},\bar{t}_f;t) = 0 \tag{6.67}$$

From above, using equs (6.29), (6.30) and (6.66), (6.67):

$$V^o_b(x_o,\bar{b},\bar{t}_f;t_o) = \psi(\bar{x}(\bar{t}_f),\bar{t}_f) \tag{6.68}$$

$$V^o_{t_f}(x_o,\bar{b},\bar{t}_f;t_o) = H(\bar{x},u^*,\lambda(\bar{t}_f),\bar{t}_f)+F_{t_f}(\bar{x},u^*,t_f) + <\bar{b},\psi_{t_f}(\bar{x}(\bar{t}_f),\bar{t}_f)> \tag{6.69}$$

NOTE We are now using the superscript o to denote that $V^o(\bar{x},\bar{b},\bar{t}_f;t)$ is the optimal cost for the state $\bar{x}$, which

the optimal trajectory for $\bar{b}$ and $\bar{t}_f$. This arises because, for fixed $\bar{b}$ and $\bar{t}_f$, the second order D.D.P algorithm of 5-II-4 finds optimal $u^o(t)$, $x^o(t)$, which we now call our new $\bar{u}(t)$, $\bar{x}(t)$.

When we reintroduce variations δx, δb, δt_f, we require

$$\psi(\bar{x}(\bar{t}_f+\delta t_f)+\delta x(\bar{t}_f+\delta t_f),\bar{t}_f+\delta t_f) = 0$$

and for V^o to be stationary w.r.t. t_f

$$V^o_{t_f}(x_o,\bar{b}+\delta b,\bar{t}_f+\delta t_f;t_o) = 0$$

Now

$$\begin{aligned} V^o(x_o,\bar{b}+\delta b,\bar{t}_f+\delta t_f;t_o) \\ = V^o+\langle V^o_b,\delta b\rangle+V^o_{t_f}\delta t_f+\langle V^o_{bt_f},\delta b\rangle\delta t_f \\ + \tfrac{1}{2}\langle\delta b,V^o_{bb}\delta b\rangle + \tfrac{1}{2}V^o_{t_ft_f}\delta t_f^2 \end{aligned} \tag{6.70}$$

provided δb and δt_f are small enough.

Thus

$$V^o_b(x_o,\bar{b}+\delta b,\bar{t}_f+\delta t_f;t_o) = \psi(\bar{x}(\bar{t}_f+\delta t_f)+\delta x(\bar{t}_f+\delta t_f);\bar{t}_f+\delta t_f) = 0$$

$$= V^o_b + V^o_{bb}\delta b + V^o_{bt_f}\delta t_f$$

and

$$V^o_{t_f}(x_o,\bar{b}+\delta b,\bar{t}_f+\delta t_f;t_o) = 0$$

$$= V^o_{t_f} + V^o_{bt_f}\delta b + V^o_{t_ft_f}\delta t_f$$

i.e.
$$\begin{bmatrix}\delta b\\ \delta t_f\end{bmatrix} = -\begin{bmatrix}V^o_{bb} & V^o_{bt_f}\\ V^o_{t_fb} & V^o_{t_ft_f}\end{bmatrix}^{-1}\begin{bmatrix}V^o_b\\ V^o_{t_f}\end{bmatrix} \tag{6.71}$$

where V_b^o, $V_{t_f}^o$ are given by (6.68), (6.69).

All quantities above are evaluated at x_o, $\bar{b}$, $\bar{t}_f$; t_o. ε : $(0 \leq \varepsilon \leq 1)$ is present to ensure these changes are not too large.

Remark: The cases

(i) terminal equality constraints with fixed terminal time

(ii) free terminal time (i.e. no terminal constraints)

are just special cases of the analysis of the last two sections and the relevant differential equations and boundary conditions can be got quite easily from the equations of the two sections given above.

THE COMPUTATIONAL PROCEDURE

Step 0. Using a nominal final time $\bar{t}_f$ and nominal control $\bar{u}(t), t \in [t_o, \bar{t}_f]$ run a nominal trajectory. Using a nominal set of multipliers $\bar{b}$, calculate the nominal cost $V^{\bar{u}}(x_o, \bar{b}, \bar{t}_f; t_o)$ from equ (6.4).

Step 1. Using boundary conditions (6.27), (6.28), (6.34) integrate $\dot{a}$, $\dot{\lambda}$, $\dot{p}$ backwards in time from t_f to t_o all the while storing $u^*(t)$ and k_x (given by equ (6.61)).

If $|a(x_o, \bar{b}, \bar{t}_f; t_o)| < \eta$ go to 2.

Otherwise, use the computational procedure of 5-II-4 to reduce $|a(x_o, \bar{b}, \bar{t}_f; t_o)|$.

(i.e. treat the problem as a free endpoint problem by keeping $b = \bar{b}, t_f = \bar{t}_f$).
When $|a(x_o, \bar{b}, \bar{t}_f, t_o)| < \eta$ replace the old nominal trajectory by the new one and go to step 2.

Step 2. Using boundary conditions (6.31), (6.32), (6.33), (6.35), (6.36) integrate eqns (6.21), (6.22), (6.23), (6.25), (6.26) backwards along this trajectory all the while storing k_b, k_{t_f} (given in eqns (6.62), (6.63)) in the process. Use eqn (6.60).

Step 3. Calculate δb and δt_f from eqn (6.71).

Step 4. Set $\varepsilon = 1$

Step 5. Apply the control $u(t) = k_x(t)\delta x(t) + k_b \delta b + k_{t_f} \delta t_f$
and compute a new $x(t)$ trajectory and corresponding $\psi(x(t_f), t_f)$ $\quad t_f = \bar{t}_f + \delta t_f$

Step 6. If $|\psi(x(t_f), t_f)| < \eta_2$ compute optimal V. Stop.
Otherwise go to 7.

Step 7. If there has been an improvement in the endpoint error as measured by $|\psi(x(t_f), t_f)| - |\psi(\bar{x}(\bar{t}_f), \bar{t}_f)|$ then set $\bar{x}(t) = x(t)$

$$\bar{b}_{new} = \bar{b}_{old} + \delta b$$

$$\bar{t}_{f_{new}} = \bar{t}_{f_{old}} + \delta t_f. \quad \text{Go to 1.}$$

Otherwise, set $\varepsilon = \varepsilon/2$ and go to 5.

REFINEMENTS TO STEP 7 FOR FIXED TERMINAL TIME. [9].

TEST 1

Although δb is chosen according to (6.70) (where ε is s.t. $\psi(x(t_f), t_f)$ is reduced) it may lie outside the validity of the expansion (6.70) (where $\delta t_f = 0$).
We have

$$V^o(x_o, \bar{b}+\delta b, t_o)$$

$$= a^o(x_o, \bar{b}; t_o) + V^{\bar{u}}(x_o, \bar{b}; t_o) + [V^o_b(x_o, \bar{b}; t_o)]^T \delta b$$

$$+ \tfrac{1}{2}\delta b^T V^o_{bb}(x_o, \bar{b}, t_o)\,\delta b \qquad (6.72)$$

Equ (6.70) and (6.72) coincide at $t = t_o$.
Choose δb and evaluate the R.H.S. of (6.72). Integrate (6.3) as described previously and evaluate $V(x_o, \bar{b}+\delta b, t_o)$.
If δb is given by:

$$\delta b = -\varepsilon[V^o_{bb}]^{-1} V^o_b$$

$$= -\varepsilon[V^o_{bb}]^{-1} \psi(\bar{x}(t_f), t_f)$$

then

$$V^o(x_o, \bar{b}+\delta b; t_o) - V^{\bar{u}}(x_o, \bar{b}; t_o)$$

$$= a^o(x_o, \bar{b}; t_o) - (\varepsilon - \tfrac{1}{2}\varepsilon^2)[\psi(\bar{x}(t_f), t_f)]^T [V^o_{bb}(x_o, \bar{b}, t_o)]^{-1} \psi(\bar{x}(t_f), t_f) \qquad (6.73)$$

If equation (6.73) does not predict the change in V to within a given tolerance, ε should be reduced until it does.

TEST 2

This is useful if it is desirable to keep the new trajectory in the immediate neighbourhood of the nominal trajectory (e.g. we may wish to prevent the trajectory from jumping to another local minimum). We have

$$V^{o}(x,b;t) = \int_{t}^{t_f} L(x^{o},u^{o},\tau)\,d\tau + F(x^{o}(t_f),t_f) + b^{o}\psi(x^{o}(t_f),t_f)$$

and

$$V^{\bar{u}}(\bar{x},\bar{b};t) = \int_{t}^{t_f} L(\bar{x},\bar{u},\tau)\,d\tau + F(\bar{x}(t_f),t_f) + \bar{b}\psi(\bar{x}(t_f),t_f)$$

Then

$$\begin{aligned}\Delta V(t) &= V^{o}(x,b,t) - V^{\bar{u}}(\bar{x},b,t)\\ &= \int_{t}^{t_f} [L(x^{o},u^{o},\tau) - L(\bar{x},\bar{u},\tau)]\,d\tau + F(x^{o}(t_f),t_f) - F(\bar{x}(t_f),t_f)\\ &\quad + b^{o}\psi(x^{o}(t_f),t_f) - \bar{b}\psi(\bar{x}(t_f),t_f) \qquad (6.74)\end{aligned}$$

From equs (6.6) and (6.42), with $\delta t_f = 0$:

$$\begin{aligned}\Delta V(t) = a^{o}(\bar{x},\bar{b},t) + V_{x}^{o}(\bar{x},\bar{b},t)\,\delta x + V_{b}^{o}(\bar{x},\bar{b};t)\,\delta b + \tfrac{1}{2}\delta x^{T} V_{xx}^{o}(\bar{x},\bar{b},t)\,\delta x\\ + \delta x^{T} V_{xb}^{o}(\bar{x},\bar{b},t)\,\delta b + \tfrac{1}{2}\delta b^{T} V_{bb}^{o}(\bar{x},\bar{b},t)\,\delta b \qquad (6.75)\end{aligned}$$

Equations (6.74) and (6.75) are theoretically equal. Their difference is a test of the size of δx and δb. From (6.74):

$$\Delta V(t) - \Delta V(t_o) = \int_{t_o}^{t} (L(x^{o},u^{o},\tau) - L(\bar{x},\bar{u},\tau))\,d\tau \qquad (6.76)$$

and from (6.75):

$$\begin{aligned}\Delta V(t) - \Delta V(t_o) = {} & a^o(\bar{x},\bar{b},t) + V_x^o(\bar{x},\bar{b},t)\,\delta x + V_b^o(\bar{x},\bar{b},t)\,\delta b \\ & + \tfrac{1}{2}\delta x^T V_{xx}^o(\bar{x},\bar{b},t)\,\delta x + \delta x^T V_{xb}^o(\bar{x},\bar{b},t)\,\delta b \\ & + \tfrac{1}{2}\delta b^T V_{bb}^o(\bar{x},\bar{b},t)\,\delta b - [a^o(x_o,\bar{b},t_o) \\ & + V_b^o(x_o,\bar{b},t_o)\,\delta b + \tfrac{1}{2}\delta b^T V_{bb}^o(x_o,\bar{b},t_o)\,\delta b]\end{aligned}$$

This can be simplified using equ (6.66) and (6.68), i.e. $V_b^o(x_o,\bar{b},t_o) = V_b^o(\bar{x},\bar{b},t)$ for all $t \in T$. Thus

$$\begin{aligned}\Delta V(t) - \Delta V(t_o) = {} & a^o(\bar{x},\bar{b},t) - a^o(x_o,\bar{b},t_o) + V_x^o(\bar{x},\bar{b},t)\,\delta x \\ & + \tfrac{1}{2}\delta x^T V_{xx}^o(\bar{x},\bar{b},t)\,\delta x + \delta x^T V_{xb}^o(\bar{x},\bar{b},t)\,\delta b \\ & + \tfrac{1}{2}\delta b^T V_{bb}^o(\bar{x},\bar{b},t)\,\delta b - \tfrac{1}{2}\delta b^T V_{bb}^o(\bar{x},\bar{b},t)\,\delta b. \qquad (6.77)\end{aligned}$$

Test 2 is performed by determining whether equ (6.76) agrees with equ (6.77) within a given tolerance. If the test is failed at t_i then the integration of this 'trial trajectory' can be discontinued. δb should be reduced, or, if δb is zero, δx_i should be reduced by the step size adjustment method. This test is particularly simple to apply in cases where $L(x,u,t) \equiv 0$.

III TERMINAL INEQUALITY CONSTRAINTS: Mayne [36].

Consider, for simplicity, the control problem with inequality constraint:

$$a^T x(t_f) - b \leq 0 \tag{6.78}$$

Suppose the second order D.D.P. algorithm of 5-II-4 with the terminal value of λ modified from

$$F_x(x(t_f),t_f) \text{ to } F_x(x(t_f),t_f)+ca$$

for some $c > 0$, yields $(\bar{x},\bar{u})$ satisfying Proposition 3.1.

$\bar{u}$ is the locally optimal control for a modified problem, without terminal constraints, for which the terminal cost is $\hat{F}(x(t_f),t_f)$ where:

$$\hat{F}(x(t),t) \triangleq F(x(t),t)+c(a^T x-b_1)$$

and b, is arbitrary. Let $b_1 = a^T\bar{x}(t_f)$. Then

$$F(x(t),t) = \hat{F}(x(t),t)-ca^T(x-\bar{x}(t_f))$$

Hence,

$F \geq \hat{F}$ for all x s.t.

$$ca^T(x-\bar{x}(t_f)) \leq 0 \tag{6.79}$$

Therefore, there exists an $\varepsilon > 0$ s.t.

$$V^u(x_o,t_o) \geq V^{\bar{u}}(x_o,t_o)$$

for all $u \in G$, s.t.

(i) $d(u,\bar{u}) \leq \varepsilon$

(ii) $x(t_f, x_o, t_o, u)$ satisfies (6.79)

If c can be changed to make

$$a^T \bar{x}(t_f) = b$$

then a locally optimal solution to the original problem will have been obtained.

Changing c to c + δc, $|\delta c| \leq \varepsilon$, will change $\bar{\lambda}(t)$ to $\bar{\lambda}(t) + \delta\lambda(t)$, changing $\bar{u}(t)$ to $\bar{u}(t) + \delta u(t)$, $\bar{x}(t)$ to $\bar{x}(t) + \delta x(t)$ etc.

Approximations $\delta\hat{\lambda}$, $\delta\hat{u}$, $\delta\hat{x}$ to these quantities can be obtained as follows: (see also Chapter 3-I-4).

At the end of the second order D.D.P. algorithm, we have $u^* = \bar{u}$. Therefore from equation

$$-\dot{\bar{\lambda}} = H_x(\bar{x},\bar{u},\bar{\lambda},t) + K(t)H_u(\bar{x},\bar{u},\bar{\lambda},t)$$

where

$$H_u(\bar{x},\bar{u},\bar{\lambda},t) = 0$$

and

$$\bar{\lambda}(t_f) = F_x(\bar{x}(t_f),t_f) + ca.$$

Changing c to c + δc introduces variations, giving

$$-\dot{\bar{\lambda}}-\delta\dot{\lambda} = H_x(\bar{x}+\delta x,\bar{u}+\delta u,\bar{\lambda}+\delta\lambda,t)+K^T(t)H_u(\bar{x}+\delta x,\bar{u}+\delta u,\bar{\lambda}+\delta\lambda,t)$$

$$= H_x(\bar{x}+\delta x,\bar{u}+\delta u,\bar{\lambda},t)+[f_x(\bar{x}+\delta x,\bar{u}+\delta u,t)]^T\delta\lambda$$

$$+K^T(t)\,(H_u(\bar{x}+\delta x,\bar{u}+\delta u,\bar{\lambda},t)+[f_u(\bar{x}+\delta x,\bar{u}+\delta u,t)]^T\delta\lambda)$$

$$\delta\lambda(t_f) = \delta ca$$

Expanding the R.H.S. of the above expression and ignoring all terms in δx, δu and higher order, we have, as an approximation to $\delta\lambda$

$$\dot{\delta\hat{\lambda}} = A_1^T(t)\,\delta\hat{\lambda} \tag{6.80}$$

with boundary condition

$$\delta\hat{\lambda}(t_f) = \delta c\ a. \tag{6.81}$$

where

$$A_1(t) \triangleq f_x(\bar{x}(t),\bar{u}(t),t) + f_u(\bar{x}(t),\bar{u}(t),t)K(t)$$

Let $\delta u = \arg\min_{\delta u} H(\bar{x}+\delta x, \bar{u}+\delta u, \bar{\lambda}+\delta\lambda, t)$

Then, expanding the R.H.S. of the above expression and (remembering that $H_u(\bar{x},\bar{u},\bar{\lambda},t) = 0$) ignoring any terms involving δx, $(\delta u)^3$, differentiating w.r.t. δu and setting equal to zero gives

$$\delta\hat{u} = -R^{-1}(t)B^T(t)\,\delta\hat{\lambda}(t)$$

where

$$R(t) \triangleq H_{uu}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)$$

$$B(t) \triangleq f_u(\bar{x}(t),\bar{u}(t),t).$$

So, approximation $\delta\hat{x}$ to δx is given by

$$\dot{\delta\hat{x}}(t) = A_1(t)\,\delta\hat{x}(t) + B^T(t)\,\delta\hat{u}(t)$$

$$\delta\hat{x}(t_o) = 0.$$

Then, (c.f. Chapter 3-I-4) we have

$$a^T\delta\hat{x}(t_f) = -a^T W(t_f,t_o)\,a\ \delta c$$

where

$$W(t_f,t_o) = \int_{t_o}^{t_f} \Phi(t_f,t)B(t)R^{-1}(t)B(t)\Phi^T(t_f,t)dt$$

and $\Phi(t_f,t)$ is the transition matrix corresponding to $A_1(t)$.

Noting that if $|\delta c| \leq \varepsilon$, then $\|\delta\hat{u}(t)\| \leq c_1\varepsilon$ for all $t \in T$, it can be shown that $\|\delta x(t_f)-\delta\hat{x}(t_f)\| \leq c_2\varepsilon^2$, $c_2 < \infty$. Hence equ (6.80) can be used, if $a^T w a \neq 0$, to choose a δc to reduce $a^T\bar{x}(t_f) - b$.

For each choice of c we obtain a locally optimal solution satisfying the constraint equation with b replaced by $a^T\bar{x}(t_f)$. c is changed, using equation (6.80), to make $a^T\bar{x}(t_f) = b$.

In practice one would compute:

$$a^T w(t_f,t_o)a = \int_{t_o}^{t_f} \delta\tilde{\lambda}^T(t)B(t)R^{-1}(t)B(t)\delta\tilde{\lambda}(t)dt$$

where

$$\delta\tilde{\lambda}(t) \triangleq \delta\lambda(t)/\delta c$$

$$= \Phi^T(t_f,t)a$$

is the solution of Equ (6.80) with terminal condition (Equ 6.81)) replaced by:

$$\delta\tilde{\lambda}(t_f) = a.$$

IV FIRST ORDER ALGORITHMS FOR TERMINAL INEQUALITY CONSTRAINTS: Polak and Mayne [46].

It is found that if the terminal inequality constraints are handled in a manner very similar to the way the Method of Feasible Directions handles control constraints, then the Algorithm described in Chapter 5-I-3 can be extended to handle terminal inequality constraints without much difficulty.

Let $\Omega \subset \{u \mid \|u\| \leq r\} \subset R^m$ be compact, and let $\tilde{G}$ be the space of equivalence classes of functions in G which are equal a.e.

Let $T \triangleq [0,1]$ for simplicity.

The cost functional is

$$g_o(u) \triangleq h_o(x^u(1))$$

The terminal constraints are

$$g_i(u) \triangleq h_i(x^u(1)) \leq 0 \qquad i=1, \ldots p$$

At each iteration we will only be concerned with the active terminal constraints i.e. those which are some ε away from the boundary. We define the set of active constraints as follows: Let $P = \{1, \ldots p\}$.

For any $\varepsilon \geq 0$, let

$$J_\varepsilon(g) \triangleq \{j \in P \mid g_j \geq -\varepsilon\}$$

To proceed further we make the following assumptions:

Let H1, H2 be satisfied, $s = 2$. In addition

(i) f_u, f_{ux}, L_u, L_{ux} exist and are continuous $\forall (x,u,t) \in S$. The functionals $h_j : R^n \to R$, $i=1,..p$ and their derivatives up to second order are continuous in R^n.

