SIAM J. CONTROL
Vol. 8, No. 1, February 1970

DECOUPLING AND POLE ASSIGNMENT IN LINEAR MULTIVARIABLE SYSTEMS: A GEOMETRIC APPROACH*

W. M. WONHAM† AND A. S. MORSE‡

1. Introduction. The current interest in linear multivariable control has led to several algebraic results with important applications to system synthesis. In particular, the problem of decoupling of individual system outputs by means of state variable feedback was studied by Rekasius [1], Falb and Wolovich [2] and Gilbert [3]; the problem of realizing arbitrary pole locations in the closed loop system transfer matrix was investigated by Wonham [4] and Heymann [5]. In the present article, new results are obtained along these lines. In § 3, the problem of neutralizing the effect of disturbances with respect to a specified group of output variables is solved. In § 4, the concept of a controllability subspace is introduced and its relation to pole assignability is investigated. This material is preliminary to the formulation of a general problem of output decoupling in § 5. In § 6 and § 7, necessary and sufficient conditions for decoupling are obtained in two special cases; the results of § 7 complement and extend those obtained previously in [1], [2] and [3]. In each case, the problem of pole assignment is solved completely.

Our viewpoint is that such problems are usefully treated in a geometric framework in which both definitions and results become intuitively transparent. In this way, entanglement at the outset in a thicket of algebraic calculations is avoided. Of course, for applications, it is necessary to translate the geometric criteria into matrix operations suitable for computation. This matter will be considered in a future article.

2. Notation. The control system of interest is specified by the differential equation

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{2.1}$$

with x an n-vector, u an m-vector and A, B constant matrices of dimension, respectively, $n \times n$ and $n \times m$. Here and below, all vectors and matrices have real-valued elements. Script letters denote linear subspaces; $\mathscr{E}^n$ is real n-space; $\mathscr{V}^\perp$ is the orthogonal complement of the subspace $\mathscr{V}$; 0 denotes both the vector zero and the zero subspace.

If K is a matrix, $\{K\}$ or $\mathscr{K}$ is the range of K, and $\mathscr{N}(K)$ is the null space of K. If K is of dimension $\mu \times \nu$ and $\mathscr{V} \subset \mathscr{E}^\mu$, we write $K^{-1}\mathscr{V}$ for the subspace $\{z : z \in \mathscr{E}^\nu, Kz \in \mathscr{V}\} \subset \mathscr{E}^\nu$.

The *controllable subspace* of the pair (A, B), written $\{A|\mathscr{B}\}$, is defined as

$$\{A|\mathscr{B}\} = \mathscr{B} + A\mathscr{B} + \cdots + A^{n-1}\mathscr{B}.$$

* Received by the editors February 3, 1969, and in revised form June 4, 1969.

† Office of Control Theory and Application, NASA Electronics Research Center, Cambridge, Massachusetts 02139. The work of this author was supported by the National Aeronautics and Space Administration while he held an NRC postdoctoral resident research associateship.

‡ Office of Control Theory and Application, NASA Electronics Research Center, Cambridge, Massachusetts 02139.

Thus, $\{A|\mathscr{B}\}$ is the largest subspace of $\mathscr{E}^n$ which the control $u(\cdot)$ in (2.1) can influence. Observe that $\{A|\mathscr{B}\}$ is an A-invariant subspace of $\mathscr{E}^n$.

With (2.1), we consider the auxiliary equation

$$y(t) = Hx(t), \tag{2.2}$$

where H is a constant $q \times n$ matrix. The vector y is the *output*.

Equations (2.1) and (2.2) play no essential role but serve to guide the investigation.

3. Localization of disturbances. In place of (2.1), consider the perturbed system

$$\dot{x}(t) = Ax(t) + Bu(t) + D\xi(t), \tag{3.1}$$

where D is a constant $n \times d$ matrix and $\xi(\cdot)$ is a disturbance input. If $u(t) = Cx(t) + v(t)$ (where $v(\cdot)$ is an external control input), then the output $y(\cdot)$ will be unaffected by all possible $\xi(\cdot)$ if and only if $\{A + BC|\mathscr{D}\} \subset \mathscr{N}(H)$. This suggests the problem: given A, B, $\mathscr{D} \subset \mathscr{E}^n$, $\mathscr{N} \subset \mathscr{E}^n$, under what conditions does there exist an $m \times n$ matrix C such that $\{A + BC|\mathscr{D}\} \subset \mathscr{N}$? If C exists, the effect of disturbances is, in an algebraic sense, localized to $\mathscr{N}$.