(ii) for all $\varepsilon \geq 0$,

$$\bar{U}_\varepsilon(\bar{u}) \triangleq \tilde{U}_\varepsilon(\bar{u},\cdot) \cap \tilde{G} \neq \phi \tag{6.82}$$

where we define

$$\tilde{U}_\varepsilon(\bar{u},t) \triangleq \arg\min_{w\in\Omega} \max\{H(\bar{x},w,\bar{\lambda}_o(t),t) - H(\bar{x},\bar{u},\bar{\lambda}_o(t),t);$$

$$g_j(\bar{u}) + H(\bar{x},w,\bar{\lambda}_j(t),t) - H(\bar{x},\bar{u},\bar{\lambda}_j(t),t), j\in P_\varepsilon(g)\}$$

and

$\bar{\lambda}_j$, $j=0,..$ p are the solutions to:

$$\dot{\bar{\lambda}}_j(t) = H_x(\bar{x},\bar{u},\bar{\lambda}_j(t),t) \tag{6.83}$$

$$\bar{\lambda}_j(1) = h_{jx}(\bar{x}(1)) \tag{6.84}$$

We have from Proposition 4.1 that for $u,\bar{u} \in G$

$$\Delta\hat{g}_j(u,\bar{u}) = \int_0^1 [H(\bar{x},u,\bar{\lambda}_j(t),t) - H(\bar{x},\bar{u},\bar{\lambda}_j(t),t)]dt$$

is an estimate for

$$\Delta g_j(u,\bar{u}) \triangleq g_j(u) - g_j(\bar{u})$$

such that

$$|\Delta g_j(u,\bar{u}) - \Delta\hat{g}_j(u,\bar{u})| \leq c[d(u,\bar{u})]^2 \qquad c < \infty$$

For $\bar{u} \in \tilde{G}$, define

$$\eta(z,u,g,t,\varepsilon) \triangleq \min_{w\in\Omega} \max\{H(\bar{x},w,\bar{\lambda}_o,t)-H(\bar{x},\bar{u},\bar{\lambda}_o,t);$$

$$g_j+H(\bar{x},w,\bar{\lambda}_j,t)-H(\bar{x},\bar{u},\bar{\lambda}_j,t),\ j\in J_\varepsilon(g)\}$$

where $z \in R^{n(2+p)}$ is a vector partitioned as follows

$$z \triangleq (x,\lambda_o,\lambda_1,\ldots\ \lambda_p)$$

Now, defining

$$\sigma_\varepsilon(\bar{u}) \triangleq \int_o^1 \eta(\bar{z},\bar{u},g(\bar{u}),t,\varepsilon)dt \qquad (6.85)$$

where $\bar{z}(t) \triangleq (\bar{x},\bar{\lambda}_o,\bar{\lambda}_1,\ldots\ \bar{\lambda}_p)$, our aim will be to determine an $\bar{u} \in G$ s.t. $\sigma_\varepsilon(\bar{u}) = 0$.

Step Size Rule

This is very similar to that given in Chapter 5-I-3. For $\bar{u} \in \tilde{G}$, $\varepsilon \geq 0$, define $I_\varepsilon^{\bar{u}} \in T$ by

$$I_\varepsilon^{\bar{u}} \triangleq \{t\in T \mid \eta(\bar{z},\bar{u},g(\bar{u}),t,\varepsilon) \leq \sigma_\varepsilon(u)\}$$

Let $m_\varepsilon(\bar{u}) \triangleq \mu(I_\varepsilon^{\bar{u}})$

Now for $\bar{u} \in \tilde{G}$, $c > 0, \alpha \in [0,1]$, let $I_\varepsilon^{\alpha\bar{u}}$ be the subset of T having the same properties as (i) - (v) of Chapter 5-I-3 (with the necessary notational adjustments).

For $\bar{u} \in \tilde{G}$, $\alpha \in [0,1]$, we define the new control by

$$u_\varepsilon^\alpha(t) \in \bar{U}_\varepsilon(\bar{u},t) \quad \forall t \in I_\varepsilon^{\alpha\bar{u}} \qquad (6.86)$$

$$u_\varepsilon^\alpha(t) = \bar{u}(t) \quad \forall t \in T\setminus I_\varepsilon^{\alpha\bar{u}}$$

We calculate the α which gives the new step length from

$$\alpha_\varepsilon(\bar{u}) = \{\alpha \mid \alpha = \max\{\tilde{\alpha}\in T \mid \Delta g_0(u_\varepsilon^{\alpha^1}, \bar{u}) \le \alpha^1 \sigma_\varepsilon(\bar{u})/2,$$

$$g(\bar{u}) + \Delta g(u_\varepsilon^{\alpha^1}, \bar{u}) \le 0\}, \quad \forall \alpha^1 \in [0, \tilde{\alpha}]] \tag{6.87}$$

Computional Procedure

We describe 4 different algorithms. In the algorithm description the parameter s is set equal to 1, 2, 3 or 4 according to whether algorithm 1, 2, 3, 4 is being described.

Step 0. Select an algorithm identification parameter $s \in \{1,2,3,4.\}$.
Choose (or compute) a nominal control satisfying $g(u) \le 0$.

Step 1. If $s \in \{2,3,4\}$, set $\varepsilon = \varepsilon_0$. If $s=1$, set $\varepsilon = \infty$ (i.e. we consider all the constraints for $s=1$).

Step 2. Compute the state equation $\bar{x}(t)$.

Step 3. Set $J_\varepsilon = \{j \in P \mid g_j(u) \ge -\varepsilon\}$

Step 4. For all $j \in J_\varepsilon$ compute $\bar{\lambda}_j$ by solving equations (6.83), (6.84).

Step 5. Compute a $\hat{u} \in \bar{U}_\varepsilon(\bar{u})$ according to (6.82)

Step 6. Compute $\sigma_\varepsilon(\bar{u})$ according to (6.85)
If $\sigma_\varepsilon(\bar{u}) = 0$, stop

Step 7. If s=1 or 2 go to 9
Else go to 8

Step 8. If $\sigma_\varepsilon(u) \leq -\varepsilon$ go to Step 9
Else, set $\varepsilon = \varepsilon/2$ and go to 3

Step 9. Compute an $\alpha \in \alpha_\varepsilon(\bar{u})$ according to (6.87)
Set $\bar{u}(t) = u_\varepsilon^\alpha$

Step 10. If $s \in \{2,3\}$, set $\varepsilon = \varepsilon_o$ and go 2.
Else go to 2.

Remarks

At each iteration, algorithm 1 solves (6.83) and (6.84) once for the cost and once for each constraint. Algorithm 2 solves (6.83) and (6.84) once for the cost and once for each constraint within ε of being violated, ε fixed. Algorithms 3 and 4 do the same as 2 for several reducing values of ε until a test (Step 8) is satisfied. Algorithm 3 resets ε to its initial value ε_o at each iteration, while 4 does not.

In [46], convergence has been proved for the above algorithms. To implement the algorithms the reader should consult [46], which takes into account all the approximations needed to make the feasible curve finding sub problem finitely solvable.

CHAPTER 7

CONTROL CONSTRAINTS

We consider the class of control problems with control constraints of the form

$$g(u(t),t) \leq 0 \tag{7.1}$$

where g is a p-vector function.

The problem is to find u(t), $t \in T$, satisfying (7.1) which minimizes

$$V^u(x_o,t_o) \equiv \int_{t_o}^{t_f} L(x(t),u(t),t)dt + F(x(t_f),t_f) \tag{7.2}$$

where

$$\dot{x}(t) = f(x(t),u(t),t) \qquad x(t_o) = x_o \tag{7.3}$$

Let

$$\Omega \triangleq \{u(t): g(u(t),t) \leq 0, \text{ for all } t \in T\} \tag{7.4}$$

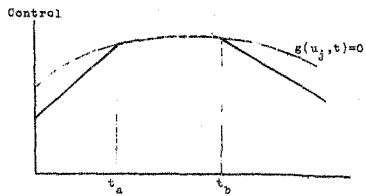
We have shown in Chapter 5 how the first order algorithms can be modified to handle control constraints of the form (7.1)

We now show how the second order algorithm of Chapter 5-II-4 is modified by using a quadratic approximation to the boundary of Ω.

JACOBSONS SECOND ORDER ALGORITHM [13], [18].

We make the following assumption: the optimal control $u^o(t)$ is continuous on the whole interval T, i.e. if and when a control hits or leaves a constraint it does so without a jump, as in Fig 7.1 (This clearly excludes the bang-bang problems which are treated in Chapter 8).

Fig. 7.1



t_a, t_b are times at which the constraint becomes active or inactive, respectively.

Assume that the optimal cost $V^o(x(t),t)$ is 'smooth enough' to allow a power series expansion about $\bar{x}(t)$.

$$V^o(\bar{x}+\delta x,t)=V^o(\bar{x},t) + V^o_x(\bar{x},t)\delta x + \tfrac{1}{2}\langle \delta x, V^o_{xx}(\bar{x},t)\delta x\rangle + \text{h.o.t.} \quad (7.5)$$

We may truncate the power series provided the truncated terms are negligible, i.e. we must limit the size

of the δx's. The Bellman equation in the neighbourhood of the nominal trajectory:

$$-\frac{\partial}{\partial t}V^{\bar{u}}(\bar{x},t) - \frac{\partial}{\partial t}a(t) - \langle\frac{\partial}{\partial t}V_x(\bar{x},t),\delta x\rangle - \tfrac{1}{2}\langle\delta x,\frac{\partial}{\partial t}V_{xx}(\bar{x},t)\delta x\rangle$$

$$= \min_{\delta:\bar{u}+\delta u\in\Omega}[L(\bar{x}+\delta x,\bar{u}+\delta u,t) + \langle V_x(\bar{x},t)+V_{xx}(\bar{x},t)\delta x, f(\bar{x}+\delta x,\bar{u}+\delta u,t)\rangle] \tag{7.6}$$

where the $V_{xxx}\delta x\delta x$ terms are neglected for the same reasons as given in 2-IV and:

$$V(\bar{x}+\delta x,t) = V^{\bar{u}}(\bar{x},t)+a(t)+V_x(\bar{x},t)\,\delta x+\tfrac{1}{2}\langle\delta x,V_{xx}(\bar{x},t)\delta x\rangle \tag{7.7}$$

The superscript o has been dropped for the following reasons:

The cost described by the truncated power series is optimal subject to the constraint that δx remain small. It is therefore not the truly optimal cost given that any size of δx is allowed. If the nominal trajectory is not sufficiently close to the optimal one (and we need to restrict the size of δx using the S.S.A.M. of Chapter 5-II), then it is not the optimal cost.

Let $\hat{g}(u,t)$ denote the set of $\hat{p}(t)$ active constraints at time t, for $u \in \Omega$.

PROPOSITION 7.1 Let $\bar{u} \in \Omega$. Assume that:

(i) $\hat{g}_u(\hat{u}(t),t)$ has full rank $\hat{p}$ and $[\hat{g}_u(\hat{u},t) : H_u(\bar{x},\hat{u},V_x(\bar{x},t)t)]$ has rank $\hat{p}$ (7.8)

(ii) $[H_{uu}(\bar{x},\hat{u},V_x(\bar{x},t),t) + b(t)\hat{g}_{uu}(\hat{u},t)]^{-1} > 0$ (7.9)

where

$$\hat{u}(t) = \arg\min_{u\in\Omega} H(\bar{x},u,V_x(\bar{x},t),t) \tag{7.10}$$

and b is calculated using equations (7.24), (7.25).

Then $a(t), V_x(\bar{x},t), V_{xx}(\bar{x},t)$ satisfy:

$$-\dot{a}(t) = H(\bar{x},\hat{u},V_x(\bar{x},t),t - H(\bar{x},\tilde{u},V_x(\bar{x},t),t) \tag{7.11}$$

$$-\dot{V}_x(\bar{x},t) = H_x(\bar{x},\hat{u},V_x(\bar{x},t),t)+V_{xx}(\bar{x},t)\Delta f(t) \tag{7.12}$$

$$-\dot{V}_{xx}(\bar{x},t) = H_{xx}(\bar{x},\hat{u},V_x(\bar{x},t),t)+f_x^m(\bar{x},\hat{u},t)P(t)+P(t)f_x(\bar{x},\hat{u},t) - \hat{R}^m(t)[H_{uu}(\bar{x},\hat{u},V_x(\bar{x},t),t)+bg_{uu}(\hat{u},t)]\hat{R}(t) \tag{7.13}$$

with boundary conditions:

$$a(t_f) = 0 \tag{7.14}$$

$$V_x(\bar{x}(t_f),t_f) = F_x(\bar{x}(t_f),t_f) \tag{7.15}$$

$$V_{xx}(\bar{x}(t_f),t_f) = F_{xx}(\bar{x}(t_f,t_f) \tag{7.16}$$

where

$$\Delta f(t) = f(\bar{x},\hat{u},t) - f(\bar{x},\tilde{u},t) \tag{7.17}$$

and

$$\dot{\hat{K}}(t) = -[H_{\hat{u}\hat{u}}(\bar{x},\hat{u},V_x(\bar{x},t),t)+b\hat{g}_{\hat{u}\hat{u}}(\hat{u},t)]^{-1} z(t)[H_{\hat{u}x}(\bar{x},\hat{u},V_x(\bar{x},t),t + f_{\hat{u}}^T(\bar{x},\hat{u},t)F(t)] \tag{7.18}$$

$$z(t) = I_m - g_{\hat{u}}^T(\hat{u},t)[\hat{g}_{\hat{u}}(u,t)[H_{\hat{u}\hat{u}}(\bar{x},\hat{u},V_x(\bar{x},t),t)+b\hat{g}_{\hat{u}\hat{u}}(\hat{u},t)]^{-1} g_{\hat{u}}^T(\hat{u},t)]^{-1} \times \hat{g}_{\hat{u}}(\hat{u},t)[H_{\hat{u}\hat{u}}(\bar{x},\hat{u},V_x(\bar{x},t),t)+b\hat{g}_{\hat{u}\hat{u}}(\hat{u},t)]^{-1} \tag{7.19}$$

Proof From (7.6):

$$-\frac{\partial}{\partial t}V^{\bar{u}}(\bar{x},t) - \frac{\partial a}{\partial t} - \frac{\partial}{\partial t}V_x(\bar{x},t)\delta x - \tfrac{1}{2}<\delta x\frac{\partial}{\partial t}V_{xx}(\bar{x},t)\delta x>$$

$$= \min_{\delta u:\bar{u}+\delta u\in\Omega}[H(\bar{x}+\delta x,\bar{u}+\delta u,V_x(\bar{x},t),t)+<V_{xx}(\bar{x},t)\delta x,f(\bar{x}+\delta x,\bar{u}+\delta u,t)>] \tag{7.20}$$

For $\delta x = 0$ the R.H.S. of equ (7.20) becomes

$$\min_{\delta u:\bar{u}+\delta u\in\Omega} H(\bar{x},\bar{u}+\delta u,V_x(\bar{x},t),t) \tag{7.21}$$

This is in general a difficult non-linear programming problem, which we shall assume can be solved. Let $\delta\hat{u}$ minimize the above expression, and $\hat{u} = \bar{u} + \delta\hat{u}$. If $g(\hat{u},t) < 0$, so $\hat{u} = u^*$, then the algorithm is the same as that of Chapter 5-II-4.

Assume that $\hat{p}(t)$ of the constraints are active at

time t, $0 < \hat{p}(t) \leq p$.

So
$$\hat{g}(\hat{u},t) = 0 \tag{7.22}$$

Adjoint (7.22) to (7.21) using a Lagrange multiplier b(t) of dimension $\hat{p}(t)$ to give:

$$J(\vec{x},\hat{u},b,V_x(\vec{x},t),t) = H(\vec{x},\hat{u},V_x(\vec{x},t),t) + \langle b(t),\hat{g}(\hat{u},t)\rangle \tag{7.23}$$

Using (i), (ii) we have, from McCormick [28], the following equations which enable us to determine b and $\hat{u}$

$$\frac{\partial J}{\partial \hat{u}} = H_u(\vec{x},\hat{u},V_x(\vec{x},t),t) + \hat{g}_{\hat{u}}^T(\hat{u},t)b = 0 \tag{7.24}$$

$$\frac{\partial J}{\partial b} = \hat{g}(\hat{u},t) = 0 \tag{7.25}$$

Now, reintroducing δx into (7.21) and δu to retain optimality gives

$$\min_{\delta u:\hat{u}+\delta u\in\hat{\Omega}} [H(\vec{x}+\delta x,\hat{u}+\delta u,V_x(\vec{x},t),t)+\langle V_{xx}(\vec{x},t)\delta x,f(\vec{x}+\delta x,\hat{u}+\delta u,t)\rangle] \tag{7.26}$$

where the minimizing δu still has to be determined. Assume that at time t all the constraints $\hat{g}(\hat{u}+\delta u,t)$ remain active. To ensure this b must change from b to $b + \delta b$.

Analogous conditions to (7.24), (7.25) are:

$$H_u(\bar{x}+\delta x,\hat{u}+\delta u,V_x(\bar{x},t),t)+f_u^T(\bar{x}+\delta x,\hat{u}+\delta u,t)V_{xx}(\bar{x},t)\delta x$$
$$+g_u^T(\hat{u}+\delta u,t)(b+\delta b) \quad = 0 \qquad (7.27)$$

$$g(\hat{u}+\delta u,t) \quad = 0 \qquad (7.28)$$

Expanding to first order about $\bar{x}$, $\hat{u}$ (to obtain a linear relationship between δx and δu) gives, using equations (7.24), (7.25),

$$[H_{uu}(\bar{x},\hat{u},V_x(\bar{x},t),t) + b\hat{g}_{uu}(\hat{u},t)]\delta u + \hat{g}_u^T\delta b$$
$$+ [H_{ux}(\bar{x},\hat{u},V_x(x,t),t)+f_u^T(\bar{x},\hat{u},t)P(t)]\delta x = 0 \qquad (7.29)$$

$$\hat{g}_u(\hat{u},t)\delta u \quad = 0 \qquad (7.30)$$

Thus

$$\delta u(t) = -[H_{uu}(\bar{x},\hat{u},V_x(\bar{x},t),t)+b\hat{g}_{uu}(\hat{u},t)]^{-1}[\hat{g}_u(\hat{u},t)\delta b$$
$$+(H_{ux}(\bar{x},\hat{u},V_x(\bar{x},t),t)+f_u^T(\bar{x},\hat{u},t)P(t))\delta x] \qquad (7.31)$$

where assumption (ii) gaurantees the minimization of (7.26). Using (7.31) in (7.30) gives:

$$\delta b(t) = -[\hat{g}_u^T(\hat{u},t)(H_{uu}(\bar{x},\hat{u},V_x(\bar{x},t),t)+b\hat{g}_{uu}(\hat{u},t))^{-1}\hat{g}_u(\hat{u},t)]^{-1}\times$$
$$\hat{g}_u(\hat{u},t)[H_{uu}(\bar{x},\hat{u},V_x(\bar{x},t),t)+b\hat{g}_{uu}(\hat{u},t)]^{-1}[H_{ux}(\bar{x},\hat{u},V_x(\bar{x},t),t)$$
$$+ f_u^T(\bar{x},\hat{u},t)P(t)]\delta x \qquad (7.32)$$

and substituting this into (7.31) leads to:

$\delta u = \hat{K}(t)\delta x$, $\hat{K}(t)$ defined in (7.18)

Expand the expression in equation (7.26) to second order and substitute for δu, δb. Then, equating the coefficients of δx with those on the L.H.S of (7.20) gives equations (7.11), (7.12), (7.13).

Terminal conditions : See Proposition 2.4

□

Remarks

(i) When $\hat{g}(t) = 0$, $\forall t \in T$, then $z = I_m$ and the equations reduce to those of Chapter 5-II-4.

(ii) Local, strict convexity at $\hat{u}$ is sufficient to ensure Assumption (7.9) is satisfied.

(iii) If only one control is used, from equation (7.30) $\delta u = 0$ if $\hat{g}_u \neq 0$; and $z = \hat{K}(t) = 0$ even if $(H_{uu}+b\hat{g}_{uu}) \leq 0$. Equation (7.13) is then a linear matrix equation.

(iv) At points where a constraint ceases to be active or inactive, z will change discontinuously. However, as $\hat{u}$ is continuous (by assumption), only $\hat{V}_{xx}(\bar{x},t)$ will have a discontinuity.

(v) On the forward run, u(t) will be discontinuous at times of boundary points of $\hat{u}(t)$ due to the presence

of the discontinuous z(t). However, this discontinuity does not affect the cost to second order. See [13].

The discontinuity in the forwards u(t) can be overcome easily by using the computational trick of Chapter 5-II-4.

i.e. compute

$$u(t) = \arg \min_{u \in \Omega} H(\bar{x}+\delta x, u, V_x + V_{xx}\delta x, t)$$

Since V_x, V_{xx} are continuous, the u(t) generated is continuous.

Computational Procedure

This is much the same as the second order D.D.P. algorithm given in Chapter 5-II-4, with modifications.

Step 0. Choose a nominal control $\bar{u}(t), t \in T$ satisfying $g(\bar{u},t) \leq 0, \quad \forall t \in T$.

Step 1. Standard.

Step 2. Using boundary conditions (7.14) - (7.16) integrate equs (7.11) - (7.13) backward in time from t_f to t_o, all the while minimizing H w.r.t. u to obtain $\hat{u}(t)$, and $\hat{\xi}$. Use equations (7.24), (7.25) to calculate b, which enables z and $\hat{R}(t)$ to be calculated using equations (7.18), (7.19) Store $\hat{u}$, $\hat{R}(t)$.

Note the time N_{eff} when $|a(\bar{x},t)|$ becomes greater than η.

Steps 3, 4, 5. Standard.

Step 3 of S.S.A.M. must be modified in part to

$$u(t) = \bar{u}(t) \qquad t \in [1,N.]$$

$$= \hat{u}(t) + \hat{R}(t)\delta x \qquad t \in [N_1,N].$$

Otherwise the rest of S.S.A.M. remains the same.

In this Chapter we compared the nominal control $\bar{u}$ with the control $\hat{u}$, where

$$\hat{u}(t) = \arg \min_{u \in \hat{\Omega}} H(\bar{x},u,V_x(\bar{x},t),t).$$

We have not yet derived results which allow us to compare two arbitrary controls, as in previous sections.

CHAPTER 8

BANG-BANG CONTROL PROBLEMS

I PROPERTIES OF THE COST FUNCTION AT SWITCHING POINTS

Consider

$$\dot{x}(t) = f(x(t;x_o,t_o,u),u(t),t) \qquad x(t_o) = x_o \tag{8.1}$$

$$V^u(x(t),t) \triangleq \int_t^{t_f} L(x(\tau),u(\tau),\tau)d\tau + F(x(t_f),t_f) \tag{8.2}$$

where u(t) satisfies for all $t \in T$,

$$u_j^b \leq u_j(t) \leq u_j^a \qquad j = 1,.. m \tag{8.3}$$

and u_j^b, u_j^a, $j = 1,.. m$ are constant values.

In this Chapter we consider the case where the control law switches between the upper and lower bounds defined in (8.3).

Assume we have a nominal control $\bar{u}(t) \in G$ satisfying (8.3), with resultant nominal trajectory

$$\bar{x}(t) \triangleq x(t;x_o,t_o,\bar{u})$$

Let u*(t) denote the control:

$$\begin{aligned} u^*(t) &= v^- \quad t \in [t_a,t_1) \\ &= v^+ \quad t \in (t_1,t_f] \qquad t_o \leq t_a \leq t_1 \leq t_f \end{aligned} \tag{8.4}$$

where $u^*(t)$ has left and right limits of v^- and v^+ respectively at t_1, where v^- and v^+ are constant vectors with components either u_j^b or u_j^a $j = 1,.. m$ i.e. travelling backwards in time from t_f, t_1 is the time when at least one of the components of $u^*(t)$ changes discontinuously from one of the bounds to the other.

We now use a dynamic programming approach to investigate the properties of $V^{u^*}(\bar{x}(t),t)$ further, in particular at t_1, where $u^*(t)$ is restricted to a piecewise continuous function of the form (8.4).

For the control $u^*(t) \in G$ defined by (8.4), $t \in [t_a, t_f]$ define

$$\theta(\bar{x}(t),t) \triangleq V^{u^*(t)}(\bar{x}(t),t) \tag{8.5}$$

Let superscripts +, - denote the right and left limits respectively of θ at t_1.

We have from equ (2.9), for $t \in \theta(u^*)$

$$-\theta_t(\bar{x}(t),t) = H(\bar{x},u^*,\theta_x(\bar{x}(t),t),t) \tag{8.6}$$

so

$$-\theta_{xt}(\bar{x}(t),t) = H_x(\bar{x},u^*,\theta_x(\bar{x},t),t)+\theta_{xx}(\bar{x},t)f(\bar{x},u^*,t) \tag{8.7}$$

$$-\theta_{tt}(\bar{x}(t),t) = H_t(\bar{x},u^*,\theta_x(\bar{x},t),t) + \theta_{xt}(\bar{x},t)f(\bar{x},u^*,t)$$

$$= H_t(\bar{x},u^*,\theta_x(\bar{x},t),t) - \langle H_x(\bar{x},u^*,\theta_x(\bar{x},t),t), f(\bar{x},u^*,t)\rangle$$

$$-\langle f(\bar{x},u^*,t)\theta_{xx}(\bar{x},t)f(\bar{x},u^*,t)\rangle \qquad (8.8)$$

where $u^* = v^- \quad t \in [t_a,t_1)$

$\quad\quad = v^+ \quad t \in (t_1,t_f]$

For $t \in [t_a,t_f]$ the cost to go is given by

$$\int_t^{t_1} L(\bar{x}(\tau),u^*(\tau),\tau)d\tau + \theta(\bar{x}(t),t) \qquad (8.9)$$

which, as we can see, can be considered a function of t_1. So, for the cost to go for $t \leq t_1$, define

$$\theta(\bar{x}(t),t_1,t) = \int_t^{t_1} L(\bar{x}(\tau),u^*(\tau),\tau)d\tau + \theta(\bar{x}(t_1),t_1)$$

$$= \int_t^{t_1} L(\bar{x}(\tau),u^*(\tau),\tau)d\tau + \theta(\bar{x}(t)+\Delta x(t_1),t_1) \qquad (8.10)$$

where

$$\Delta x(t_1) = \int_t^{t_1} f(\bar{x}(\tau),u^*(\tau),\tau)d\tau \qquad (8.11)$$

PROPOSITION 8.1 Let H1, H2, H3 be satisfied, s = 3. Let $u^*, \bar{u} \in G$. Let u^* be given by (8.4). Then

$$\theta^-_{t_1}(\bar{x}(t_1),t_1;t_1) = \theta^+_{t_1}(\bar{x}(t_1),t_1;t_1)$$

$$= H(\bar{x}(t_1),v^-,\theta_x(\bar{x}(t_1),t_1),t_1) - H(\bar{x}(t_1),v^+,\theta_x(\bar{x}(t_1),t_1),t_1) \tag{8.12}$$

$$\theta^-_{xt_1}(\bar{x}(t_1),t_1;t_1) = \theta^+_{xt_1}(\bar{x}(t_1),t_1;t_1)$$

$$= H_x(\bar{x}(t_1),v^-,\theta_x(\bar{x}(t_1),t_1),t_1) - H_x(\bar{x}(t_1),v^+,\theta_x(\bar{x}(t_1),t_1),t_1)$$

$$+ \theta_{xx}(\bar{x}(t_1),t_1)\Delta f(t) \tag{8.13}$$

$$\theta^-_{t_1t_1}(\bar{x}(t_1),t_1;t_1) = \theta^+_{t_1t_1}(\bar{x}(t_1),t_1;t_1)$$

$$= H_t(\bar{x}(t_1),v^-,\theta_x(\bar{x}(t_1),t_1),t_1) - H_t(\bar{x}(t_1),v^+,\theta_x(\bar{x}(t_1),t_1),t_1)$$

$$+ [H_x(\bar{x}(t_1),v^-,\theta_x(\bar{x},t_1),t_1) - H_x(\bar{x}(t_1)v^+,\theta_x(\bar{x},t_1),t_1)]f(\bar{x},v^-,t_1)$$

$$-\Delta f^T(t_1)H_x(\bar{x}(t_1),v^+,\theta_x(\bar{x},t_1),t_1) + \Delta f^T(t_1)\theta_{xx}(\bar{x},t_1)\Delta f(t_1) \tag{8.14}$$

where

$$\Delta f(t) \triangleq f(\bar{x},v^-,t) - f(\bar{x},v^+,t)$$

Proof Introducing δx and δt in the switching time into equation (8.10) gives:

$$\theta(\bar{x}+\delta x, t_1+\delta t_1, t) = \int_t^{t_1+\delta t_1} L(\bar{x}+\delta x, u^e, \tau)\,d\tau$$

$$+\theta(\bar{x}(t)+\delta x(t)+\Delta x(t_1+\delta t_1), t_1+\delta t_1)$$

Now for $\delta t_1 > 0$, we have for $t = t_1$

$$\theta(\bar{x}+\delta x, t_1+\delta t_1, t_1) = \int_{t_1}^{t_1+\delta t_1} L(\bar{x}+\delta x, \bar{v}, \tau)\,d\tau \qquad (8.15)$$

$$+\theta(\bar{x}(t_1)+\delta x(t_1)+\Delta x(t_1+\delta t_1); t_1+\delta t_1)$$

where

$$\Delta x(t_1+\delta t_1) = \int_t^{t_1+\delta t_1} f(\bar{x}+\delta x, \bar{v}, \tau)\,d\tau$$

Expanding the L.H.S. of (8.15) to second order about $\bar{x}, t_1$ for $\delta t_1 > 0$

$$\theta(\bar{x}+\delta x, t_1+\delta t_1; t_1) = \theta^+(\bar{x}, t_1, t_1) + \langle\theta^+_x, \delta x\rangle + \theta^+_{t_1}\delta t_1 + \langle\theta^+_{xt_1}, \delta x\rangle \delta t_1$$

$$+ \tfrac{1}{2}\langle\delta x, \theta^+_{xx}\delta x\rangle + \tfrac{1}{2}\theta^+_{t_1 t_1}\delta t_1^2 \qquad (8.16)$$

The superscript + applies as we are approaching t_1 from the right.

Expand the R.H.S. of (8.15) to second order about $\bar{x}$, t_1. Equation (8.15) is precisely the same form as equ (6.45) in Chapter 6 with ψ absent and θ^+ replacing F. (To see the actual derivation in detail refer to Jacobson [16], Appendix).

Then equating like coefficients of δx, δt, one obtains equations

$$\theta^+_{t_1}(\bar{x},t_1;t_1) = H(\bar{x},v^-,\theta^+_x(\bar{x},t_1),t_1) + \theta^+_t(\bar{x},t_1) \tag{8.17}$$

$$\theta^+_{xt_1}(\bar{x},t_1;t_1) = \theta^+_{xt}(\bar{x},t_1) + H_x(\bar{x},v^-,\theta^+_x(\bar{x},t_1),t_1) + \theta^+_{xx}(\bar{x},t)f(\bar{x},v^-,t) \tag{8.18}$$

$$\theta^+_{t_1t_1}(\bar{x},t_1;t_1) = H_t(\bar{x},v^-,\theta^+_x(\bar{x},t),t) + \theta^+_{tt}(\bar{x}(t_1),t_1) + <H_x(\bar{x},v^-,\theta^+_x(\bar{x},t_1),t_1),f(\bar{x},v^-,t_1)> \tag{8.19}$$
$$+2<\theta^+_{xt}(\bar{x},t_1),f(\bar{x},v^-,t_1)>+<f(\bar{x},v^-,t_1),\theta^+_{xx}(\bar{x},t_1)f(\bar{x},v^-,t_1)>$$

Now we have from Proposition 2.1(i) that θ_x and θ_{xx} are continuous in (x,t) at t_1. Hence we can dispense with the superscripts + (and -) for θ_x, θ_{xx}.

Substituting equs (8.6) - (8.8) with $u^* = v^+$ at t_1^+ into equs (8.17) - (8.19) gives equs (8.12) - (8.14).

Consider $\delta t < 0$. Then we have for $t \geq t_1$

$$\theta(\bar{x}+\delta x,t_1+\delta t_1;t_1) = -\int_{t_1+\delta t_1}^{t_1} L(\bar{x}+\delta x,v^+,\tau)d\tau + \theta(\bar{x}(t_1)+\delta x(t_1)+\Delta x(t_1+\delta t_1);t_1+\delta t_1) \tag{8.20}$$

where

$$\Delta x(t_1+\delta t_1) = -\int_{t_1+\delta t_1}^{t_1} f(\bar{x}+\delta x,v^+,\tau)d\tau$$

Expanding the L.H.S. of equ (8.20) about $\bar{x},t_1$ for $\delta t < 0$ gives equ (8.16) with superscript + replaced by -. Expanding the R.H.S. of equ (8.20) in the same way as equ (6.45) in Chapter 6 was expanded, and equating coefficients of δx, δt_1 as before, gives

$$\theta^-_{t_1}(\bar{x},t_1,t_1) = -H(\bar{x},v^+,\theta_x(\bar{x},t_1),t_1) - \theta^-_t(\bar{x},t_1)$$

$$\theta^-_{xt_1}(\bar{x},t_1;t_1) = -\theta^-_{xt}(\bar{x},t_1) - H_x(\bar{x},v^+,\theta_x(\bar{x},t_1),t_1)$$
$$-\theta_{xx}(\bar{x},t_1)f(\bar{x},v^+,t_1)$$

$$\theta^-_{t_1t_1}(\bar{x},t_1;t_1) = -H_t(\bar{x},v^+,\theta_x(\bar{x},t),t) - \theta^-_{tt}(\bar{x},t_1)$$
$$-\langle H_x(\bar{x},v^+,\theta_x(\bar{x},t_1),t_1),f(\bar{x},v^+,t_1)\rangle$$
$$-2\langle\theta^-_{xt}(\bar{x},t_1),f(\bar{x},v^+,t_1)\rangle-\langle f(\bar{x},v^+,t_1),\theta_{xx}f(\bar{x},v^+,t_1)\rangle$$

Substituting equs (8.6) - (8.8) with $u^* = v^-$ at t_1^- into the above equations gives the result.

□

Remark: In general, if more than one component (say r, where $r \leq m$) of u^* changes at t_1, it may be necessary to consider the separate changes in switch times $\delta t_{1_1}, \ldots \delta t_{1_r}$ for each component in the analysis above. However, here we treat the case where only one component of the control switches at t_1.

II THE OPTIMAL COST FOR STATE $(\bar{x}(t),t)$

We consider minimizing (8.2) subject to (8.1). We assume that the optimal control switches between the upper and lower bounds defined in (8.3). We investigate what happens to $V^o(x,t)$ and its partial derivatives at the switch points, in view of the fact that the Bellman P.D.E. was derived on the assumption that the $V^o(x,t)$ had continuous first and second partial derivatives w.r.t. x and t. At points where the control is discontinuous neither $\dot{x}(t)$ nor the derivatives of $V^o(x,t)$ will necessarily be continuous.

Consider the state $x = \bar{x}$.

Assume we have a single discontinuity at time t_1. Let $V^+(\bar{x},t)$ denote the optimal return function to the right of t_1, where a continuous control v^+ is used for $t \in (t_1,t_f]$. i.e. v^+ is the control that minimizes the R.H.S. of the Bellman P.D.E. for the state $x = \bar{x}(t)$, $t \in (t_1,t_f]$.

Suppose constant control v^- is used to the left of the discontinuity. Then, for $t \leq t_1$, denote the cost to go for state $\bar{x}(t)$ by

$$\theta^-(\bar{x}(t),t_1,t) = \int_t^{t_1} L(x,v^-,\tau)d\tau + V^+(\bar{x}(t_1),t_1)$$

(we drop the superscript for convenience).

We have from Proposition 2.1(i) that at $t = t_1$

$$\theta^{-}(\bar{x}(t_1);t_1,t_1) = V^{+}(\bar{x}(t_1),t_1) \tag{8.21}$$

$$\theta^{-}_{x}(\bar{x}(t_1);t_1;t_1) = V^{+}_{x}(\bar{x}(t_1),t_1) \tag{8.22}$$

$$\theta^{-}_{xx}(\bar{x}(t_1),t_1;t_1) = V^{+}_{xx}(\bar{x}(t_1),t_1) \tag{8.23}$$

and from Proposition (8.1)

$$\theta^{-}_{t_1}(\bar{x}(t_1),t_1;t_1) = V^{+}_{t_1}(\bar{x}(t_1);t_1) \tag{8.24}$$

$$\theta^{-}_{xt_1}(\bar{x}(t_1),t_1;t_1) = V^{+}_{xt_1}(\bar{x}(t_1),t_1) \tag{8.26}$$

$$\theta^{-}_{t_1t_1}(\bar{x}(t_1),t_1;t_1) = V^{+}_{t_1t_1}\bar{x}(t_1),t_1) \tag{8.26}$$

where + and - denote right and left limits respectively.

Now, θ^{-} is not the optimal return function $V^{-}(\bar{x},t)$ to the left of t_1 unless t_1 is chosen optimally. *i.e. θ^{-} must be minimized w.r.t. t_1.*

NECESSARY CONDITIONS FOR A MINIMUM

A necessary condition for $\theta^{-}(\bar{x}(t_1),t_1,t_1)$ to be minimized w.r.t. t_1 is that

$$\theta^{-}_{t_1}(\bar{x}(t_1),t_1;t_1) = 0 \tag{8.27}$$

A further necessary condition is

$$\theta^{-}_{t_1 t_1}(\bar{x}(t_1),t_1;t_1) \geq 0 \tag{8.28}$$

Now, introduce variations δx. In order to maintain the necessary condition of optimality for the case where variations δx are present it is required that we introduce variations δt_1 to ensure that

$$\theta^{-}_{t_1}(\bar{x}+\delta x,t_1+\delta t_1;t_1) = 0 \tag{8.29}$$

But

$$\theta^{-}_{t_1}(\bar{x}+\delta x,t_1+\delta t_1;t_1) = \theta^{-}_{t_1}+<\theta^{-}_{xt_1},\delta x>+\theta^{-}_{t_1 t_1}\delta t_1 \tag{8.30}$$

Using (8.21) - (8.24) gives:

$$\delta_{t_1} = -[\theta^{-}_{t_1 t_1}(\bar{x},t_1;t_1)]^{-1}<\theta^{-}_{xt_1}(\bar{x},t_1;t_1),\delta x> \tag{8.31}$$

where

$$\theta^{-}_{t_1 t_1}(\bar{x},t_1;t_1) > 0$$

This is an optimal local linear feedback controller relating the required switch time δt_1 to the δx appearing at $t = t_1$.

JUMP CONDITIONS FOR $V^{o}(\bar{x},t)$

We now relate $V^{+}_{x}(\bar{x}(t_1),t_1)$, $V^{+}_{xx}(\bar{x}(t_1),t_1)$ and $V^{-}_{x}(\bar{x}(t_1),t_1)$, $V^{-}_{xx}(\bar{x}(t_1),t_1)$. We have

A further necessary condition is

$$\theta^{-}_{t_1 t_1}(\bar{x}(t_1),t_1;t_1) \geq 0 \tag{8.28}$$

Now, introduce variations δx. In order to maintain the necessary condition of optimality for the case where variations δx are present it is required that we introduce variations δt_1 to ensure that

$$\theta^{-}_{t_1}(\bar{x}+\delta x,t_1+\delta t_1;t_1) = 0 \tag{8.29}$$

But

$$\theta^{-}_{t_1}(\bar{x}+\delta x,t_1+\delta t_1;t_1) = \theta^{-}_{t_1}+<\theta^{-}_{xt_1},\delta x>+\theta^{-}_{t_1 t_1}\delta t_1 \tag{8.30}$$

Using (8.21) - (8.24) gives:

$$\delta_{t_1} = -[\theta^{-}_{t_1 t_1}(\bar{x},t_1;t_1)]^{-1}<\theta^{-}_{xt_1}(\bar{x},t_1;t_1),\delta x> \tag{8.31}$$

where

$$\theta^{-}_{t_1 t_1}(\bar{x},t_1;t_1) > 0$$

This is an optimal local linear feedback controller relating the required switch time δt_1 to the δx appearing at $t = t_1$.

JUMP CONDITIONS FOR $V^{o}(\bar{x},t)$

We now relate $V^{+}_{x}(\bar{x}(t_1),t_1), V^{+}_{xx}(\bar{x}(t_1),t_1)$ and $V^{-}_{x}(\bar{x}(t_1),t_1)$, $V^{-}_{xx}(\bar{x}(t_1),t_1)$. We have

$$\bar{\theta}(\bar{x}+\delta x, t_1+\delta t_1, t_1)$$

$$= \theta^- + \langle\theta^-_x, \delta x\rangle + \theta^-_{t_1}\delta t_1 + \langle\theta^-_{xt_1}, \delta x\rangle \delta t_1$$

$$+ \tfrac{1}{2}\langle\delta x, \theta^-_{xx}\delta x\rangle + \tfrac{1}{2}\theta^-_{t_1t_1}\delta t_1^2 \qquad (8.32)$$

Now, substituting (8.31) into (8.32) to eliminate δt_1 gives:

$$\theta^-(\bar{x}+\delta x, t_1 - (\theta^-_{t_1t_1})^{-1}\langle\theta^-_{xt_1}, \delta x\rangle, t_1)$$

$$= \theta^- + \langle\theta^-_x, \delta x\rangle + \tfrac{1}{2}\langle\delta x, (\theta^-_{xx} - \frac{\theta^-_{xt_1}}{\theta^-_{t_1t_1}}\theta_{t_1x})\delta x\rangle \qquad (8.33)$$

Renaming the L.H.S. (which is now independent of t_1) of (8.27) as $V^-(\bar{x}+\delta x, t)$, we have from equs (8.21) - (8.26)

$$V^-(\bar{x}, t_1) = V^+(\bar{x}, t_1) \qquad (8.34)$$

$$V^-_x(\bar{x}, t_1) = V^+_x(\bar{x}, t_1) \qquad (8.35)$$

$$V^-_{xx}(\bar{x}, t_1) = V^+_{xx}(\bar{x}, t_1) - (V^+_{xt_1} \cdot V^+_{t_1x})/V^+_{t_1t_1} \qquad (8.36)$$

These results were derived independently by Jacobson [16], and Dyer, McReynolds [6].

III ALGORITHMS FOR SOLVING BANG-BANG CONTROL PROBLEMS.

Consider the dynamic system described by

$$\dot{x}(t) = f(x,u,t) = f_1(x,t)+f_2(x,t)u; \quad x(t_o) = x_o$$

where f_1 is an n-dimensional, nonlinear vector function of x and t, f_2 is an nxm matrix function of x and t, and u is a m-dimensional control vector which is required to satisfy (8.3).

The problem is to choose u(t), $t \in T$ to satisfy (8.3) and to minimize

$$V(x_o,t_o) = \int_{t_o}^{t_f} L(x,t)dt + F(x(t_f):t_f)$$

Then, we have

$$H(x,u,V_x(x,t),t) = L(x,t)+\langle V_x, f_1(x,t)+f_2(x,t)u\rangle$$

If $(f_2^T V_x)_j$, $j=1,..m$ is not zero on a finite interval (i.e. the problem is singular or partially singular - see Chapter 10), then H is minimized w.r.t. u if

$$u_j = u_j^b \qquad (f_2^T V_x)_j > 0 \qquad j = 1,..m. \tag{8.37}$$

$$u_j = u_j^a \qquad (f_2^T V_x)_j < 0 \qquad j = 1,...m.$$

So our control is of the type which we have been

discussing. We now discuss how the existing algorithms are adapted to handle the above problem.