THEOREM 3.1. *There exists C such that $\{A + BC|\mathscr{D}\} \subset \mathscr{N}$ if and only if $\mathscr{D} \subset \mathscr{V}$, where $\mathscr{V}$ is the maximal subspace such that*

$$\mathscr{V} \subset \mathscr{N} \cap A^{-1}(\mathscr{B} + \mathscr{V}). \tag{3.2}$$

Furthermore $\mathscr{V}$ is given by $\mathscr{V} = \mathscr{V}^{(\nu)}$, where

$$\mathscr{V}^{(0)} = \mathscr{N}, \qquad \mathscr{V}^{(i)} = \mathscr{V}^{(i-1)} \cap A^{-1}(\mathscr{B} + \mathscr{V}^{(i-1)}), \tag{3.3}$$

$$i = 1, 2, \cdots, \nu,$$

and $\nu = \dim \mathscr{N}$.

Here and below, "maximal" ("minimal") mean l.u.b. (g.l.b.) with respect to the usual partial ordering of subspaces by inclusion.

To prove the theorem we need two auxiliary facts.

LEMMA 3.1. *Let $x_i \in \mathscr{E}^n$, $u_i \in \mathscr{E}^m$, $i = 1, \cdots, N$, and write $X = (x_1, \cdots, x_N)$, $U = (u_1, \cdots, u_N)$. There exists an $m \times n$ matrix C such that $Cx_i = u_i, i = 1, \cdots, N$, if and only if $\mathscr{N}(X) \subset \mathscr{N}(U)$. C always exists if the x_i are linearly independent.*

The simple proof is omitted.

LEMMA 3.2. *Let $\mathscr{V} \subset \mathscr{E}^n$. There exists an $m \times n$ matrix C such that $(A + BC)\mathscr{V} \subset \mathscr{V}$ if and only if $A\mathscr{V} \subset \mathscr{B} + \mathscr{V}$.*

Proof. Necessity is clear. For sufficiency, let $v_1, \cdots, v_\mu$ be a basis of $\mathscr{V}$. Then $Av_i = Bu_i + w_i$ for some $u_i \in \mathscr{E}^m$ and $w_i \in \mathscr{V}$. Choose C, by Lemma 3.1, such that $Cv_i = -u_i$, $i = 1, \cdots, \mu$; then $(A + BC)v_i = w_i$.

Proof of Theorem 3.1. For sufficiency, (3.2) implies $\mathscr{V} \subset \mathscr{N}$ and $A\mathscr{V} \subset \mathscr{B} + \mathscr{V}$. By Lemma 3.2, there exists C such that $(A + BC)\mathscr{V} \subset \mathscr{V}$. Then

$$\{A + BC|\mathscr{D}\} \subset \{A + BC|\mathscr{V}\} = \mathscr{V} \subset \mathscr{N}.$$

The maximal property of $\mathscr{V}$ was not required.

For necessity write $\{A + BC|\mathscr{D}\} = \mathscr{W}$. Then

$$\mathscr{W} \subset \mathscr{N}, \qquad A\mathscr{W} \subset \mathscr{B} + \mathscr{W}. \tag{3.4}$$

If $\mathscr{W}$ is the class of all $\mathscr{W} \subset \mathscr{E}^n$ which satisfy (3.4), then clearly $0 \in \mathscr{W}$ and $\mathscr{W}$ is closed under addition. Hence, $\mathscr{W}$ contains a (unique) maximal member $\mathscr{V}$. Then $\mathscr{D} \subset \mathscr{W} \subset \mathscr{V}$ and $\mathscr{V}$ satisfies (3.2).

To prove the second statement of the theorem, observe that $\mathscr{V}^{(0)} \supset \mathscr{V}$, and if $\mathscr{V}^{(i-1)} \supset \mathscr{V}$, then $\mathscr{V}^{(i)} \supset \mathscr{V} \cap A^{-1}(\mathscr{B} + \mathscr{V}) = \mathscr{V}$. Thus, $\mathscr{V}^{(i)} \supset \mathscr{V}$ for all i; and since $\mathscr{V}^{(i)} \subset \mathscr{V}^{(i-1)}$, there is a least integer j such that $\mathscr{V}^{(i)} = \mathscr{V}^{(j)}$ if $i \geqq j$. Since $\mathscr{V}^{(j)} \supset \mathscr{V}$ and $\mathscr{V}^{(j)}$ satisfies (3.4), $\mathscr{V}^{(j)} = \mathscr{V}$. Clearly, $0 \leqq j \leqq \nu$; and if $\mathscr{D} \subset \mathscr{V}$ we even have $0 \leqq j \leqq \nu - \dim \mathscr{D}$.