1. The Gradient Method, [6], [7], [8].

Assume we have a nominal control $\bar{u}(t)$ of the form (8.4) with a switching point at $t = t_1$, and associated nominal trajectory $\bar{x}(t)$, $t \in T$. Now, because the switching time t_1 has not necessarily been chosen optimally, denote the cost by $\theta(\bar{x},t_1,t)$ for state $\bar{x}(t)$, where $u^*(t) = \bar{u}(t)$. The change in switching times δt_1 is chosen to ensure that $\theta^-(\bar{x}+\delta x,t_1+\delta t_1;t_1)$ is less than $\theta^-(\bar{x},t_1;t_1)$.

Expanding to second order:

$$\theta^-(\bar{x}+\delta x,t_1+\delta t_1,t_1)$$

$$= \theta^- + <\theta^-_x,\delta x> + \theta^-_{t_1}\delta t_1 + <\theta^-_{xt_1},\delta x>\delta t_1$$

$$+ \tfrac{1}{2}<\delta x,\theta^-_{xx}\delta x> + \tfrac{1}{2}\theta^-_{t_1t_1}\delta t_1^2 \qquad (8.38)$$

We can choose, for example

$$\delta t_1 = -\varepsilon\theta^-_{t_1}(\bar{x},t_1;t_1) \qquad (8.39)$$

or

$$\delta t_1 = -\varepsilon\,\mathrm{sgn}(\theta^-_{t_1}(\bar{x},t_1,t_1)) \qquad (8.40)$$

ε is chosen to ensure the validity of expansion (8.38).

If δt_1 is chosen according to (8.39), we have, eliminating δt_1 and using equations (8.21) - (8.26), the following jump conditions at t_1

$$\theta_x^-(\bar{x},t_1,t_1) = \theta_x^+(\bar{x},t_1;t_1) - \varepsilon\theta_{t_1}(\bar{x},t_1;t_1)\theta_{xt_1}(\bar{x},t_1;t_1) \qquad (8.41)$$

$$\theta_{xx}^-(\bar{x},t_1,t_1) = \theta_{xx}^+(\bar{x},t_1,t_1) \qquad (8.42)$$

If δt_1 is chosen according to (8.40), we have

$$\theta_x^-(\bar{x},t_1;t_1) = \theta_x^+(\bar{x},t_1;t_1) - \varepsilon\mathrm{sgn}(\theta_{t_1}(\bar{x},t_1;t_1))\theta_{xt_1}(\bar{x},t_1;t_1) \qquad (8.43)$$

$$\theta_{xx}^-(\bar{x},t_1;t_1) = \theta_{xx}^+(\bar{x},t_1;t_1) \qquad (8.44)$$

Computational Procedure

Step 0. Choose nominal control $\bar{u}(t)$, $t \in T$ of the form (8.4). Run $\bar{x}(t)$. Calculate $V^{\bar{u}}(x_o,t_o)$ $(\equiv\theta(x_o,t_o))$ Store $\bar{x}(t)$, $\bar{u}(t)$, $V^{\bar{u}}$.

Step 1. Choose $\varepsilon > 0$.

Step 2. Using boundary conditions (2.17), (2.19) integrate equations (2.16), (2.18) backwards in time until the first switching point of $\bar{u}(t)$, to obtain θ_x^+, θ_{xx}^+.

Step 3. Compute $\theta_{t_1}, \theta_{xt_1}$ from equations (8.12), (8.13). Store.
Calculate $\theta_{\tilde{x}}$, $\theta_{\tilde{x}\tilde{x}}$ from equations (8.41), (8.42) or equations (8.43), (8.44).
Return to Step 2 and continue integrating until the initial time.

Step 4. Calculate the new control u(t) with change in switch times of $\tilde{u}(t)$ given by (8.39) or (8.40).
Calculate x(t) using this control, and the new cost $V^u(x_o,t_o)$.

Step 5. If $V^u \geq V^{\tilde{u}}$, set $\varepsilon = \varepsilon/2$. Go to step 2.
Otherwise, set $\tilde{u}(t) = u(t)$

$$\tilde{x}(t) = x(t)$$

$$V^{\tilde{u}}(x_o,t_o) = V^u(x_o,t_o).$$ Store.

Go to 1.

The above procedure is repeated until the optimal solution is found or no further improvements can be made.

2. JACOBSON'S FIRST ORDER METHOD, [13], [16], [18].

Suppose we have arbitrary nominal control $\tilde{u}(t)$ satisfying constraints (8.3), with associated trajectory $\tilde{x}(t)$. Use equations (2.34) - (2.37).

Suppose

$$u^*(t) = \arg\min_u H(\tilde{x}(t),u,\theta_x(\tilde{x}(t),t),t)$$

then u'(t) is given by equation (8.37).

The computation procedure is the same as the first order method described in Chapter 5-I-2.

3. THE SUCCESSIVE SWEEP ALGORITHM, [6], [7], [8].

Suppose we have some nominal control $\bar{u}(t)$ of the form (8.4) with switching time at t_1, and associated trajectory $\bar{x}(t)$. Denote the cost by $\theta(\bar{x},t_1,t)$, $u^*(t) = \bar{u}(t)$. We solve equations (2.16) - (2.19) (with $\bar{\lambda}(t)=\theta_x$, $\bar{P}(t)=\theta_{xx}$).

If $\theta_{t_1} \neq 0$, choose δt_1 to minimize the R.H.S. of (8.32). This gives

$$\delta_{t_1} = -\theta_{t_1t_1}^{-1}(\theta_{t_1}+\theta_{xt_1}\delta x) \tag{8.45}$$

giving jump conditions

$$\theta_x^- = \theta_x^+ - \theta_{t_1}^+\theta_{xt_1}^+/\theta_{t_1t_1}^+ \tag{8.46}$$

$$\theta_{xx}^- = \theta_{xx}^+ - \theta_{xt_1}^+\cdot\theta_{t_1x}^+/\theta_{t_1t_1}^+ \tag{8.47}$$

where θ_{t_1}, θ_{xt_1}, $\theta_{t_1t_1}$ are given by (8.12) - (8.14).

Computational Procedure

Step 0. Choose nominal control $\bar{u}(t)$ of the form (8.4).
Run $\bar{x}(t)$. Calculate $\theta(\bar{x}_o,t_o)$.
Store $\bar{x}$, $\bar{u}$, 0.

Step 1. Using boundary conditions (2.17), (2.19) integrate

equations (2.16), (2.18) backward in time as far as the first switching time of $\bar{u}(t)$.

Compute and store θ_{t_1}, θ_{xt_1}, $\theta_{t_1t_1}$ using equations (8.12)-(8.14). Compute θ_x^-, θ_{xx}^- using (8.46), (8.47).

Step 2. Continue integrating to next switching time.

Step 3. Repeat 1, 2 until initial time.

Step 4. Integrate the state equation forward to the first switch.

Compute δt_1 given in equation (8.45) (NB at first switch $\delta x = 0$).

Step 5. Continue integrating and storing $x(t)$ to t_f.

Compute the new performance index $V^u(x_o,t_o)$ where $u(t)$ differs from $\bar{u}(t)$ in respect of the switcing times.

Set $\bar{u}(t) = u(t)$

$\bar{x}(t) = x(t)$

$\theta = V^u$. Go to 1.

Repeat Steps 1 - 5 until no further improvement is made.

Remarks

(i) In general several gradient steps may have to be taken before the full Newton - Raphson step could be used, because equation (8.45) may be invalid (δt may be too large, or $\theta_{t_1t_1} \leq 0$), for nominal trajectory.

(ii) The choice of the nominal control is important.

4. JACOBSON'S SECOND ORDER ALGORITHM [13], [16], [18].

Suppose we have a nominal control $\bar{u}(t)$, $t \in T$ satisfying (8.3), with associated trajectory $\bar{x}(t)$. Let

$$u^*(t) = \arg\min_u H(\bar{x}(t),u,V_x(\bar{x},t),t)$$

where the components of $u^*(t)$ are given by (8.37).

Now suppose that the minimizing control for $\bar{u}+\delta u$ remains $u^*(t)$ except in the neighbourhood of switch points $u^*(t)$.

Then, we solve, for $t \in \theta(u^*)$,

$$-\dot{a}(t) = H(\bar{x},u^*,V_x(\bar{x},t),t)-H(\bar{x},\bar{u},V_x(\bar{x},t),t) \tag{8.48}$$

$$-\dot{V}_x(\bar{x},t) = H_x(\bar{x},u^*,V_x(\bar{x},t),t)+V_{xx}(\bar{x},t)[f(\bar{x},u^*,t)-f(\bar{x},\bar{u},t)] \tag{8.49}$$

$$-\dot{V}_{xx}(\bar{x},t) = H_{xx}(\bar{x},u^*,V_x(\bar{x},t),t)+f_x^T V_{xx}(\bar{x},t)+V_{xx}(\bar{x},t)f_x^T + V_{xxx}(\bar{x},t)[f(\bar{x},u^*,t)-f(\bar{x},\bar{u},t)]. \tag{8.50}$$

with boundary conditions

$$a(t_f) = 0 \tag{8.51}$$

$$V_x(\bar{x},t_f) = F_x(\bar{x}(t_f),t_f) \tag{8.52}$$

$$V_{xx}(\bar{x},t_f) = F_{xx}(\bar{x}(t_f),t_f). \tag{8.53}$$

The jump conditions for $t \in T - \theta(u^*)$ are given by (8.34) - (8.36).

Equations (8.45) - (8.53) cannot be used computationally as equ. (8.50) involves V_{xxx}. In the same way as in Chapter 2-V, we estimate the error in $a(t)$, $V_x(t)$ and $V_{xx}(t)$ if the V_{xxx} terms are omitted, [17], [18].

We have from Proposition 2.7, for $d(\bar{u},u^*) \leq \varepsilon$ or $d_1(\bar{u},u^*) \leq \varepsilon$, the error introduced into $a(t)$ is of order ε^3, and into $V_x(t)$ is of order ε^2.

Owing to the neglect of the V_{xxx} terms an error in the switch times of u^*, of magnitude

$$\Delta t_i = -V_{t_i t_i}^{-1} f_u^T V_x\Big|_{t_i}$$

is introduced; where we have ΔV_x is of order ε^2. The error introduced into $a(t_i)$ is, to second order in Δt_i

$$V_{t_i}\Delta t_i + \tfrac{1}{2}V_{t_i t_i}\Delta t_i^2.$$

However, at a switch time $V_{t_i} = 0$. Thus the error introduced into $a(t_i)$ by Δt_i is of order ε^4.

We thus solve the following differential equations; for $t \in \theta(u^*)$

$$-\dot{\hat{a}}(t) = H(\bar{x},u^*,\hat{V}_x(\bar{x},t),t) - H(\bar{x},\bar{u},\hat{V}_x(\bar{x},t),t) \tag{8.54}$$

$$-\dot{\hat{V}}_x(\bar{x},t) = H_x(\bar{x},u^*,\hat{V}_x(\bar{x},t),t) + \hat{V}_{xx}(\bar{x},t)\Delta f(t) \tag{8.55}$$

$$-\dot{\hat{V}}_{xx}(\bar{x},t) = H_{xx}(\bar{x},u^*,\hat{V}_x(\bar{x},t),t)+f_x^T\hat{V}_{xx}(\bar{x},t)+\hat{V}_{xx}(\bar{x},t)f_x^T \qquad (8.56)$$

where $u^*_j = u_j^b \quad (f_1^T(\bar{x},t),\hat{V}_x(\bar{x},t))_j > 0 \quad j=1,..m \qquad (8.57)$

$$= u_j^a \quad (f_1^T(\bar{x},t),\hat{V}_x(\bar{x},t))_j < 0$$

Equations (8.54) - (8.56) have boundary conditions:

$$\hat{a}(t_f) = 0 \qquad (8.58)$$

$$\hat{V}_x(\bar{x},t_f) = F_x(\bar{x}(t_f),t_f) \qquad (8.59)$$

$$\hat{V}_{xx}(\bar{x},t_f) = f_{xx}(\bar{x}(t_f),t_f). \qquad (8.60)$$

$\hat{a}(t)$, V_x, V_{xx} are estimates for $a(t)$, V_x, V_{xx} such that

$$\|a(t) - \hat{a}(t)\| \leq c_1\varepsilon^3$$

$$\|V_x(\bar{x},t)-\hat{V}_x(\bar{x},t)\| \leq c_2\varepsilon^2$$

$$\|V_{xx}(\bar{x},t)-V_{xx}(\bar{x},t)\| \leq c_3\varepsilon \qquad c_1,c_2,c_3 < \infty.$$

The jump conditions for $T \sim \theta(u^*)$ are given by (8.34) - (8.36) (where all the partial derivatives of V have a $\sim$).

Computational Procedure

The algorithm of 5-II-4 is adapted as follows:

Step 0.1. Standard.

Step 2. Using boundary conditions (8.53) - (8.60) integrate

equs (8.54) - (8.56) backwards from t_f to t_o all the while minimizing H w.r.t. u to obtain $u^*(t)$. Calculate the jumps in V_{xx} at switch times of $u^*(t)$. Store $u^*(t)$, t_{1_i}, i=1,..p the switch times of $u^*(t)$, and $\beta_i = -V_{t_1 t_1}^{-1} V_{xt_1}$ at the switch times, where $V_{t_1 t_1}$, V_{xt_1} are given by equ (8.13), (8.14). Note the time N_{eff} when $|a(\bar{x},t)|$ becomes greater than η.

Step 3. Standard.

Step 4. Apply the *S.S.A.M.* with step 3 revised in part as follows:
Define, $u = \bar{u}$ on $[1,N_1]$
and u, as shown below, on $[N_1,N]$

Step 5. Standard.

DETERMINING A NEW CONTROL

Step 0. Set u^*, t_{1_i}, β_i from main algorithm.

Step 1. Run forward in time using new control $u_j(t) = u_j^*(t)$, j=1,..m. If t_f is reached return to Step 3 of S.S.A.M.
Otherwise, when a switch point of $u_j^*(t)$ is reached, measure $\delta x(t_{1_i})$ and calculate δt_{1_i} using equ. (8.31)

Step 2. If $\delta t_{1_i} = 0$, go to 3

If $\delta t_{1_i} < 0$, backspace integration procedure by amount δt_{1_i} and set

$$u_j(t) = u_j^*(t_{1_i}^+), \quad t \in (t_{1_i} - \delta t_{1_i}, t_{1_i}]$$

integrate forward to $t = t_{1_i}$. After time $t = t_{1_i}$.

Again set $u_j(t) = u_j^*(t)$. Go to 3.

If $\delta t_{1_i} > 0$, set

$$u_j(t) = u_j^*(t_{1_i}^-), \quad t \in [t_{1_i}, t_{1_i} + \delta t_{1_i}]$$

Integrate forward to $t = t_{1_i} + \delta t_{1_i}$.

(If $t_{1_i} + \delta t_{1_i} > t_f$ integrate only to $t = t_f$).

After time $t = t_{1_i} + \delta t_{1_i}$, again set $u_j(t) = u_j^*(t)$.

Go to 3.

Step 3. If $t = t_f$, return to Step 3 SSAM

Otherwise return to 1.

This algorithm can be extended to handle bang-bang control problems with Fixed Endpoints and Implicit final time in the same way that Jacobson's second-order algorithm was extended in Chapter 6. See [13], [16], [18].

CHAPTER 9

STATE CONSTRAINED PROBLEMS

In this Chapter we use the approach of Mayne to derive some recent results due to Jacobson, Lele, Speyer [26] for state constrained problems using, in part, the expressions for ΔV derived previously. This avoids the difficulty of treating the differential equation as an equality constraint. In [26] the differential equation is added to the cost function by means of Lagrange multipliers.

The basic problem considered is the same as that previously considered except that there is no integral cost ($L \equiv 0$), only terminal cost F, and there is an added constraint of the form

$$S(x(t)) \leq 0, \quad \text{for all } t \in T \tag{9.1}$$

where $S : R^n \to R^m$ and its derivatives up to order p are continuous. The constraint is assumed to be of p^{th} order, i.e. the p^{th} time derivative of the constraint is the first to contain the control variable explicitly. We also have

$$\dot{x}(t) = f(x(t),u(t)); \quad x(t_o) = x_o \tag{9.2}$$

Let $(\bar{x},\bar{u})$ denote the optimal solution, assumed to exist. Let $\bar{y}(t)$, $y(t)$ denote respectively $S(\bar{x}(t))$, $S(x(t))$.

Let

$$\delta y(t) \triangleq y(t)-\bar{y}(t). \tag{9.3}$$

An approximation $\delta\hat{y}(t)$ to $\delta y(t)$ can be obtained as the solution to

$$\dot{\delta\hat{x}}(t) = A(t)\delta\hat{x}(t)+B(t)\delta u(t) \tag{9.4}$$

$$\delta\hat{y}(t) = C(t)\delta\hat{x}(t) \tag{9.5}$$

$$\delta\hat{x}(t_o) = 0 \tag{9.6}$$

where

$$A(t) \triangleq f_x(\bar{x}(t),\bar{u}(t)) \tag{9.7}$$

$$B(t) \triangleq f_u(\bar{x}(t),\bar{u}(t)) \tag{9.8}$$

$$C(t) \triangleq S_x(\bar{x}(t)) \tag{9.9}$$

$$\delta u(t) \triangleq u(t)-\bar{u}(t). \tag{9.10}$$

Let $\bar{x}(t) \triangleq v_x^{\bar{u}}(\tilde{x}(t),t)$ be the solution to equ (2.16), (2.17).

PROPOSITION 9.1: Let H1A, H2A be satisfied, s = 2. Let $u, \bar{u} \in G$. Then, if $\max_{t\in T} \|\delta u(t)\| \leq \varepsilon$,

$$\Delta V = \Delta\hat{V}+e_1$$

where

$$\Delta\hat{V} = \int_{t_o}^{t_f} \bar{w}^T(t)\delta u(t)\,dt \tag{9.11}$$

and $|e_1| \leq c_1\varepsilon^2$

$$\bar{w}^T(t) \triangleq H_u(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t). \tag{9.12}$$

Proof: Follows from Proposition 4.2. □

PROPOSITION 9.2: Let H1A, H2A be satisfied, s = 2, $\bar{u}$, u ∈ G. Then, if $\max_{t\in T} \|\delta u(t)\| \leq \varepsilon$,

(i) $\|\delta y(t)\| \leq c_1\varepsilon$ $\quad c_1 < \infty$ for all t ∈ T

(ii) $\|\delta y(t)-\delta\hat{y}(t)\| \leq c_2\varepsilon$ $\quad c_2 < \infty$ for all t ∈ T

Proof:
$$\begin{aligned}\delta y(t) &= y(t)-\bar{y}(t)\\ &= S(x(t))-S(\bar{x}(t))\\ &= S_x(\bar{x}(t))\delta x + o(\delta x)\end{aligned}$$

$$\delta y(t)-\delta\hat{y}(t) = S_x(\bar{x}(t))(\delta x-\delta\hat{x})+o(\delta x)$$

Now, because our hypothesis, we have from Chapter I - III that

$$\|\delta x\| \leq c_3\varepsilon$$

$$\|\delta x-\delta\hat{x}\| \leq c_4\varepsilon^2 \qquad c_3, c_4 < \infty$$

Result follows. □

Let $y \in C^m$, the space of all continuous functions on T with norm

$$\|y\|_C \triangleq \max_{t \in T} \|y(t)\| . \tag{9.13}$$

The dual space is B^m, the space of m-dimensional functions of bounded variation, continuous from the right and usual norm $\|z^*\|_B$, so that $\langle z^*, y\rangle$ has the representation

$$\langle z^*, y\rangle = \int_{t_o}^{t_f} y^T(t)\, d\eta(t) \tag{9.14}$$

for some $\eta \in B^m$, and Stieltjies integration is implied. In the sequel we assume that Ω is convex.

Then we have from Luenberger [32]:

PROPOSITION 9.3: Let H1A, H2A be satisfied, $s = 2$, $\bar{u} \in G$. Let $(\bar{x}, \bar{u})$ be optimal. Then there exists a $r_o \geq 0$, $z^* \geq 0$ where $r_o \in R^1$, $z^* \in B^m$ and $|r_o| + \|z^*\|_B > 0$ s.t.

(i) $$r_o \Delta\hat{V} + \langle z^*, y\rangle \geq 0 \quad \text{for all } u \in G \tag{9.15}$$

(ii) $$\langle z^*, \bar{y}\rangle = 0 \tag{9.16}$$

(so that (i) becomes $r_o \Delta\hat{V} + \langle z^*, \delta y\rangle \geq 0$)

(iii) η, the representation of z^* is nondecreasing and is constant when $\bar{y} < 0$.

Proof: In the space $W = R^1 \times C^m$ define the sets

$A \triangleq \{(r,z): r \geq \Delta\hat{V},\ z \geq \hat{y} \text{ for some } u \in G\}$

$B \triangleq \{(r,z): r \leq 0,\ z \leq \theta\}$

where 0 denotes the null function.

A and B are convex sets

B has interior points as C^m has an interior

Then $A \cap \text{Int. } (B) = \phi$

Suppose this is false. Then there exists some u satisfying the hypothesis such that

$$\Delta V < 0, \quad \hat{y} < 0.$$

We have from proposition 9.1 that, for $\max_t \|\delta u(t)\| \leq \varepsilon$,

$$\Delta V = \Delta\hat{V} + e_1$$

where

$$\Delta\hat{V} = \int_{t_o}^{t_f} w^T \delta u \, dt, \quad |e_1| \leq c_1 \varepsilon^2$$

Define $u_1 \in G$ by

$$u_1 = (1-\alpha)\bar{u} + \alpha u \qquad 0 \leq \alpha \leq 1$$

giving

$$\delta u_1 = u_1 - \bar{u}(t) = \alpha \delta u$$

Then

$$\Delta\hat{V}_1 = \alpha \Delta\hat{V} < 0$$

and $|e_1| < L\varepsilon^2$ for α of the order of ε. This implies that

$$\Delta V = V^{u_1} - V^{\bar{u}} < 0 \tag{9.17}$$

If $\hat{y} < 0$, then there exists a sphere of radius ρ centred on $\hat{y}$ which is contained in N- the negative cone in C^m. For $0 < \alpha < 1$ the point $\alpha\hat{y}$ is the centre of a sphere of radius $\alpha\rho$ which is contained in N.

Then, because $\bar{y} = S(\bar{x}(t)) \leq 0$, $(1-\alpha) > 0$ $(1-\alpha)\bar{y}+\alpha\hat{y} = \bar{y}+\alpha\delta\hat{y}$ is the centre of a sphere of radius $\alpha\rho$ which is contained in N. For fixed δx,

$$\| S(\bar{x}+\alpha\delta\hat{x})-\bar{y}-\alpha\delta\hat{y}\| = O(\alpha)$$

Then, for α sufficiently small,

$$S(\bar{x}+\alpha\delta\hat{x}) < 0 \tag{9.18}$$

where $\alpha\delta\hat{x} = \alpha\int_{t_o}^{t_f} \Phi(t,\tau)B(\tau)\delta u(\tau)\,d\tau$

$$= \int_{t_o}^{t_f} \Phi(t,\tau)B(\tau)\delta u_1(\tau)\,d\tau$$

Then (9.17) and (9.18) contradict the optimality of $\bar{u}$.

So A ∩ Int (B) = ϕ and Int (B) ≠ ϕ

Therefore, there exists a closed hyperplane separating A and B. Hence there exists r_o, z^*, δ s.t.

$$r_o r+\langle z^*, z\rangle \geq \delta \quad \text{for all } (r,z) \in A$$

$$r_o r+\langle z^*, z\rangle \leq \delta \quad \text{for all } (r,z) \in B$$

As $(0,0) \in A \cap B$, $\delta = 0$

Setting $r = \Delta V$, $z = \hat{y}$ gives (i).

From the nature of B, $r_o \geq 0$, $z^* \geq 0$.

For $u \approx \bar{u}$, i.e. $\hat{y} = \bar{y}$ (i) implies $\langle z^*, \bar{y} \rangle \geq 0$.

But $z^* \geq 0$, $\bar{y} \leq 0$ implies $\langle z^*, \bar{y} \rangle \leq 0$.

Hence, $\langle z^*, \bar{y} \rangle = 0$.

$\langle z^*, z \rangle \leq 0$, for all $z \leq 0$ implies via equation (9.14) that η is nondecreasing. □

Let $\Phi(t,t_o)$ be the transition matrix associated with $A(t)$, defined in equ (9.7). Then, for $u \in G$:

$$\langle z^*, \delta y \rangle = \int_{t_o}^{t_f} \delta\hat{y}^T(t)\, d\eta(t)$$

$$= \int_{t_o}^{t_f} \delta\hat{x}^T(t) C^T(t)\, d\eta(t)$$

$$= \int_{t_o}^{t_f} \left(\int_{t_o}^{t} \delta u^T(\tau) B^T(\tau) \Phi^T(t,\tau)\, d\tau \right) C^T(t)\, d\eta(t)$$

$$= \int_{t_o}^{t_f} \left(\int_{\tau}^{t_f} \delta u^T(\tau) B^T(\tau) \Phi^T(t,\tau) C^T(t)\, d\eta(t) \right) d\tau$$

$$= \int_{t_o}^{t_f} \delta u^T(\tau) B^T(\tau) \lambda'(\tau)\, d\tau \qquad (9.19)$$

where

$$\lambda'(\tau) \triangleq \int_{\tau}^{t_f} \Phi^T(t,\tau) C^T(t)\, d\eta(t) \qquad (9.20)$$

Note that if η were differentiable, then λ' would satisfy the differential equation

$$-\dot{\lambda}'(t) = A^T(t)\lambda'(t)+C^T(t)\dot{\eta}(t)$$

but such a representation is not valid if η is bounded variation. However, we shall see that by assuming certain differentiability properties we can obtain such a representation.

We have that $\bar{\lambda}(t)$ is the solution to equs (2.16) and (2.17). Then

$$-\dot{\bar{\lambda}}(t) = A^T(t)\bar{\lambda}(t) \qquad \bar{\lambda}(t_f) = F_x(\bar{x}(t_f),t_f)$$

Define

$$\lambda(t) \triangleq r_o\bar{\lambda}(t)+\lambda'(t) \tag{9.21}$$

$$w(t) \triangleq B^T(t)\lambda(t) = H_u(\bar{x}(t),\bar{u}(t),\lambda(t)) \tag{9.22}$$

Then $w(t)$ is bounded variation, because $\lambda'(t)$ is bounded variation. Note that if η were differentiable, we could regard $w(t)$ as being the output of the following system

$$-\dot{\lambda}(t) = A^T(t)\lambda(t)+C^T(t)\dot{\eta}(t)$$
$$w(t) = B^T(t)\lambda(t)$$

If w is smooth (e.g. $w \equiv 0$) and this system is invertible we could deduce the smoothness of $\dot{\eta}$. This is what in effect will be done:

PROPOSITION 9.4: Let H1A, H2A be satisfied, $s = 2$. Let $\bar{u} \in G$. Let $(\bar{x},\bar{u})$ be optimal. Assume

(i) $\bar{u}(t) \in$ interior of Ω for all $t \in T$.

(ii) $r_o > 0$ (so we set it equal to unity)

Then

$$w(t) \equiv 0 \qquad \text{for all } t \in T.$$

Proof: Because of our hypothesis we have from Proposition 9.3 (setting $r_o = 1$), that there exists $z^* \geq 0$, $z^* \in B^m$ and $\|z^*\|_B + 1 > 0$ such that

$$\Delta\hat{V} + \langle z^*, \delta\hat{y}\rangle \geq 0$$

from (9.19) we have

$$0 \leq \Delta\hat{V} + \int_{t_o}^{t_f} \delta u^T(t) B^T(t) \lambda'(t)\, dt$$

$$= \int_{t_o}^{t_f} \delta u^T(t) [\bar{w}(t) + B^T(t)\lambda'(t)]\, dt$$

$$= \int_{t_o}^{t_f} \delta u^T(t) B^T(t) (\bar{\lambda}(t) + \lambda'(t))\, dt$$

$$= \int_{t_o}^{t_f} \delta u^T(t) B^T(t) \lambda(t)\, dt$$

$$= \int_{t_o}^{t_f} \delta u^T(t) w(t)\, dt$$

We have from (i) that $\bar{u} \in$ int (Ω), $\forall t \in T$. Therefore

$$\int_{t_o}^{t_f} v^T(t)w(t)dt = 0$$

for all $v \in G$. This implies that

$$w(t) \equiv 0 \qquad \text{a.e.} \quad t \in T$$

However, $w(t) \in B^m$, and is therefore continuous from the right. Therefore

$$w(t) \equiv 0 \quad \text{for all} \quad t \in T. \qquad \square$$

Define $B_o, \ldots B_{p-1}$, $C_o, \ldots C_{p-1}$ as follows:

$$B_o(t) \triangleq B(t) \tag{9.23}$$

$$B_{r+1}(t) \triangleq -\dot{B}_r(t) + A(t)B_r(t) \tag{9.24}$$

$$C_o(t) \triangleq C(t) \tag{9.25}$$

$$C_{r+1}(t) \triangleq \dot{C}_r(t) + C_r(t)A(t) \tag{9.26}$$

PROPOSITION 9.5: Assume C and B are p-1 times and A is p-2 times continuously differentiable. If C_rB_o is constant (zero) on T_i for $r = 0,1,..p-2$, then

$$C_{p-1}B_o \equiv C_{p-2}B_1 \equiv \ldots \equiv C_1B_{p-2} \equiv C_oB_{p-1} \text{ on } T_i \tag{9.27}$$

Proof: C_oB_o is constant by hypothesis.
Assume C_iB_j is constant for $i+j = n < p-2$.
Then $\dot{C}_iB_j + C_i\dot{B}_j \equiv 0$

i.e. $(C_{i+1}-C_iA)B_j \equiv C_i(B_{j+1}-AB_j)$

so $C_{i+1}B_j \equiv C_iB_{j+1}$

C_nB_o constant implies $C_{n+1}B_o \equiv C_nB_1$

$C_{n-1}B_1$ constant implies $C_nB_1 \equiv C_{n-1}B_2$

"

"

C_oB_n constant implies $C_1B_n \equiv C_oB_{n+1}$

i.e. $C_{n+1}B_o \equiv C_nB_1 \equiv \ldots \equiv C_oB_{n+1}$

i.e. C_iB_j *is constant for* $i+j = n+1$.

Hence, by induction C_iB_j is constant for all i,j such that $i+j \leq p-2$. But C_jB_j constant for $i+j = p-2$ implies (9.27).

□

Define $\left(\frac{x}{u}\right)(t) \triangleq \left(\frac{d}{dt}\right)^r \bar{u}(t)$ (9.28)

PROPOSITION 9.6: Let $(\bar{x},\bar{u})$ be optimal. Assume (without loss of generality) that $\bar{x}$ consists of one boundary and 2 interior trajectories with entry and exit time t_a and t_b respectively.

$T_i \triangleq (t_a,t_b)$. Then, let

(i) f be continuously differentiable up to $(p+1)^{th}$ order in x and u, S be p times continuously

differentiable in x.

(ii) $\left(\frac{F}{u}\right)$, r=1,...p be continuous on T_1.

(iii) $C_x B \equiv 0$ on T_1, r=0,..p-2.

$C_{p-1} B$ nonsingular on T_1.

(iv) $r_o > 0$ (so set $r_o = 1$).

(v) $\bar{u}(t) \in$ interior of Ω, for all $t \in T$.

Then a necessary condition for optimality of $(\bar{x},\bar{u})$ is

$$H_u(\bar{x}(t),\bar{u}(t),\lambda(t)) = 0 \quad \text{for all } t \in T \tag{9.29}$$

where

$$-\dot{\lambda}(t) = A^T(t)\lambda(t) \qquad t \in [t_o,t_a)\cup(t_b,t_f] \tag{9.30}$$

$$-\dot{\lambda}(t) = A^T(t)\lambda(t)+C^T(t)\dot{\eta}(t) \qquad t \in T_1$$

with

$$\lambda(t_f) = F_x(\bar{x}(t_f),t_f) \tag{9.31}$$

$$\lambda(t_\alpha^-) = \lambda(t_\alpha^+) + C^T(t_\alpha)\mu(t_\alpha) \; ; \quad \alpha = a,b. \tag{9.32}$$

$$\mu(t_\alpha) \geq 0$$

$$\dot{\eta}(t) = -([C(t)B_{p-1}(t)]^T)^{-1}B_p^T\lambda(t) \qquad t \in T_1 \tag{9.33}$$

$$= 0 \qquad \text{elsewhere.}$$

if also

(vi) $\bar{u}$ is continuous across t_a, t_b,

(vii) $H_{uu}(t_\alpha)$, $\alpha = a,b$, is positive definite,

then

$\left(\frac{x}{u}\right)$, $r = 0,..\ p-2$ is continuous across t_a, t_b and $\mu(t_\alpha)$, $\alpha = a, b$ is given by

$$0 = \left(B^{p-1}\right)(t_\alpha^-) - \left(B^{p-1}\right)(t_\alpha^+)]^T \lambda(t_\alpha^-) + (-1)^{p-1}[C(t_\alpha)B_{p-1}(t_\alpha)]^T \mu(t_\alpha) \tag{9.34}$$

Proof: Assume (i) - (v) are satisfied

(9.29) follows from proposition 9.4.

Also:

$$\begin{aligned} 0 &= B^T(t)\lambda(t) \\ &= B^T(t)(\bar{\lambda}(t) + \lambda'(t)) \\ &= B^T(t)\phi^T(t_b,t)\lambda(t_b^-) + \int_t^{t_b^-} B^T(t)\phi^T(\tau,t)C^T(\tau)d\eta(\tau). \end{aligned}$$

Integrating by parts gives

$$\begin{aligned} 0 &= B^T(t)\phi^T(t_b,t)\lambda(t_b^-) + B^T(t)\phi^T(\tau,t)C^T(\tau)\eta(\tau)\Big]_{\tau=t}^{\tau=t_b^-} \\ &\quad - \int_t^{t_b^-} B^T(t)\frac{d}{d\tau}(\phi^T(t,\tau)C^T(\tau))\eta(\tau)d\tau \\ &= -B^T(t)C^T(t)\eta(t) + B^T\phi^T(t_b,t)[\lambda(t_b^-) + C^T(t_b^-)\eta(t_b^-)] \\ &\quad - \int_t^{t_b^-} B^T(t)\phi^T(\tau,t)C_1^T(\tau)\eta(\tau)d\tau. \end{aligned}$$

We have from (iii), $CB \equiv 0$. Hence both sides of the above can be differentiated w.r.t. t, giving

$$0 = (\dot{B}^T\phi^T + B^T\dot{\phi}^T)[\lambda(t_b^-) + C^T(t_b^-)\eta(t_b^-)]$$

$$- \int_t^{t_b^-} [\dot{B}^T\phi^T + B^T\dot{\phi}^T]C_1^T(\tau)\eta(\tau)\,d\tau + B^T(t)C_1^T(t)\eta(t)$$

$$= B^T(t)C_1^T(t)\eta(t) - B_1^T(t)\phi^T(t_b,t)[\lambda(t_b^-) + C^T(t_b^-)\eta(t_b^-)]$$

$$+ \int_t^{t_b^-} B_1^T(t)\phi^T(\tau,t)C_1^T(\tau)\eta(\tau)\,d\tau. \qquad (9.35)$$

From (iii) $C_1B \equiv 0$ on T_1, so both sides of equ (9.35) can be differentiated, giving

$$0 = -B_1^T(t)C_1^T(t)\eta(t) + B_2^T(t)\phi^T(t_b,t)[\lambda(t_b^-) + C^T(t_b^-)\eta(t_b^-)]$$

$$- \int_t^{t_b^-} B_2^T(t)\phi^T(\tau,t)C_1^T(\tau)\eta(\tau)\,d\tau.$$

From (iii) $C_2(t)B(t) \equiv 0$ for all $t \in T$. So we have from Proposition 9.5 that $C_1(t)B_1(t) \equiv 0$. Proceeding iteratively as above, we have, using (iii)

$$0 = (-1)^p B_{p-2}^T(t)C_1^T(t)\eta(t) + (-1)^{p-1}B_{p-1}^T(t)\phi^T(t_b,t)[\lambda(t_b^-) + C^T(t_b^-)\eta(t_b^-)]$$

$$+ (-1)^p \int_t^{t_b} B_{p-1}^T(t)\phi^T(\tau,t)C_1^T(\tau)\eta(\tau)\,d\tau \qquad (9.36)$$

Now, from Proposition 9.5 if $C_{p-2}B \equiv 0$, then

$$C_{p-1}B \equiv C_{p-2}B_1 \equiv \cdots \equiv C_1B_{p-2} \equiv C_oB_{p-1}.$$

Therefore, from (iii) we have that $C_i B_{p-2}$ is non singular. Then, from (9.36) it can be seen that η is absolutely continuous for all $t \in T_1$.

Also, from (9.16) we have

$$\int_{t_o}^{t_f} S(\tilde{x}(t)) d\eta(t) = 0$$

Thus

$$\int_{t_o}^{t_a} S d\eta + \int_{t_a}^{t_b} S d\eta + \int_{t_o}^{t_f} S d\eta = 0$$

i.e. $$\int_{t_o}^{t_a} S d\eta + \int_{t_b}^{t_f} S d\eta = 0$$

So η is constant on $[t_o, t_a]$ and $(t_b, t_f]$, since $S < 0$ in these intervals and η is nondecreasing.

Hence λ is the solution of

$$-\dot{\lambda}(t) = A^T(t)\lambda(t) \qquad t \in [t_o, t_a) \cup (t_b, t_f]$$

$$-\dot{\lambda}(t) = A^T(t)\lambda(t) + C^T(t)\dot{\eta}(t) \qquad t \in T_1$$

with possible jumps of the form

$$\lambda(t_\alpha^-) = \lambda(t_\alpha^+) + C^T(t_\alpha)\mu(t_\alpha) \qquad \alpha = a.b.$$

where $\mu(t_\alpha) > 0$, because η is nondecreasing.

We now derive the representation for $\dot{\eta}(t)$ given in equation (9.33). Certainly $\dot{\eta}(t) = 0$ on $[t_o, t_a) \cup (t_b, t_f]$, because we have from above that η is constant over these intervals.

Again

$w(t) = B^T(t)\lambda(t)$ where $w(t) \equiv 0$.

Differentiating the above w.r.t. t gives

$$0 = \frac{d}{dt}[B^T(t)\lambda(t)] = -B_1^T(t)\lambda(t) - [C(t)B(t)]^T\dot{\eta}(t)$$

where $\dot{\eta}(t) = 0 \quad t \in T \setminus T_1$.

Now $CB \equiv 0$ on T_1, and differentiation gives

$$0 = B_2^T(t)\lambda(t) + [C(t)B_1(t)]^T\dot{\eta}(t).$$

From (iii) and Proposition 9.5, we have $CB_1 \equiv C_1B \not\equiv 0$. Proceeding iteratively as above, we have, using $C_rB \equiv 0$ $r = 0,..p-2$

$$0 = (-1)^p(B_p^T(t)\lambda(t) + [C(t)B_{p-1}(t)]^T\dot{\eta}(t)).$$

From (iii) $C_{p-1}B$ is nonsingular on T_1, i.e. from equ (9.27) CB_{p-1} is nonsingular on T_1. Therefore

$$\dot{\eta}(t) = -([C(t)B_{p-1}(t)]^T)^{-1}B_p^T(t)\lambda(t).$$

We are now left with the problem of finding the possible jumps at t_a, t_b. Consider t_a. Denote $H(\bar{x}(t_a), u(t_a^-), \lambda(t_a^-))$ by $H(t_a^-)$. Assumptions (vi), (vii) are now also assumed to hold. We have from (9.29)

$$\begin{aligned} 0 &= H_u(t_a^-) - H_u(t_a^+) \\ &= B^T(t_a^-)\lambda(t_a^-) - B^T(t_a^+)\lambda(t_a^+) \\ &= [B(t_a^-) - B(t_a^+)]^T\lambda(t_a^-) + B^T(t_a^+)[\lambda(t_a^-) - \lambda(t_a^+)] \end{aligned} \qquad (9.37)$$

where

$$\lambda(t_a^-) - \lambda(t_a^+) = C^T(t_a)\mu(t_a), \quad \text{where} \quad \mu(t_a) \geq 0. \tag{9.38}$$

We cannot substitute 9.38 into 9.37 to give us $\mu(t_a)$, because $CB \equiv 0$. Therefore, evaluating $\dot{H}_u(t_a^-) - \dot{H}_u(t_a^+)$ gives

$$\begin{aligned}
0 &= \dot{H}_u(t_a^-) - \dot{H}_u(t_a^+) \\
&= [\dot{B}(t_a^-) - \dot{B}(t_a^+)]^T\lambda(t_a^-) + [B(t_a^-) - B(t_a^+)]^T\dot{\lambda}(t_a^-) \\
&\quad + \dot{B}^T(t_a^+)[\lambda(t_a^-) - \lambda(t_a^+)] + B^T(t_a^+)[\dot{\lambda}(t_a^-) - \dot{\lambda}(t_a^+)] \\
&= [\dot{B}(t_a^-) - \dot{B}(t_a^+)]^T\lambda(t_a^-) + \dot{B}^T(t_a^+)[\lambda(t_a^-) - \lambda(t_a^+)] \\
&\quad + B^T(t_a^-)\dot{\lambda}(t_a^-) - B^T(t_a^+)\dot{\lambda}(t_a^+) \\
&= [\dot{B}(t_a^-) - \dot{B}(t_a^+)]^T\lambda(t_a^-) + \dot{B}^T(t_a^+)[\lambda(t_a^-) - \lambda(t_a^+)] \\
&\quad - B^T(t_a^-)A^T(t_a^-)\lambda(t_a^-) - B^T(t_a^+)[-A^T(t_a^+)\lambda(t_a^+) - C^T(t_a^+)\dot{\eta}(t_a^+)] \\
&= [\dot{B}(t_a^-) - \dot{B}(t_a^+)]^T\lambda(t_a^-) + \dot{B}^T(t_a^+)[\lambda(t_a^-) - \lambda(t_a^+)] \\
&\quad - B^T(t_a^+)A^T(t_a^+)[\lambda(t_a^-) - \lambda(t_a^+)] \quad \text{from (i), the continuity} \\
&\qquad\qquad \text{of } \tilde{u} \text{ across } t_a,\ CB = 0 \text{ on } T_1. \\
&= [\dot{B}(t_a^-) - \dot{B}(t_a^+)]^T\lambda(t_a^-) - B_1^T(t_a^+)[\lambda(t_a^-) - \lambda(t_a^+)]
\end{aligned} \tag{9.39}$$

Substituting (9.38) into (9.39) will not give $\mu(t_a)$ because $CB_1 \equiv C_1 B \equiv 0$ on T_1. We also have

$$[\dot{B}(t_a^-) - \dot{B}(t_a^+)]^T \lambda(t_a^-) = [\dot{\bar{u}}(t_a^-) - \dot{\bar{u}}(t_a^+)] H_{uu}(t_a^-)$$

Hence, positive definiteness of $H_{uu}(t_a^-)$ implies that $\dot{\bar{u}}$ is continuous across t_a. Therefore, evaluating $\ddot{H}_u(t_a^-) - \ddot{H}_u(t_a^+)$ and making use of the continuity of $\dot{\bar{u}}$ across t_a, $CB_1 \equiv 0$, gives:

$$0 = [\ddot{B}(t_a^-) - \ddot{B}(t_a^+)]^T \lambda(t_a^-) + B_2^T(t_a^+)[\lambda(t_a^-) - \lambda(t_a^+)]$$

we also have

$$[\ddot{B}(t_a^-) - \ddot{B}(t_a^+)]^T \lambda(t_a^-) = [\ddot{\bar{u}}(t_a^-) - \ddot{\bar{u}}(t_a^+)] H_{uu}(t_a^-)$$

and, by (vii), we have that $\ddot{\bar{u}}$ is continuous across t_a. Proceeding iteratively, we have, *because* of our assumptions, that

$(\overset{r}{\bar{u}})(t)$ is *continuous* across t_a, r = 0,..p-2, and

$$0 = [(\overset{p-1}{B})(t_a^-) - (\overset{p-1}{B})(t_a^+)]^T \lambda(t_a^-)$$

$$+ (-1)^{p-1} B_{p-1}^T(t_a^+)[\lambda(t_a^-) - \lambda(t_a^+)] \qquad (9.40)$$

Substituting (9.38) into (9.40) gives (9.34). Similar results hold for t_b. □

We now show that Assumption (iv) of the above Proposition is unnecessary as it is implied by the other assumptions.