Remark 1. Theorem 3.1 depends essentially on the fact that the class $\mathscr{W}$ determined by (3.4), or equivalently

$$\mathscr{W} = \{\mathscr{W} : \mathscr{W} \subset \mathscr{N} \cap A^{-1}(\mathscr{B} + \mathscr{W})\},$$

has a maximal element $\mathscr{V}$. Furthermore, $\mathscr{V}$ is defined constructively by means of (3.3). This fact will be used without special comment in the following sections.

4. Controllability subspaces. In regard to the system (2.1), suppose that a subspace $\mathscr{R} \subset \mathscr{E}^n$ is selected and that it is desired to modify the system in such a way that $\mathscr{R}$, but no larger subspace, is completely controllable. This aim is to be realized by feedback of state variables and by forming suitable linear combinations of control variables: that is, by setting $u = Cx + Kv$, where K is an $m \times m'$ matrix for some $m' \leqq m$. Then (2.1) becomes

$$\dot{x} = (A + BC)x + BKv$$

and we require

$$\{A + BC|\{BK\}\} = \mathscr{R}. \tag{4.1}$$

Condition (4.1) can be expressed more neatly by noting that $\{BK\} \subset \mathscr{B}$ and the following.

LEMMA 4.1. *If $\hat{\mathscr{B}} \subset \mathscr{B}$ and $\{A|\hat{\mathscr{B}}\} = \mathscr{R}$, then $\{A|\mathscr{B} \cap \mathscr{R}\} = \mathscr{R}$. Conversely, if $\{A|\mathscr{B} \cap \mathscr{R}\} = \mathscr{R}$, there exists a matrix K such that $\{A|\{BK\}\} = \mathscr{R}$.*

Proof. $\{A|\hat{\mathscr{B}}\} = \mathscr{R}$ implies $\hat{\mathscr{B}} \subset \mathscr{R}$, so $\hat{\mathscr{B}} \subset \mathscr{B} \cap \mathscr{R}$, and thus $\mathscr{R} = \{A|\hat{\mathscr{B}}\} \subset \{A|\mathscr{B} \cap \mathscr{R}\}$. Also, $A\mathscr{R} \subset \mathscr{R}$ implies $A(\mathscr{B} \cap \mathscr{R}) \subset \mathscr{R}$; by induction $A^j(\mathscr{B} \cap \mathscr{R}) \subset \mathscr{R}$, $j = 1, 2, \cdots$, and so $\{A|\mathscr{B} \cap \mathscr{R}\} \subset \mathscr{R}$.

For the converse, let b_i, $i = 1, \cdots, m$, be the ith column of B and let $\{r_j, j = 1, \cdots, m'\}$ be a basis of $\mathscr{B} \cap \mathscr{R}$. Then

$$r_j = \sum_{i=1}^{m} k_{ij} b_i, \qquad j = 1, \cdots, m',$$

for suitable k_{ij}, and we set $K = [k_{ij}]$. This completes the proof of the lemma.

By Lemma 4.1, we can pose the synthesis problem as follows:

Given A, B and $\mathscr{R}$, find conditions for the existence of C such that

$$\{A + BC|\mathscr{B} \cap \mathscr{R}\} = \mathscr{R}. \tag{4.2}$$

If such a C exists, we call $\mathscr{R}$ a *controllability subspace* of the pair (A, B). Observe that $\mathscr{R} = 0$ and $\mathscr{R} = \{A|\mathscr{B}\}$ are controllability subspaces.

Controllability subspaces can be characterized as follows.

THEOREM 4.1. *Let A, B, $\mathscr{R} \subset \mathscr{E}^n$ be fixed. $\mathscr{R}$ is a controllability subspace of (A, B) if and only if*

$$A\mathscr{R} \subset \mathscr{B} + \mathscr{R} \tag{4.3}$$

and

$$\mathscr{R} = \hat{\mathscr{R}}, \tag{4.4}$$

where $\hat{\mathscr{R}}$ is the minimal subspace such that

$$\hat{\mathscr{R}} = \mathscr{R} \cap (A\hat{\mathscr{R}} + \mathscr{B}). \tag{4.5}$$

Furthermore, $\hat{\mathscr{R}} = \mathscr{R}^{(\rho)}$, where $\rho = \dim \mathscr{R}$ and

$$\begin{aligned} \mathscr{R}^{(0)} &= 0, \\ \mathscr{R}^{(i)} &= \mathscr{R} \cap (A\mathscr{R}^{(i-1)} + \mathscr{B}), \qquad i = 1, 2, \cdots, n. \end{aligned} \tag{4.6}$$

Write $\mathbf{C}$ for the class of matrices C such that $(A + BC)\mathscr{R} \subset \mathscr{R}$. To prove the theorem we need two preliminary results.

LEMMA 4.2. *Let $\tilde{\mathscr{R}} \subset \mathscr{R}$. For all $C \in \mathbf{C}$,*

$$\mathscr{R} \cap \mathscr{B} + (A + BC)\tilde{\mathscr{R}} = \mathscr{R} \cap (A\tilde{\mathscr{R}} + \mathscr{B}).$$

Proof. Let $C \in \mathbf{C}$. Then $(A + BC)\tilde{\mathscr{R}} \subset \mathscr{R}$ and $A\tilde{\mathscr{R}} + \mathscr{B} = (A + BC)\tilde{\mathscr{R}} + \mathscr{B}$. By the modular distributive rule for subspaces,

$$\begin{aligned} \mathscr{R} \cap (A\tilde{\mathscr{R}} + \mathscr{B}) &= \mathscr{R} \cap [(A + BC)\tilde{\mathscr{R}} + \mathscr{B}] \\ &= (A + BC)\tilde{\mathscr{R}} + \mathscr{R} \cap \mathscr{B}. \end{aligned}$$

LEMMA 4.3. *If $C \in \mathbf{C}$ then*

$$\sum_{j=1}^{i} (A + BC)^{j-1}(\mathscr{B} \cap \mathscr{R}) = \mathscr{R}^{(i)}. \qquad i = 1, \cdots, n, \tag{4.7}$$

where the sequence $\mathscr{R}^{(i)}$ is defined by (4.6).

Proof. Equation (4.7) is true for $i = 1$. If it is true for $i = k - 1$, then by Lemma 4.2,

$$\begin{aligned} \sum_{j=1}^{k} (A + BC)^{j-1}(\mathscr{B} \cap \mathscr{R}) &= \mathscr{B} \cap \mathscr{R} + (A + BC)\mathscr{R}^{(k-1)} \\ &= \mathscr{R} \cap (A\mathscr{R}^{(k-1)} + \mathscr{B}) \\ &= \mathscr{R}^{(k)}. \end{aligned}$$

Proof of Theorem 4.1. By Lemma 3.2, $\mathbf{C}$ is nonempty if and only if (4.3) is true. Let

$$\mathscr{R} = \{A + BC | \mathscr{B} \cap \mathscr{R}\}. \tag{4.8}$$

Then $C \in \mathbf{C}$. By Lemma 4.3,

$$\mathscr{R} = \sum_{j=1}^{n} (A + BC)^{j-1}(\mathscr{B} \cap \mathscr{R}) = \mathscr{R}^{(n)} = \mathscr{R}^{(p)}.$$

Conversely, if $\mathscr{R} = \mathscr{R}^{(n)}$, then (4.8) is true for every $C \in \mathbf{C}$. It remains to show that (4.5) has the minimal solution $\mathscr{R}^{(\rho)}$. By induction on i in (4.6), it is seen that $\mathscr{R}^{(i)} \subset \hat{\mathscr{R}}$, $i = 1, 2, \cdots$, for every solution $\hat{\mathscr{R}}$ of (4.5), and that the sequence $\mathscr{R}^{(i)}$ is monotone nondecreasing. Hence, there is $\mu \leqq \rho$ such that $\mathscr{R}^{(i)} = \mathscr{R}^{(\mu)}$ for $i \geqq \mu$; in particular, $\mathscr{R}^{(\rho)} \subset \hat{\mathscr{R}}$ and $\mathscr{R}^{(\rho)}$ satisfies (4.5).