PROPOSITION 9.7: Let assumptions (i) - (iii), (v) - (vii) of Proposition 9.6 be satisfied. Then

$$r_o > 0 \quad \text{(and can be set equal to unity)}.$$

Proof: Assume $r_o = 0$.

Then, from equation (9.21) we have

$$\lambda(t) = r_o\bar{\lambda}(t) + \lambda'(t).$$

Thus

$$\lambda(t_f) = r_o F_x(\bar{x}(t_f), t_f).$$
$$= 0.$$

Assume, with no loss of generality, that we have only one boundary arc with entry and exit times t_a, t_b respectively. Therefore, from equ (9.30)

$$\lambda(t_b^+) = 0$$

Therefore, from equ (9.34) (because $CB_{p-1} \neq 0$)

$$\mu(t_b) = 0$$

so

$$\lambda(t_b^-) = 0$$

Hence, $\lambda(\cdot) = 0$ along the boundary. Thus (9.33)

$$\dot{\pi}(t) = 0 \quad \text{and} \quad r_o = 0.$$

But $r_o = 0$, $z^* = 0$, $\mu = 0$ contradicts Proposition 9.3 that there exists a nontrivial separating hyperplane. Thus $r_o \neq 0$, and from Proposition 9.3, $r_o > 0$. □

We now relate the properties of C_rB to S.

Let

$$(\dot{S}) \triangleq S_x(x)f(x,u,t)$$

i.e. $(\dot{S})(\bar{x}(t),\bar{u}(t)) = {}^{d}/_{dt}S(\bar{x}(t))$

and

$$(\overset{1}{S}) \triangleq (\dot{S})$$

$$(\overset{r}{S}) \triangleq (\overset{r-1}{S}{}^{\cdot})$$

$$(\overset{r}{S})_x \triangleq \frac{\partial}{\partial x}(\overset{r}{S})(x)$$

PROPOSITION 9.8: If S is p times continuously differentiable, then, for $r = 0,..p$

$$(\overset{r}{S})_x(\bar{x}(t),\bar{u}(t)) = C_r(t) \tag{9.41}$$

$$(\overset{r}{S})_u(\bar{x}(t),\bar{u}(t)) = C_{r-1}(t)B_0(t) + ((\overset{r-1}{S})_x f)(\bar{x}(t),\bar{u}(t)). \tag{9.42}$$

Proof: $(S_x)(\bar{x}(t),\bar{u}(t)) = C(t) = C_0(t)$ from (9.9), (9.25).

Suppose

$$(\overset{r}{S})_x(\bar{x}(t),\bar{u}(t)) = C_r(t)$$

Then

$$\begin{aligned}(\overset{r+1}{S})_x &= ((\overset{r}{S})_x f)_x \\ &= ((\overset{r}{S})_{xx} f) + (\overset{r}{S})_x f_x \\ &= (\overset{\cdot}{\overset{r}{S}})_x + (\overset{r}{S})_x f_x \\ &= \dot{C}_r(t) + C_r(t)A(t) = C_{r+1}(t)\end{aligned}$$

Hence, by induction we obtain (9.41). Also

$$\begin{aligned}(\overset{r}{S})_u &= ((\overset{r-1}{S})_x f)_u \\ &= (\overset{r-1}{S})_{xu} f + (\overset{r-1}{S})_x f_u\end{aligned}$$

which gives (9.42). □

COROLLARY 9.9: Let S be p times continuously differentiable. Then, if $(\overset{r}{S})(\bar{x}(t),\bar{u}(t))$ is independent of u, r = 0,..p-1, then

$$\begin{aligned}(\overset{p}{S})_u(\bar{x}(t),\bar{u}(t)) &= C_{p-1}(t)B_0(t) \qquad (9.43)\\ &= C_0(t)B_{p-1}(t).\end{aligned}$$

Proof: From equ (9.42) □

Combining the results of Proposition 9.6, 9.7, 9.9 gives us the necessary conditions of Jacobson, Lele and Speyer [26]:

PROPOSITION 9.10: If assumptions (i), (ii), (iii), (v) of Proposition 9.6 are satisfied, then necessary conditions for $(\bar{x},\bar{u})$ to be optimal for the state constrained problem described at the beginning of the Chapter are:

$$\frac{\partial H}{\partial u} = 0 = f_u^T \lambda$$

where the $\lambda(\cdot)$ are given by

$$-\dot{\lambda} = \frac{\partial H}{\partial x}$$

$$= f_x^T \lambda + S_x^T \hat{\eta}$$

with

$$\lambda(t_f) = F_x(\bar{x}(t_f), t_f)$$

and

$$\hat{\eta} \geq 0 \qquad S(\bar{x}(t)) = 0$$

$$= 0 \qquad S(\bar{x}(t)) < 0$$

is a bounded function for $t \in [t_o, t_f]$.

We also have, from [3], [48], that

$$H(t_a^-) = H(t_a^+).$$

Definition: The Hamiltonian H is said to be regular if along a given $x(t)$, $\lambda(t)$ trajectory (say $\bar{x}(t)$, $\bar{\lambda}(t)$), $H(\bar{x},u,\bar{\lambda})$ has an unique minimum in u, $t \in T$.

For the case of a regular Hamiltonian (so condition (vii) of Proposition 9.6 holds), Speyer [48] and McIntyre and Paiewonsky [39] have shown that $\bar{u}$ must be continuous across the junction, i.e. condition (vi) of Proposition 9.6 is a necessary condition for $(\bar{x},\bar{u})$ to be optimal.

Then, for the case p = 1 and H regular we have, from the continuity of $\bar{u}$ that $\mu(t_a) = 0$. i.e. there is no jump in λ at the junction points.

Jacobson et al relate their results [26] to those of [3],[48]. Denham and Bryson [5], used the results of [3] for a steepest ascent algorithm, while Speyer [48] proposed a second order sweep algorithm. Other important computational techniques are of the penalty function type.

CHAPTER 10

SINGULAR PROBLEMS

Consider the control problem as originally defined. Let $(\bar{x},\bar{u})$ be optimal, and assume that $H(\bar{x}(t),u,\bar{\lambda}(t),t)$ (where $\bar{\lambda}(t)$ is the solution of equs (2.16) and (2.17)), is independent of u for all $t \in T$. Obviously, the estimate $\int_{t_o}^{t_f} \Delta H(t)dt$ of ΔV is zero, and therefore of no use. However, Proposition 4.2 can be used to give a necessary condition of optimality (Mayne [36]).

PROPOSITION 10.1: Let $\bar{u} \in G$. If either

(i) H1A, H2A are satisfied, $s=3$ and $d(u,\bar{u}) \leq \varepsilon$, or

(ii) H1, H2 are satisfied, $s=3$ and $d_1(u,\bar{u}) \leq \varepsilon$ and if $(\bar{x},\bar{u})$ are optimal, $H(\bar{x}(t),u,\bar{\lambda}(t),t)$ is independent of u for all $t \in T$, then

$$\phi(t) \stackrel{\Delta}{=} \Delta f^T(t)\Delta H_x(t) + \Delta f^T(t)\bar{P}(t)\Delta f(t) \geq 0 \qquad (10.1)$$

for all $u \in \Omega$, all $t \in T$, where

$\Delta f(t) = f(\bar{x}(t),u,t) - f(\bar{x}(t),\bar{u}(t),t)$

$\Delta H_x(t) = H_x(\bar{x}(t),u,\bar{\lambda}(t),t) - H_x(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t)$

and $\bar{\lambda}(t)$, $\bar{P}(t)$ are the solutions to equs (2.16)-(2.19).

Proof: Assume the inequality 10.1 is violated at $t_1 \in \theta(\bar{u})$, for $u = v$. Define the control $u(t)$ by

$$u(t) = \bar{u}(t), \quad t \notin T_\varepsilon \triangleq [t_1-\varepsilon, t_1]$$

$$u(t) = \bar{u}(t) + v, \quad t \in T_\varepsilon.$$

Hence, for $t \in T_\varepsilon$

$$\delta\hat{x}(t) = \Delta f(t^*) \qquad t \in [t_1-\varepsilon, t_1]$$

where

$$t^*(t) \in [t_1-\varepsilon, t].$$

Since $\phi(t_1) \leq C < 0$, there exists an $\varepsilon = \varepsilon_1 > 0$ s.t. $\Delta f^T(t^*(t))[\Delta H_x(t) + \bar{P}(t)\Delta f(t)] < C/2 < 0$ for all $t \in T_\varepsilon$.

Hence, from Proposition 4.2, there exists an $\varepsilon \leq \varepsilon_1$ such that $\Delta V < 0$. This contradicts optimality. □

The above result is a strong version of the condition of optimality given by Jacobson [19].

To proceed further (in the present state of knowledge) we have to impose the further restriction that f and L are linear in u or that weak variations only are permitted.

Thus under conditions of Proposition 4.5, and the independence of H w.r.t. u, we easily obtain a weak version of Corollary 4.3:

$$\Delta\hat{V} = \int_{t_0}^{t_f} v^T(t)[C(t) + B^T(t)\bar{P}(t)]z(t)dt \qquad (10.2)$$

$$|\Delta V - \Delta\hat{V}| \leq c\varepsilon^3, \quad c < \infty$$

where

$$C(t) \triangleq H_{ux}(\bar{x}(t),\bar{u}(t),\bar{\lambda}(t),t) \tag{10.3}$$

$$A(t) \triangleq f_x(\bar{x}(t),\bar{u}(t),t) \tag{10.4}$$

$$B(t) \triangleq f_u(\bar{x}(t),\bar{u}(t),t) \tag{10.5}$$

where $\bar{\lambda}(t)$ is the solution to equations (2.16), (2.17), and $z(t)$ replaces $\delta\hat{x}$, $v(t)$ replaces δu. i.e.

$$\dot{z}(t) = A(t)z(t) + B(t)v(t) \tag{10.6}$$

$$z(t_o) = 0 \tag{10.7}$$

Define

$$\bar{c}(t) \triangleq c(t) + B^T(t)\bar{F}(t) \tag{10.8}$$

$$y(t) \triangleq \bar{c}(t)z(t) \tag{10.9}$$

Then (10.2) becomes

$$\Delta\hat{V} = \int_{t_o}^{t_f} v^T(t)y(t)\,dt \tag{10.10}$$

One way of obtaining sufficient conditions for the nonnegativity of $\Delta\hat{V}$ has been demonstrated by Jacobson [20]. Add to the integrand of equ (10.2)

$$\tfrac{1}{2}z^T(t)\bar{F}(t)[A(t)z(t) + B(t)v(t) - \dot{z}(t)]$$

where

$$-\dot{\tilde{P}}(t) = A^T(t)\tilde{P}(t) + \tilde{P}(t)A(t) - \tilde{Q}(t) \tag{10.11}$$

$$\tilde{P}(t_f) = P_f \tag{10.12}$$

so

$$\begin{aligned}\Delta\hat{V} &= \int_{t_o}^{t_f} [v^T(t)\tilde{C}(t)z(t) + \tfrac{1}{2}z^T(t)\tilde{P}(t)[A(t)z(t) \\ &\qquad + B(t)v(t) - \dot{z}(t)]]\,dt \\ &= \int_{t_o}^{t_f} (v^T(t)\tilde{C}(t)z(t) + \tfrac{1}{2}z^T(t)\tilde{P}(t)A(t)z(t) \\ &\quad + \tfrac{1}{2}z^T(t)\tilde{P}(t)B(t)v(t) + \tfrac{1}{2}z^T(t)\tilde{P}(t)z(t) + \tfrac{1}{2}z^T(t)\dot{\tilde{P}}(t)z(t))\,dt \\ &\quad - \tfrac{1}{2}z^T(t_f)\tilde{P}_f z(t_f) \\ &= \int_{t_o}^{t_f} (\tfrac{1}{2}z^T(t)\tilde{Q}(t)z(t) + v^T(t)[\tilde{C}(t) + B^T(t)\tilde{P}(t)]z(t))\,dt \\ &\quad - \tfrac{1}{2}z^T(t_f)\tilde{P}_f z(t_f)\end{aligned} \tag{10.13}$$

Clearly if

(i) there exists $\tilde{P}$ satisfying (10.11) and boundary condition $\tilde{P}_f \geq 0$, and

(ii) $\tilde{C}(t) + B^T(t)\tilde{P}(t) \equiv 0 \qquad \forall t \in T$ (10.14)

(iii) $\tilde{Q}(t) \geq 0 \qquad \forall t \in T$ (10.15)

then $\Delta\hat{V} \geq 0$ on T.

Another way of obtaining sufficient conditions for the nonnegativity of $\Delta\hat{V}$ has been demonstrated by Jacobson [22].

Add to the integrand of (10.2)

$$\tfrac{1}{2}z^T(t)\tilde{P}(t)[A(t)z(t) + B(t)v(t) - \dot{z}(t)]$$

where, now, $\tilde{P}(t)$ is any nxn symmetric, continuously differentiable, matrix function of time. So

$$\Delta\hat{V} = \int_{t_o}^{t_f} (\tfrac{1}{2}z^T[\dot{\tilde{P}}(t) + \tilde{P}(t)A(t) + A^T(t)\tilde{P}(t)]z(t) + v^T[\tilde{C}(t) + B^T(t)P(t)]z(t))dt - \tfrac{1}{2}z^T(t_f)\tilde{P}_f z(t_f)$$

Then, if

(i) $\tilde{C}(t) + B^T(t)\tilde{P}(t) \equiv 0 \ \forall t \in T$ (10.16)

(ii) $\dot{\tilde{P}}(t) + \tilde{P}(t)A(t) + A^T(t)\tilde{P}(t) \geq 0$ (10.17)

(iii) $\tilde{P}(t_f) \leq 0$ (10.18)

then $\Delta\hat{V} \geq 0$ on T.

In [25], Jacobson and Speyer obtain an 'integrated' *version of the above conditions* which allows for possible nondifferentiability of $\tilde{P}$. Necessary conditions are also obtained, where the gap between the necessary and sufficient conditions is minimal.

We now show how (10.2) can be used to obtain necessary conditions for the nonnegativity of $\Delta\hat{V}$. We will restrict attention to the case where f is linear in u, $L_{uu} \equiv 0$.

Let $\eta(t)$ denote the adjoint variable for the linearized system (10.6), (10.7)

$$-\dot{\eta}(t) = A^T(t)\eta(t) + \bar{C}^T(t)v(t) \qquad (10.19)$$

$$\eta(t_f) = 0$$

PROPOSITION 10.2: Let H1, H2 be satisfied, ś=2. Let $\bar{u}, u \in G$. Then

(i) $$-\dot{\lambda}(t) = A^T(t)\lambda(t) + \bar{C}^T(t)\delta u(t)$$

$$\lambda(t_f) = F_x(\bar{x}(t_f), t_f)$$

(ii) $$\eta(t) = \lambda(t) - \bar{\lambda}(t)$$

where $\bar{\lambda}(t)$ is the solution to equations (2.16), (2.17).

Proof: (i) From Chapter 2

$$-\dot{\lambda}(t) = H_x(\bar{x}(t), u(t), \lambda(t), t) + P(t)\Delta f(t)$$

$$= f_x^T(\bar{x}(t), u(t), t)\lambda(t) + P(t)B(t)\delta u(t)$$

$$= A^T(t)\lambda(t) + \bar{C}^T(t)\delta u(t).$$

(ii) $$-(\dot{\lambda}(t) - \dot{\bar{\lambda}}(t)) = A^T(t)\lambda(t) + \bar{C}^T(t)\delta u(t) - A^T(t)\bar{\lambda}(t)$$

$$= A^T(t)(\lambda(t) - \bar{\lambda}(t)) + \bar{C}^T(t)\delta u(t)$$

$$\lambda(t_f) - \bar{\lambda}(t_f) = 0.$$

Hence result. □

Define

$$H(z,v,\eta,t) \triangleq v^T\bar{C}(t)z + \eta^T(t)[A(t)z(t) + B(t)v(t)]$$

$$H_u(t) \triangleq H_u(\bar{z},v,\bar{\eta},t)$$

$$= \bar{C}(t)\bar{z}(t) + B^T(t)\bar{\eta}(t) \qquad (10.20)$$

where $\bar{z}(t)$ and $\bar{\eta}(t)$ are zero on T.

Define (c.f. Chapter 9)

$$B_0(t) = B(t) \qquad (10.21)$$

$$B_{r+1}(t) = -\dot{B}_r(t) + A(t)B_r(t) \qquad (10.22)$$

$$C_0(t) = \bar{C}(t) \qquad (10.23)$$

$$C_{r+1}(t) = \dot{C}_r(t) + C_r(t)A(t) \qquad (10.24)$$

PROPOSITION 10.3: If $\overset{r}{(C)}(t)$, $\overset{r}{(B)}(t)$ exist for all $t \in T$, $r=0,\ldots p-1$ and

$$[C_{r-1}(t)B_0(t) + (-1)^r C_0(t)B_{r-1}(t)] \equiv 0 \quad r=1,..p-1$$

then

$$\overset{p}{(H_u)}(t) = [C_{p-1}(t)B_0(t) + (-1)^p[C_0(t)B_{p-1}(t)]^T]v(t) \qquad (10.25)$$

Proof: Recall that

$$\overset{r}{(C)}(t) \triangleq (d/dt)^r C(t) \quad \text{etc.}$$

From (10.20)

$$\dot{H}_u(t) = C_1(t)\bar{z}(t) - B_1^T(t)\bar{\eta}(t)$$

$$+[C_0(t)B_0(t) - [C_0(t)B_0(t)]^T]v(t).$$

From our assumptions the term in square brackets is equal to 0. Therefore

$$\ddot{H}_u(t) = C_2(t)\bar{z}(t) + B_2^T(t)\bar{\eta}(t)$$

$$+[C_1(t)B_0(t) - [C_0(t)B_1(t)]^T]v(t).$$

Using our hypothesis again and proceeding iteratively we obtain (10.25) □

Define

$$\Phi_{ij}(t) \triangleq C_i(t)B_j(t) + (-1)^{i+j+1}B_i^T(t)C_j^T(t) \tag{10.26}$$

Then, provided the assumptions of Proposition 10.3 are satisfied, we have from (10.25), (10.26)

$$[(H)_u^r(t)]_u = \Phi_{r-1,0}(t) \tag{10.27}$$

Also, define

$$R_r(t) \triangleq (-1)^r[C_{r-1}(t)B_r(t) + B_r^T(t)C_{r-1}^T(t) + \frac{d}{dt}[C_{r-1}(t)B_{r-1}(t)]$$

$$= (-1)^r[C_r(t)B_{r-1}(t) + B_r^T(t)C_{r-1}^T(t)]$$

$$= (-1)^r\Phi_{r,r-1}(t) \tag{10.28}$$

$$S_r(t) \triangleq (-1)^r[C_{r-1}(t)B_{r-1}(t) - B_{r-1}^T(t)C_{r-1}^T(t)]$$

$$= (-1)^r \Phi_{r-1,T-1}(t) \qquad (10.29)$$

Making use of the defining property of C_r, B_r, $r = 0,1,2..$ we have:

PROPOSITION 10.4: If $\Phi_{r,o}$ is constant on T, r=0,..p-1, then

$$\Phi_{p,o} \equiv \Phi_{p-1,1} \equiv \quad \cdots \quad \equiv \Phi_{1,p-1} \equiv \Phi_{o,p} \quad \text{on T}$$

Proof: $\Phi_{o,o}$ is constant by hypothesis.

Assume $\Phi_{i,j}$ is constant for $i+j = n < p-1$

Φ_{ij} constant, implies, by (10.26)

$$\dot{C}_iB_j + C_i\dot{B}_j + (-1)^{i+j+1}\dot{B}_i^TC_j^T + (-1)^{i+j+1}B_i^T\dot{C}_j = 0$$

i.e. $(C_{i+1} - C_iA)B_j + C_i(-B_{j+1} + AB_j)$

$$+ (-1)^{i+j+1}(-B_{i+1} + AB_i)^TC_j^T + (-1)^{i+j+1}B_i^T(C_{j+1} - C_jA)^T = 0$$

so

$$C_{i+1}B_j - C_iB_{j+1} + (-1)^{i+j+1}(-B_{i+1}^T)C_j^T + (-1)^{i+j+1}B_i^TC_{j+1}^T = 0$$

i.e. $$\Phi_{i+1,j} = \Phi_{i,j+1}$$

Continuing as in Proposition 9.5, we obtain the result. □

We now consider transforming the problem defined by equations (10.2) (10.6) as follows:

$$\dot{v}_1(t) \triangleq \quad (t) \qquad v_1(t_o) = 0$$

$$\dot{v}_r(t) \triangleq \quad _{r-1}(t) \qquad v_r(t_o) = 0$$

$$z_r(t) \triangleq z_{r-1}(t) - B_{r-1}(t)v_r(t)$$

$$z_o(t) = z(t) \qquad B_o(t) = B(t)$$

The above transformation equations are due to Goh [40], [41] and are used to transform the singular problem into a nonsingular one, as will be shown below.

PROPOSITION 10.5: Let $(\bar{x},\bar{u})$ be optimal.

If $[(\overset{r}{H}_u)(t)]_{\bar{u}} = 0$ on T, r=0,..2p-2, and $S_r(t)$ and $R_r(t)$ are identically zero on T, r=1,..p-1. Then

(i) $S_p(t) \equiv 0$ on T

(ii) $R_p(t) \geq 0$ on T

are 2 necessary conditions for the nonnegativity of $\Delta\hat{V}$.

Proof: Using the transformation given above for r=1, the system equations becomes:

$$\dot{z}_1(t) = A(t)z(t) + B_1(t)v_1(t) \tag{10.30}$$

and $\Delta\hat{V}$ becomes:

$$\Delta\hat{V} = \Delta\hat{V}_1 = \int_{t_o}^{t_f} v_1^T C(t)[z_1(t) + B_o(t)v_1(t)]dt$$

Integrating by parts to rearrange the terms in $\dot{v}_1$ we obtain

$$\Delta\hat{V}_1 = \int_{t_1}^{t_f} (v_1^T[-C_1(t)]z_1(t) - \tfrac{1}{2}v_1^T[C_o(t)B_o(t) - B_o^T(t)C_o^T(t)]\dot{v}_1(t)$$

$$- \tfrac{1}{2}v_1^T(t)[C_o(t)B_1(t) + B_1^T(t)C_o(t) + \frac{d}{dt}[C_o(t)B_o(t)]]dt$$

$$+ v_1^T(t_f)C_o(t_f)z_1(t_f) + \tfrac{1}{2}v_1^T(t_f)C_o(t_f)B(t_f)v_1(t_f)$$

$$= \int_{t_1}^{t_f} (v_1^T[-C_1(t)]z_1(t) + \tfrac{1}{2}v_1^T R_1(t)v_1(t)$$

$$+ \tfrac{1}{2}v_1^T S_1(t)\dot{v}_1(t))dt$$

$$+ v_1^T(t_f)C_o(t_f)z_1(t_f) + \tfrac{1}{2}v_1^T(t_f)C_o(t_f)B(t_f)v_1(t_f) \qquad (10.31)$$

From our hypothesis $S_1(t)$, $R_1(t)$ are zero. By repeatedly applying the transformation and, and the fact that $R_r(t) \equiv 0$, $S_r(t) \equiv 0$, r=1,..p-1, then

$$\Delta\hat{V} = \Delta\hat{V}_p = \int_{t_o}^{t_f} (v_p^T(t)[(-1)^p C_p(t)]z_p(t) + \tfrac{1}{2}v_p^T R_p(t)v_p(t)$$

$$+ \tfrac{1}{2}v_p^T S_p(t)\dot{v}_p(t))dt$$

$$+ \sum_{r=1}^{p} (v_r^T(t_f)C_{r-1}(t_f)z_r(t_f)$$

$$+ \tfrac{1}{2}v_r^T(t_f)C_{r-1}(t_f)B_{r-1}(t_f)v_r(t_f)](-1)^{r-1}$$

Now, from (10.29), $S_p(t)$ is antisymmetric, and if it is nonzero, an appropriate control v_p can be found such that the term involving $\dot{v}_p$ dominates the other quantities, giving $\Delta\hat{V} < 0$.

Hence $S_p(t) \equiv 0$ on T is a necessary condtion for the non negativity of $\Delta\hat{V}$.

Given $S_p(t) \not\equiv 0$ on T, if $R_p(t) \not\equiv 0$, then a control, zero except on $[t_i - \epsilon_i, t_i]$, can be found such that $\Delta\hat{V} < 0$. □

PROPOSITION 10.6: Let A, $\bar{C}$, B be 2p times continuously differentiable. Then

(i) if $[(\overset{r}{H_u})(t)]_u \equiv 0$ on T $r=1,..,2p-2$,

$$S_p(t) = (-1)^p [(\overset{2p-1}{R_u})(t)]_u \quad \text{on } T \tag{10.32}$$

(ii) if $[(\overset{r}{H_u})(t)]_u \equiv 0$ on T $r=1,..2p-1$,

$$R_p(t) = (-1)^p [(\overset{2p}{H_u})(t)]_u \quad \text{on } T \tag{10.33}$$

Proof: (i) $[(\overset{r}{H_u})(t)]_u = 0$, $r=1,..2p-2$ implies that

$$\Phi_{r,o} \equiv 0 \qquad r=0,\ldots 2p-3.$$

From (10.29)

$$S_p(t) = (-1)^p \phi_{p-1,p-1}(t)$$

$$= (-1)^p \phi_{2p-2,0}(t) \quad \text{using Proposition 10.4}$$

$$= (-1)^p [(H_u^{2p-1})(t)]_u \quad \text{from (10.27).}$$

(ii) Is similarly proved. □

We see from Proposition 10.5, 10.6 that

$$(-1)^p [H_u^{2p}]_u \geq 0 \tag{10.34}$$

is a necessary condition for optimality. Kelley [29] obtained a new necessary condition (generalized Legendre-Uebsch) for non-negativity of the singular second variation. The general form of this condition (10.34) was obtained (apparently independently) by Kelley et al [31], Robbins [47] and Tait [50].

It was shown ([31], pg 75) that the control u cannot appear in an odd time derivative of H_u if u is a scalar.

The generalized Legendre-Clebsch condition for the case of vector controls was obtained by Robbins [47] and Goh [11], which took the form of equ (10.34) and

$$(-1)^p [H_u^{2p-1}]_u = 0 \qquad \forall t \in T$$

(see Propositions 10.5 (i), 10.6 (i)). Note that for the case p=1, we have that

$$S_1(t) = (-1)^K C_o B_o - B_o^T C_o(t) = 0$$

i.e. $\quad \bar{C}B = B^T\bar{C}^T$

$$(C+B^T\bar{F})B = B^T(C+B^T\bar{F})^T$$

so $\quad CB = B^TC^T.$

So CB is symmetric (see Jacobson [23], Robbins [47], Goh [10]).

Jacobson [21] showed by means of a counter example that satisfaction of the generalized Legendre-Uebsch necessary conditions and the necessary conditions of [19] are, in general, insufficient for optimality in singular problems.

We now make use of these results in obtaining sufficient conditions of optimality.

Consider the case when the singular control is of first order, and $R_1(t) > 0$, $S_1(t) = 0 \quad \forall t \in T$. We use the system and cost function defined by equations (10.30), (10.31) to obtain a control law, and assess the performance of this law.

Applying linear optimal control theory yields

$$v_1(t) = -K(t)z_1(t) \tag{10.35}$$

$$K(t) = R_1^{-1}(t)[B_1^T(t)P_1(t) - C_1(t)] \tag{10.36}$$

where $P_1(t)$ is the solution, assumed to exist on T, of the following Riccati equation:

$$-\dot{P}_1(t) = A^T(t)P_1(t) + P_1(t)A(t) - K^T(t)R_1(t)K(t) \qquad (10.37)$$

and $P_1(t_f)$ *is still to be determined.*

Using this control law, $\Delta\hat{V}$ becomes

$$\begin{aligned}\Delta\hat{V} &= \tfrac{1}{2}v_1^T(t_f)C_0(t_f)z_1(t_f) + \tfrac{1}{2}z_1(t_f)C_0^T(t_f)v_1^T(t_f) \\ &\quad + \tfrac{1}{2}v_1^T(t_f)C_0(t_f)B(t_f)v_1(t_f) - \tfrac{1}{2}z_1(t_f)P_1(t_f)z_1(t_f) \\ &= z_1^T(t_f)[-\tfrac{1}{2}K^T(t_f)C_0(t_f) - \tfrac{1}{2}C_0^T(t_f)K(t_f) \\ &\quad + \tfrac{1}{2}K^T(t_f)C_0(t_f)B(t_f)K(t_f) - \tfrac{1}{2}P_1(t_f)]z_1(t_f) \qquad (10.38)\end{aligned}$$

Therefore, choosing

$$\begin{aligned}P_1(t_f) = -[&K^T(t_f)\bar{C}(t_f) + \bar{C}^T(t_f)K(t_f) - K^T(t_f)\bar{C}(t_f)B(t_f)K(t_f)] \\ &- K^T\Gamma K \qquad (10.39)\end{aligned}$$

if $r \geq 0$, we have from (10.38) that $\Delta\hat{V} \geq 0$.

We now discuss our choice of $\tilde{K}$. From (10.37)

$$\begin{aligned}-\dot{P}_1B &= A^TP_1B + P_1AB - K^TR_1KB \\ &= A^TP_1B + P_1AB - (B^TP_1 - C_1)^TR_1^{-1}(B^TP_1 - C_1)B \\ &= A^TP_1B + P_1AB - (B^TP_1 - C_1)^TR_1^{-1}(B^TP_1B - C_1B) \\ &= A^TP_1B + P_1AB - (B^TP_1 - C_1)^TR_1^{-1}[B^TP_1B - (\dot{\bar{C}} + \bar{C}A)B]\end{aligned}$$

Now

$$R_1 = (-1)(C_1 B + B_1^T \bar{C})$$

$$= (-1)[(\dot{\bar{C}} + \bar{C}A)B + (AB - \dot{B})^T \bar{C}^T]$$

$$= (-1)[\dot{\bar{C}}B + \bar{C}AB + B^T A^T \bar{C}^T - \dot{B}^T \bar{C}^T]$$

so

$$-\bar{C}AB = R_1 + \dot{\bar{C}}B + B^T A^T \bar{C}^T - \dot{B}^T \bar{C}^T$$

Therefore

$$-\dot{P}_1 B = A^T P_1 B + P_1 AB - (B_1^T P_1 - C_1)^T R_1^{-1}[(B_1^T) P_1 B + R_1 + \dot{\bar{C}}B + B^T A^T \bar{C}^T - \dot{B}^T \bar{C}^T]$$

$$= A^T P_1 B + P_1 AB - (B_1^T P_1)^T + C_1^T - K(t) B_1^T P_1 B - K(t) B_1^T \bar{C}$$

$$= A^T P_1 B + P_1 AB - P_1 (AB - \dot{B}) + \dot{\bar{C}}^T + A^T \bar{C}^T - K(t) B_1^T P_1 B - K(t) B_1^T \bar{C}$$

Therefore

$$-\frac{d}{dt}[P_1 B + \bar{C}^T] = [A - B_1 K(t)]^T [P_1 B + \bar{C}^T]$$

So that if

$$P_1(t_f) B(t_f) + \bar{C}^T(t_f) = 0 \qquad (10.40)$$

then

$$P_1(t) B(t) + \bar{C}^T(t) = 0 \quad \text{for all } t \in T$$

satisfying the requirement for $\dot{AV} \geq 0$ (sec 10.14). See also Anderson [23].

Now

$$R_1 = (-1)(C_1 B + B_1^T \bar{C})$$

$$= (-1)[(\dot{\bar{C}} + \bar{C}A)B + (AB - \dot{B})^T \bar{C}^T]$$

$$= (-1)[\dot{\bar{C}}B + \bar{C}AB + B^T A^T \bar{C}^T - \dot{B}^T \bar{C}^T]$$

so

$$-\dot{\bar{C}}AB = R_1 + \dot{\bar{C}}B + B^T A^T \bar{C}^T - \dot{B}^T \bar{C}^T$$

Therefore

$$-\dot{P}_1 B = A^T P_1 B + P_1 AB - (B_1^T P_1 - C_1)^T R_1^{-1}[(B_1^T) P_1 B + R_1 + \dot{\bar{C}}B + B^T A^T \bar{C}^T - \dot{B}^T \bar{C}^T]$$

$$= A^T P_1 B + P_1 AB - (B_1^T P_1)^T + C_1^T - K(t) B_1^T P_1 B - K(t) B_1^T \bar{C}$$

$$= A^T P_1 B + P_1 AB - P_1 (AB - \dot{B}) + \dot{\bar{C}}^T + A^T \bar{C}^T - K(t) B_1^T P_1 B - K(t) B_1^T \bar{C}$$

Therefore

$$-\frac{d}{dt}[P_1 B + \bar{C}^T] = [A - B_1 K(t)]^T [P_1 B + \bar{C}^T]$$

So that if

$$P_1(t_f) B(t_f) + \bar{C}^T(t_f) = 0 \,, \tag{10.40}$$

then

$$P_1(t) B(t) + \bar{C}^T(t) = 0 \quad \text{for all } t \in T$$

satisfying the requirement for $\Delta \hat{V} \geq 0$ (see 10.14). See also Jacobson [22].

Now, if $P_1(t_f)$ satisfies (10.40), we have

$$K(t_f)B(t_f) = R_1^{-1}(t_f)[B_1^T(t_f)P(t_f)B(t_f) - C_1(t_f)B(t_f)]$$

$$= -R_1^{-1}(t_f)[B_1^T(t_f)\bar{C}^T(t_f) + C_1(t_f)B(t_f)]$$

$$= R_1^{-1}(t_f)R_1(t_f) = I.$$

Therefore, let $\tilde{K}$ be given by:

$$\begin{bmatrix} K(t_f) \\ \tilde{K} \end{bmatrix} [B(t_f) \; : \; \tilde{B}] = I_n$$

$\tilde{B}$ is any $(n-m)\times n$ matrix such that $[B(t_f) : \tilde{B}]$ is non singular.

$$K(t_f)B(t_f) = I_m$$

$$\tilde{K}\,B(t_f) = 0$$

$$\tilde{K}\,\tilde{B} = I_{n-m}.$$

Also

$$-\begin{vmatrix} B^T(t_f) \\ \tilde{B}^T \end{vmatrix} [P_1(t_f)][B(t_f) \; : \; \tilde{B}] = \begin{vmatrix} \bar{C}(t_f)B(t_f) & \bar{C}(t_f)\tilde{B} \\ \tilde{B}^T\bar{C}^T(t_f) & \Gamma \end{vmatrix}$$

Now we require $P_1(t_f) \leq 0$. (Compare this and equations (10.37), (10.40) with (10.11) - (10.15).)

Sufficient conditions for the positive definiteness (semi-definiteness) of $-P_1(t_f)$ are

(i) $\bar{C}(t_f)B(t_f) > 0$

(ii) $\Gamma - \tilde{B}^T\bar{C}(t_f)[\bar{C}(t_f)B(t_f)]^{-1}\bar{C}(t_f)\tilde{B} > 0 \qquad (\geq 0)$

We have from Proposition 10.1 and the remarks after Proposition 10.6 that $\tilde{C}B \geq 0$ in T is a necessary condition of optimality. Hence we have (Mayne [36]):

PROPOSITION 10.7: Let the hypothesis of Proposition 4.2 be satisfied. Let f, L be linear in u. Assume the solution $(\tilde{x},\tilde{u})$ is singular with order 1, i.e.

$$S_1(t) = -[C_0(t)B(t) - B^T(t)C_0(t)] = 0$$

$$R_1(t) = -[C_1(t)B(t) + B_1^T(t)C_0^T(t)] > 0 \quad \text{for all } t \in T,$$

C and B are assumed differentiable. Then sufficient conditions for $\Delta\hat{V} \geq 0$ are:

(i) $\tilde{C}(t_f)B(t_f) = B^T(t_f)\tilde{P}(t_f)B(t_f) + C(t_f)B(t_f) > 0$

where $\tilde{P}(t_f)$ is the solution to equations (2.18), (2.19)

(ii) $\Gamma - \tilde{B}^T\tilde{C}(t_f)[\tilde{C}(t_f)B(t_f)]^{-1}\tilde{C}(t_f)\tilde{B} \geq 0$

(iii) The Matrix Riccati differential equation, equ (10.37) with terminal condition satisfying (10.39), (10.40), has no conjugate point in T.

We now obtain the control law for v: From (10.35)

$$v_1(t) = -K(t)z_1(t)$$

so

$$\begin{aligned} v(t) &= \dot{v}_1(t) \\ &= -\dot{K}(t)z_1(t) - K(t)\dot{z}_1(t) \\ &= -[\dot{K}(t) + K(t)A(t)]z_1(t) - K(t)B_1(t)v_1(t) \\ &= -K_1(t)z_1(t) - K(t)B_1(t)v_1(t) \end{aligned}$$

where

$$K_1(t) \triangleq \dot{K}(t) + K(t)A(t)$$

Now, for $P_1(t_f)B(t_f) + \bar{C}^T(t_f) = 0$, then

$$P_1(t)\ B(t) + \bar{C}^T(t) = 0 \text{ for all } t \in T, \text{ and}$$

so

$$K(t)B(t) = I \text{ for all } t \in T$$

i.e., for all $t \in T$

$$K_1(t)B(t) = K(t)B_1(t).$$

Therefore

$$\begin{aligned} v(t) &= -K_1(t)z(t) + K_1(t)B(t)V_1(t) - K(t)B_1(t)v_1(t) \\ &= -K_1(t)z(t). \end{aligned}$$

With this control law

$$\begin{aligned} \frac{d}{dt}[K(t)z(t)] &= [\dot{K}(t)z(t) + K(t)\dot{z}(t)] \\ &= [K_1(t) - K(t)A(t)]z(t) + K(t)[A(t)z(t) + B(t)v(t)] \\ &= K_1(t)z(t) + K(t)B(t)v(t) \\ &= [K_1(t) - K(t)B(t)K_1(t)]z(t) \\ &= [K_1(t) - IK_1(t)]z(t) = 0 \end{aligned}$$

So that

$$K(t)z(t) = 0 \qquad \text{if} \qquad K(t_o)z(t_o) = 0.$$

To obtain necessary and sufficient conditions for the optimality of the original system (i.e. $\Delta V \geq 0$) from the necessary and sufficient conditions for the nonnegativity of $\Delta\dot{V}$, the reader should see, for example [25].

An alternative approach to the above problem is the transformation approach using the transformation of Kelley.

Details of Kelley's transformation are in Kelley [30], Kelley et al [31], Speyer and Jacobson [49] and Mayne [34]. This has the virtue of a lower dimension Riccati equation. However, Speyer and Jacobson [49] point out that if the problem is singular of order higher than one, then a reduction of the state space to achieve a nonsingular problem requires repeated application of the transformation technique; which is cumbersome, especially if there are multiple control variables.

A survey of necessary and sufficient conditions for the optimality for a class of time-varying singular quadratic minimization problems is presented in [23].

Researchers have concerned themselves mainly with necessary and sufficient conditions of optimality, and the computation of singular extremals, with a few exceptions, has largely been ignored. An algorithm, similar to penalty function techniques of solving state constrained problems, with proven convergence, has been presented by Jacobson, Gershwin and Lele [24] for a certain class of problems. They point out that the technique described is equally applicable to the bang-bang control problems of Chapter 8, and provides alternatives to the methods described in [6], [16]. The algorithm makes use of the Second Order D.D.P. Algorithm described in Chapter 5-II-4.

CONCLUSION

The main purpose of this work has been to expand on the paper by Mayne, [36], who exhibited the central role that the exact expressions for the change in cost due to arbitrary controls play in both control theory and numerical optimization.

By applying the differential dynamic programming technique to the partial differential equation

$$-\frac{\partial V(x,t)}{\partial t} = H(x,\omega,V_x(x,t),t)$$

where ω denotes either some arbitrary control $u \in G$, or a control policy $k(x(\cdot),\cdot)$ which generates $v(\cdot)$. We obtained differential equations satisfied by

$$a(t) = V(\bar{x}(t),t) - V^{\bar{u}}(\bar{x}(t),t), \quad t \in T,$$

for some arbitrary nominal control $\bar{u}(\cdot) \in G$ with associated trajectory $\bar{x}(\cdot)$. V denotes V^u or V^k according to the context. Differential equations satisfied by V_x, V_{xx} were also derived. These equations are the same as those derived in [36].

For use in first-order algorithms the error arising in $a(t)$ was shown to be of order ε^2 for $d(u,\bar{u}) \leq \varepsilon$, $d_1(u,\bar{u}) \leq \varepsilon$ when the terms involving V^u_{xx} were omitted from the differential equation satisfied by V^u_x. The error in $a(t)$, when the terms involving V^k_{xxx} were omitted from the differential equations satisfied by V^k_{xx}, was found to be of order ε^3 for

$d(\bar{u},u^*)\leq\varepsilon$, $d_1(\bar{u},u^*)\leq\varepsilon$, where

$$u^* = \arg\min_{u\in G} H(\bar{x},u,V_x^{\bar{k}}(\bar{x},t),t).$$

It was seen in chapter 5 that the neglect of these terms was of no serious consequence in the computational procedures outlined because the size of the δx's had, in any event, to be restricted. This restriction was accomplished via Proposition 1.4, where the size of δx could be limited by the use of ε : $0\leq\varepsilon\leq 1$ for small variations in control, or in the case of global variations in control, by making $d_1(\bar{u},u^*)\leq\varepsilon$.