Remark 2. If $\mathscr{R}$ is a controllability subspace, then it was proved incidentally that

$$\mathscr{R} = \{A + BC | \mathscr{B} \cap \mathscr{R}\}$$

for every C such that $(A + BC)\mathscr{R} \subset \mathscr{R}$. This fact will be used later without special mention.

Consider now the problem of assigning the eigenvalues of the restriction of $A + BC$ to $\mathscr{R}$. It will be shown that there is complete freedom of assignment and that simultaneously the control v introduced earlier can be made a scalar; i.e., in (4.1) K can be made an m-vector ($m' = 1$). For this, recall [4] that a subspace $\mathscr{X}$ is *A-cyclic* if there exists $x \in \mathscr{X}$ such that $\{A | \{x\}\} = \mathscr{X}$; that is, if $\mathscr{X}$ contains a *generator* x. Thus we can take $m' = 1$ if and only if $\mathscr{R}$ can be made $(A + BC)$-cyclic and $\mathscr{R} \cap \mathscr{B}$ contains a generator.

THEOREM 4.2. *Let* (4.3) *and* (4.4) *hold, and let* $\alpha_1, \cdots, \alpha_\rho$ *be arbitrary real numbers* ($\rho = \dim \mathscr{R}$). *Then C can be chosen such that* (4.2) *is true and* $\mathscr{R}$ *is* $(A + BC)$*-cyclic with characteristic polynomial*

$$\lambda^\rho - \sum_{i=1}^{\rho} \alpha_i \lambda^{i-1}. \tag{4.9}$$

If $0 \neq b \in \mathscr{B} \cap \mathscr{R}$ *is arbitrary, C can be chosen so that, in addition, b generates* $\mathscr{R}$.

Proof. By Lemma 4.3 and Theorem 4.1, $\mathbf{C}$ is nonempty and

$$\{A + BC | \mathscr{B} \cap \mathscr{R}\} = \mathscr{R} \tag{4.10}$$

for every $C \in \mathbf{C}$. Choose $C_1 \in \mathbf{C}$ arbitrarily and write $A + BC_1 = A_1$. Let $b_1 = b \in \mathscr{R} \cap \mathscr{B}$ and let ρ_1 be the largest integer such that the vectors

$$b_1, A_1 b_1, \cdots, A_1^{\rho_1 - 1} b_1$$

are independent. Put $r_1 = b_1$ and $r_j = A_1 r_{j-1} + b_1$, $j = 2, \cdots, \rho_1$. Then $r_i \in \mathscr{R}$ and the r_i are independent. If $\rho_1 < \rho$, choose $b_2 \in \mathscr{R} \cap \mathscr{B}$ such that $r_1, \cdots, r_{\rho_1}$, b_2 are independent; such a b_2 exists by (4.7). Let ρ_2 be the greatest integer such that

$$b_1, \cdots, A_1^{\rho_1 - 1} b_1, b_2, \cdots, A_1^{\rho_2 - 1} b_2$$

are independent, and define

$$r_{\rho_1 + i} = A_1 r_{\rho_1 + i - 1} + b_2, \qquad i = 1, \cdots, \rho_2.$$

Then $r_1, \cdots, r_{\rho_2}$ are independent and in $\mathscr{R}$. Continuing thus, we obtain eventually $r_1, \cdots, r_\rho$ independent and in $\mathscr{R}$, with the property

$$r_{i+1} = A_1 r_i + \tilde{b}_i, \qquad i = 1, \cdots, \rho - 1,$$

where $\tilde{b}_i \in \mathscr{R} \cap \mathscr{B}$. Now let C_2 be chosen such that

$$BC_2 r_i = \tilde{b}_i, \qquad i = 1, \cdots, \rho,$$

where $\tilde{b}_\rho \in \mathscr{R} \cap \mathscr{B}$ is arbitrary. Since $\tilde{b}_i = Bu_i$ for suitable u_i, and the r_i are independent, Lemma 3.1 guarantees that C_2 exists. The situation now is that

$$r_{i+1} = (A_1 + BC_2)r_i, \qquad i = 1, \cdots, \rho - 1,$$

and

$$(A_1 + BC_2)r_\rho \in \mathscr{R}.$$

By independence of the r_i,

$$\{A_1 + BC_2|\{r_1\}\} = \mathscr{R};$$

that is, $\mathscr{R}$ is cyclic relative to $A + B(C_1 + C_2)$ with generator $r_1 = b_1 \in \mathscr{R} \cap \mathscr{B}$. It is well known [4] that now an n-vector c can be found such that $A + B(C_1 + C_2) + b_1c'$ (restricted to $\mathscr{R}$) has the characteristic polynomial (4.9). Setting $b_1 = Bg$ for suitable $g \in \mathscr{E}^m$, it follows that the matrix

$$C = C_1 + C_2 + gc'$$

has all the required properties.