Next, differential equations were obtained which were satisfied by

$$a(t) = V^u(x(t),t) - V(x(t),t),\ t\in T,$$

where, for arbitrary control $u(\cdot)\in G$ with associated trajectory $x(\cdot)$, V denotes $V^{\bar{u}}$ or $V^{\bar{k}}$ according to the context. $\bar{u}(\cdot)$ is some arbitrary control belonging to G and $\bar{k}(\cdot,\cdot)$ denotes some control policy $k(\bar{x}(\cdot),\cdot)$. This was done to enable us to derive the exact expressions for ΔV obtained in [35], [36]. In this dissertation we have really, if anything, shown the central role the different equations for $a(t)$ have played; in that even the expressions for ΔV were derived from these equations; where $\Delta V \triangleq a(t_o)$.

It was then shown how the expressions for ΔV could be used to obtain certain well known sufficient conditions

of optimality for an optimal control $\bar{u}(\cdot)$ or a control policy $\bar{k}(\bar{x}(\cdot),\cdot)$. The expressions for ΔV given by equ. (2.86) also led directly to a local sufficient condition of optimality. In the following section, approximations $\Delta\hat{V}$ to the expression for ΔV were derived. One of the results was used later to derive a first-order algorithm with proven convergence. The result of Corollary 4.3 was used to demonstrate necessary conditions of optimality for state constrained problems and necessary and sufficient conditions for the nonnegativity of $\Delta\hat{V}$ for singular control problems. These are usually obtained via the second variation formula for the cost function derived in Proposition 4.5. The results obtained involved the comparison of arbitrary controls $u(\cdot)\epsilon G$ with the optimal control $\bar{u}(\cdot)$. The expressions derived later on in the section may prove useful in demonstrating conditions of optimality for the control policy $\bar{k}(\bar{x}(\cdot),\cdot)$. Mayne [36] expressed the hope that the second order estimates for strong variations in control will be useful in deriving further or adapting existing algorithms.

The differential equations which were derived in Chapter 2 - IV are very general. Given some nominal control $\bar{u}(\cdot)\epsilon G$, the equations of Proposition 2.5 were derived for some control policy which generates a control $u(\cdot)\epsilon G$. It was shown that the differential equations used in the algorithms outlined in Chapter 5 could be derived easily from the general differential equations of Propositions 2.4 - 2.7.

Since Jacobson [13] presented his first and second-order algorithms in 1967, and subsequently in [14], [15], [16], [18], which, for the first time enabled global variations in the control to be handled (apart from Polak and Mayne [37], [46]), there has been little further development (eg. Gershwin and Jacobson [9]) of these algorithms. Because of the generality of the new differential equations it is hoped that the results will be useful in deriving further algorithms, especially second-order algorithms. The objective of the new algorithms should be to ensure that $a(t_o) = \Delta V < 0$ at each iteration to determine the new control using the differential equations obtained in 2 - IV. In the literature, this is achieved at present either by minimizing $H(\bar{x},u,\lambda,t)$ w.r.t. u directly, or by approximating the minimizing control $u(\cdot)$. Then, because

$$-\dot{a}(t) = H(\bar{x},u,\lambda,t) - H(\bar{x},\bar{u},\lambda,t) \; ; \; a(t_f) = 0,$$

we have $a(t_o) < 0$.

Also, the existing D.D.P. algorithms may require modifications to ensure convergence in constrained problems. It was shown in [37] how this could be achieved for the first-order algorithm by, firstly, considering an approximation $\hat{\Delta V}$ to ΔV, and, secondly, by modifying the step size choice. It is not unreasonable that using the step size choice of [37], a convergent second-order algorithm may be derived, possibly using a first or second-order approximation to equation (2.86).

Since Jacobson [13] presented his first and second-order algorithms in 1967, and subsequently in [14], [15], [16], [18], which, for the first time enabled global variations in the control to be handled (apart from Polak and Mayne [37], [46]), there has been little further development (eg. Gershwin and Jacobson [9]) of these algorithms. Because of the generality of the new differential equations it is hoped that the results will be useful in deriving further algorithms, especially second-order algorithms. The objective of the new algorithms should be to ensure that $a(t_o) = \Delta V < 0$ at each iteration to determine the new control using the differential equations obtained in 2 - IV. In the literature, this is achieved at present either by minimizing $H(\tilde{x},u,\lambda,t)$ w.r.t. u directly, or by approximating the minimizing control $u(\cdot)$. Then, because

$$-\dot{a}(t) = H(\tilde{x},u,\lambda,t) - H(\tilde{x},\bar{u},\lambda,t) \; ; \; a(t_f) = 0,$$

we have $a(t_o) < 0$.

Also, the existing D.D.P. algorithms may require modifications to ensure convergence in constrained problems. It was shown in [37] how this could be achieved for the first-order algorithm by, firstly, considering an approximation $\hat{\Delta V}$ to ΔV, and, secondly, by modifying the step size choice. It is not unreasonable that using the step size choice of [37], a convergent second-order algorithm may be derived, possibly using a first or second-order approximation to equation (2.86).

In Chapter 6 - I new differential equations were obtained for the problem with terminal equality constraints and/or free terminal time, allowing comparisons of arbitrary controls to be made. These are simply extensions of the general differential equations for the unconstrained case. It was shown how the second-order D.D.P. algorithm of Jacobson could easily be extended to handle problems with terminal equality constraints and/or free terminal time, as well as problems with terminal inequality constraints. It is anticipated that any new second-order algorithm that may be presented that solves the general differential equations presented in Chapter 2 could be extended just as easily to handle terminal constraints and/or free terminal time. Note that Polak and Mayne [46], adapted the first-order algorithm of [37] to account for terminal inequality constraints. Here, the constraints are handled in a manner similar to that in which the Method of Feasible Directions handles control constraints.

Problems with control constraints received attention. For the case where the optimal control is assumed continuous for all $t \in T$, certain equations derived by Jacobson [13], [18] were derived. The second-order algorithm of Jacobson was then adapted to handle control constraints. This appears to be the only available second-order D.D.P. algorithm for handling control constraints directly. A first-order algorithm emerged as a special case of the second-order one.

We were able to compare control $\hat{u}(\cdot)$ with $\bar{u}(\cdot)$, where

$$\hat{u}(t) = \arg \min_{u\in\Omega} H(\bar{x},u,V_x(\bar{x},t),t),$$

and $\bar{x}(\cdot)$ is the trajectory associated with $\bar{u}(\cdot)$, $V_x(\bar{x},\cdot)$ satisfies equ. (7.12). We then defined

$$J(\bar{x},\hat{u},b,V_x(\bar{x},t),t) = H(\bar{x},\hat{u},V_x(\bar{x},t),t) + \langle b(t),\hat{g}(\hat{u},t)\rangle$$

where $\hat{g}(\hat{u},t)$ denotes the $\hat{p}$ active constraints $\hat{g}(\hat{u},t) = 0$, and $b(\cdot)$ is a $\hat{p}$-dimensional Lagrange multiplier; $t\in T$. From [13] or [38], we obtain equations (7.24) and (7.25) which enable us to determine $b(\cdot)$ and $\hat{u}(\cdot)$. These equations do not hold for arbitrary u. As a result general differential equations in which arbitrary controls belonging to Ω are compared, have not yet been derived.

We then considered the case where the optimal control is assumed to be of the bang-bang type. The dynamic programming technique was first applied to the cost function to obtain equations satisfied by the partial derivitives of the cost function at switching points of the bang-bang control $u^*(\cdot)$. It was then shown that, for the optimal control, the switching time was a parameter which could be chosen to minimize the cost function. This led to the jump conditions which were derived independently by Jacobson [16] and Dyer and McReynolds [6]. It was then demonstrated how, if t_1 was a nominal switching point, choice of δt_1, a change in the switching time to ensure a reduction in cost, led to differing jump conditions which were used in the Gradient Method.

We note in passing that Dyer and McReynolds, [6], [7], [8], proposed a successive sweep algorithm which made use of the earlier jump conditions. Jacobson, [13], [16], [18], showed that the second-order algorithm could be adapted to solve the bang-bang control problem. Again, a first-order algorithm emerges as a special case. In a later paper, [24], Jacobson, Gershwin and Lele added an integral quadratic functional of the control multiplied by a parameter ε to the cost function. They then outline a procedure in which the second-order algorithm of [15] is applied successively to the modified problem for a monotonically decreasing sequence of ε_k's. An important feature of this procedure is that it appears to be the first direct application of the powerful D.D.P. algorithms to solve the singular control problem; the addition of the control functional to the cost functional makes the singular control problem nonsingular.

The state constrained problem with only terminal cost ($L \equiv 0$) was considered next. It was shown by Mayne, [36], how one of the expressions for $\Delta\hat{V}$ leads directly to a generalization of the necessary conditions of optimality for state constrained problems derived by Jacobson, Lele and Speyer, [26]. This expression also enabled Mayne to prove a stronger version of the necessary conditions of optimality for singular problems derived by Jacobson, [19]. Sufficient conditions of Jacobson, [20], [22], for the non-negativity of $\Delta\hat{V}$ for singular problems were outlined.

It was then shown how certain well known necessary conditions for the nonnegativity of $\delta\hat{V}$ could be established, more particularly the generalized Legendre-Clebsch conditions. Mayne, [36], demonstrated how satisfaction of these necessary conditions enabled one to derive new sufficient conditions. Clearly, future research could be directed at establishing new conditions of optimality for both state constrained and singular control problems, new conditions for the nonnegativity of $\delta\hat{V}$ for singular problems, and investigation of the relationship between necessary and sufficient conditions (see [25], [49] for example). Also, a close connection between state constrained and singular control problems has been suggested by Jacobson and Lele†, who transformed the state variable inequality constraint by using Valentine's device of introducing a 'slack variable'. They then demonstrate that portions of the optimal trajectory lying along the constraint surface $S(x,t) = 0$ are singular in the transformed problem. A further connection between state constrained and singular control problems was hinted at in [26].

In [13] Jacobson concluded that the notion of D.D.P. could suggest directions of research which would produce efficient computational algorithms for solving state constrained and singular problems. Except for the one algorithm, [24], for the singular problem, no other approaches

† Jacobson, D.H., and M.M. Lele, 'A transformation technique for optimal control problems with state variable inequality constraints', IEEE Trans. Aut. Cont., AC-14, No. 5, 1969.

using D.D.P. appear to have been attempted.

A feature of this dissertation which distinguishes it from that of [36], is that a unified approach is brought about through differential dynamic programming.

REFERENCES.

1. Bellman, R., and S.E. Dreyfus, *Applied Dynamic Programming*, Princeton Univ. Press, 1962.

2. Bryson, A.E., and W. Denham, 'A Steepest Ascent Method for Solving Optimum Programming Problems', J. of Appl. Mech., 29, 247 - 257, 1962.

3. Bryson, A.E., W.F.Denham and S.E. Dreyfus, 'Optimal Programming Problems with Inequality Constraints I: Necessary Conditions for Extremal Solutions', J. AIAA, Vol. 1, 2544 - 2550, 1963.

4. Coddington, E.A., and N. Levinson, Theory of Ordinary Differential Equations, Mc Graw-Hill, 1955.

5. Denham, W.F., and A.E. Bryson, 'Optimal Programming Problems with Inequality Constraints II: Solution by Steepest Descent', J. AIAA, Vol. 2, 25 - 34, 1964.

6. Dyer, P., and S.R. McReynolds, 'On Optimal Control Problems with Discontinuities', J. Math. Anal. Appl., Vol. 23, No. 3, 590 - 603, Sept. 1968.

7. Dyer, P., and S.R. McReynolds, 'Optimization of Control Systems with Discontinuities and Terminal Constraints, I.E.E.E. Trans. Aut. Control, AC-14, No. 3, 223, 1969.

8. Dyer, P., and S.R. McReynolds, The Computation and Theory of Optimal Control, Academic Press, N.Y., 1970.

9. Gershwin, S.B., and D.H. Jacobson, 'A Discrete-Time Differential Dynamic Programming Algorithm with Applications to Optimal Orbit Transfer', J. AIAA, 8, 1616, 1970.

10. Goh, B.S., 'The Second Variation for the Singular Bolza Problem', SIAM J. Control, Vol. 4, 309 - 325, 1966.

11. Goh, B.S., 'Necessary Conditions for Singular Extremals involving Multiple Control Variables', SIAM J. Contol, Vol. 4, 716 - 731, 1966.

12. Halkin, H., 'Mathematical Foundations of System Optimization', in Topics in Optimization (ed. G. Leitman), Academic Press, 1967.

13. Jacobson, D.H., Ph.D. Thesis, Univ. of London, 1967.

14. Jacobson, D.H., 'Second-Order and Second Variation Methods for Determining Optimal Control', Int. J. Cont., 7, 175, 1968.

15. Jacobson, D.H., 'New Second-Order and First-Order Algorithms for Determining Optimal Control: A Differential Dynamic Programming Approach', J. Optim. Theory and Applic., Vol. 2, No. 6, 1968.

16. Jacobson, D.H., 'Differential Dynamic Programming Methods for Solving Bang-Bang Control Problems', I.E.E.E. Trans. Aut. Control, AC-13, 661, 1968.

17. Jacobson, D.H., 'A Note on Error Analysis in Differential Dynamic Programming', I.E.E.E. Trans. Aut. Control, AC-14, 197, 1969.

18. Jacobson, D.H., and D.Q. Mayne, Differential Dynamic Programming, Elsevier Press, New York, 1970.

19. Jacobson, D.H., 'A New Necessary Condition of Optimality for Singular Control Problems', SIAM J. Control, 5, 1969.

20. Jacobson, D.H., 'Sufficient Conditions for the Nonnegativity of the Second Variation in Singular and Non-Singular Control Problems', SIAM J. Control, 8, 1970.

21. Jacobson, D.H., 'On Conditions of Optimality for Singular Control Problems', I.E.E.E. Trans. Aut. Control (Corresp.), AC-15, 109 - 110, 1970.

22. Jacobson, D.H., 'A General Sufficiency Theorem for the Second Variation', J. Math. Anal. Appl., 34, No. 3, 578 - 589, 1971.

23. Jacobson, D.H., 'Totally Singular Quadratic Minimization Problems', I.E.E.E. Trans. Aut. Control, AC-16, No. 6, 651 - 658, 1971.

24. Jacobson, D.H., S.B. Gershwin and M.M. Lele, 'Computation of Optimal Singular Controls', I.E.E.E. Trans. Aut. Control, Vol AC-15, No. 1, 67 - 73, 1970.

25. Jacobson, D.H., and J.L. Speyer, 'Necessary and Sufficient Conditions for Optimality for Singular Control Problems: A Limit Approach', J. Math. Anal. Appl., Vol. 34, 239 - 266, 1971.

18. Jacobson, D.H., and D.Q. Mayne, Differential Dynamic Programming, Elsevier Press, New York, 1970.

19. Jacobson, D.H., 'A New Necessary Condition of Optimality for Singular Control Problems', SIAM J. Control, 5, 1969.

20. Jacobson, D.H., 'Sufficient Conditions for the Nonnegativity of the Second Variation in Singular and Non-Singular Control Problems', SIAM J. Control, 8, 1970.

21. Jacobson, D.H., 'On Conditions of Optimality for Singular Control Problems', I.E.E.E. Trans. Aut. Control (Corresp.), AC-15, 109 - 110, 1970.

22. Jacobson, D.H., 'A General Sufficiency Theorem for the Second Variation', J. Math. Anal. Appl., 34, No. 3, 578 - 589, 1971.

23. Jacobson, D.H., 'Totally Singular Quadratic Minimization Problems', I.E.E.E. Trans. Aut. Control, AC-16, No. 6, 651 - 658, 1971.

24. Jacobson, D.H., S.B. Gershwin and M.M. Lele, 'Computation of Optimal Singular Controls', I.E.E.E. Trans. Aut. Control, Vol AC-15, No. 1, 67 - 73, 1970.

25. Jacobson, D.H., and J.L. Speyer, 'Necessary and Sufficient Conditions for Optimality for Singular Control Problems: A Limit Approach', J. Math. Anal. Appl., Vol. 34, 239 - 266, 1971.

26. Jacobson, D.H., M.M. Lele and J.L. Speyer, 'New Necessary Conditions of Optimality for Control Problems with State-Variable Inequality Constraints', J. Math. Anal. Appl., Vol. 35, No. 2, 255 - 284, 1971.

27. Kalman, R.E., 'Contributions to the Theory of Optimal Control', Boletin de la Sociedad Matematica Mexicana, 102 - 119, 1960.

28. Kelley, H.J., 'Method of Gradients', in Optimization Techniques (ed. G. Leitman), Academic Press, 1962.

29. Kelley, H.J., 'A Second Variation Test for Singular Extremals', J. AIAA, Vol. 2, 1380 - 1382, 1964.

30. Kelley, H.J., 'A Transformation Approach to Singular Subarcs in Optimal Trajectory and Control Problems', SIAM J. Control, Vol. 2, 234 - 240, 1964.

31. Kelley, H.J., R.E. Kopp and H.G. Moyer, 'Singular Extremals', in Topics in Optimization (ed. G. Leitman), Academic Press, 1967.

32. Luenberger, D.G., Optimization by Vector Space Methods, Wiley, New York, 1969.

33. Mayne, D.Q., 'A Second-Order Gradient Method for Determining Optimal Control of Non-linear Discrete-Time Systems', Int. J. Control, 3, 85, 1966.

34. Mayne, D.Q., 'Sufficient Conditions for Optimality for Singular Control Problems', Report 18/70, C.C.D., Imperial College, London, 1970.

35. Mayne, D.Q., 'Properties of a Cost Function Employed in a Second-Order Optimization Algorithm', J. Math. Anal. Appl., Vol. 38, 1972.

36. Mayne, D.Q., 'Differential Dynamic Programming - A Unified Approach to the Optimization of Dynamic Systems', in Advances in Theory and Applications in Control and Dynamic Systems (ed. C.T. Leondes), Vol. 10, 1973.

37. Mayne, D.Q., amd E. Polak, 'First-Order, Strong Variation Algorithms for Optimal Control', Report 72/39, C.C.D., Imperial College, London, 1972, (revised 1973).

38. McCormick, G.P., SIAM J. Appl. Math. 15, 641, 1967.

39. McIntyre, J., and B. Paiewonsky, 'On Optimal Control with Bounded State Variables', in Advances in Control Sysytems (ed. C.T. Leondes), Vol. 5, 1967.

40. McReynolds, S.R., and A.E. Bryson, 'A Successive Sweep Method for Solving Optimal Programming Problems', Paper presanted at the Joint Automatic Control Conf., Troy, New York, 1965.

41. McReynolds, S.R., 'The Successive Sweep Method and Dynamic Programming', J. Math. Anal. Appl., Vol. 19, 3, 565 - 598, 1967.

42. McShane, E.J., Integration, Princeton Univ. Press, 1944.

43. Merriam, C.W., Optimization Theory and the Design of Feedback Controls, McGraw-Hill, 1964.

44. Mitter, S.K., 'Function Space Methods in Optimal Control with Applications to Power Systems', Ph.D. Thesis, Univ. of London, 1965.

45. Noton, A.R.M., P. Dyer and C.A. Markland, 'Numerical Computation of Optimal Control', I.E.E.E. Trans. Aut. Control, AC-12, No. 1, 5th, 1967.

46. Polak, E., and D.Q. Mayne, 'First-Order, Strong Variation Algorithms for Optimal Control Problems with Terminal Inequality Constraints', Report 72/45, C.C.D., Imperial College, London, 1972.

47. Robbins. H.M., 'A Generalized Legendre-Clebsch Condition for the Singular Cases of Optimal Control', IBM J. Res. Develop., Vol. 3, 361 - 372, 1967.

48. Speyer, J.L., 'Optimization and Control of Nonlinear Systems with Inflight Constraints', Ph.D. Thesis, Harvard Univ., 1968; also J.L. Speyer, A.E. Bryson, 'Optimal Programming Problems with a Bounded State Space', J. AIAA, Vol. 6, 1488 - 1491, 1968.

49. Speyer, J.L., D.H. Jacobson, 'Necessary and Sufficient Conditions for Optimality for Singular Control Problems: A Transformation Approach', J. Math. Anal. Appl., 33, 1967.

50. Tait, K.S., 'Singular Problems in Optimal Control', Ph.D. Dissert., Harvard Univ., 1965.

51. Westcott, J.H., 'The Status of Control Theory', Review Paper, Proc. 3rd I.F.A.C. Congr., London, 1966.

Author Jordi Kendal Clive
Name of thesis Differential Dynamic Programming. 1975

PUBLISHER:
University of the Witwatersrand, Johannesburg
©2013

LEGAL NOTICES:

Copyright Notice: All materials on the University of the Witwatersrand, Johannesburg Library website are protected by South African copyright law and may not be distributed, transmitted, displayed, or otherwise published in any format, without the prior written permission of the copyright owner.

Disclaimer and Terms of Use: Provided that you maintain all copyright and other notices contained therein, you may download material (one machine readable copy and one print copy per page) for your personal and/or educational non-commercial use only.

The University of the Witwatersrand, Johannesburg, is not responsible for any errors or omissions and excludes any and all liability for any errors in or omissions from the information on the Library website.