Remark 3. The result that any nonzero vector in $\mathscr{B} \cap \mathscr{R}$ can serve as generator is an extension of the useful lemma in [5].

Remark 4. If $\mathscr{R} = \mathscr{E}^n$, (4.3) holds automatically and (4.4) amounts to $\{A|\mathscr{B}\} = \mathscr{E}^n$, i.e., complete controllability of (A, B). Then Theorem 4.2 yields the known result [4] that controllability implies pole assignability. The construction just used furnishes a simpler proof of this fact than that in [4].

It will be necessary later to compute the maximal controllability subspace contained in a given subspace $\mathscr{S}$. For this, let $\bar{\mathscr{V}}$ be the maximal subspace of $\mathscr{S}$ which is $(A + BC)$-invariant for some C (recall Remark 1 following Theorem 3.1); and let $\mathbf{C}(\bar{\mathscr{V}})$ be the class of C for which $(A + BC)\bar{\mathscr{V}} \subset \bar{\mathscr{V}}$.

THEOREM 4.3. *If* $C \in \mathbf{C}(\bar{\mathscr{V}})$, *the subspace*

$$\bar{\mathscr{R}} = \{A + BC|\mathscr{B} \cap \bar{\mathscr{V}}\} \tag{4.11}$$

is the maximal controllability subspace in $\mathscr{S}$.

Proof. By (4.2) and Lemma 4.1, $\bar{\mathscr{R}}$ is a controllability subspace. Furthermore, by Lemma 4.3 with $\mathbf{C}(\bar{\mathscr{V}})$ in place of $\mathbf{C}$, $\bar{\mathscr{R}}$ is independent of $C \in \mathbf{C}(\bar{\mathscr{V}})$ and so is uniquely defined. Now suppose

$$\hat{\mathscr{R}} = \{A + B\hat{C}|\mathscr{B} \cap \hat{\mathscr{R}}\}, \qquad \hat{\mathscr{R}} \subset \mathscr{S}.$$

Since $\hat{\mathscr{R}}$ is $(A + B\hat{C})$-invariant and $\bar{\mathscr{V}}$ is maximal, there follows $\hat{\mathscr{R}} \subset \bar{\mathscr{V}}$. Let $\bar{\mathscr{V}} = \hat{\mathscr{R}} \oplus \hat{\mathscr{V}}$. By the construction used in proving Lemma 3.2, a matrix C exists such that

$$Cx = \hat{C}x, \quad x \in \hat{\mathscr{R}}; \qquad (A + BC)\hat{\mathscr{V}} \subset \bar{\mathscr{V}}.$$

Then $C \in \mathbf{C}(\bar{\mathscr{V}})$, and

$$\begin{aligned} \hat{\mathscr{R}} &= \{A + BC|\mathscr{B} \cap \hat{\mathscr{R}}\} \\ &\subset \{A + BC|\mathscr{B} \cap \bar{\mathscr{V}}\} \\ &= \bar{\mathscr{R}}; \end{aligned}$$

that is, $\bar{\mathscr{R}}$ is maximal.

5. Decoupling of output variables: Problem statement. Consider the output equation (2.2), with

$$H = \begin{bmatrix} H_1 \\ \vdots \\ H_k \end{bmatrix}, \tag{5.1}$$

where H_i is of dimension $q_i \times n$, $i = 1, \cdots, k$, $k \geqq 2$, $q_1 + \cdots + q_k = q$. Then (2.2) can be written

$$y_i = H_i x, \qquad i = 1, \cdots, k, \tag{5.2}$$

where y_i is a q_i-vector. The vectors y_i may be regarded as physically significant groups of scalar output variables. It may therefore be desirable to control completely each of the output vectors y_i individually, without affecting the behavior of the remaining y_j, $j \neq i$. This end is to be achieved by linear state-variable feedback together with the assignment of a suitable group of control inputs to each y_i. That is, in (2.1) we set

$$u = Cx + \sum_{i=1}^{k} K_i v_i. \tag{5.3}$$

For v_i to control y_i completely, we must have

$$H_i\{A + BC|\{BK_i\}\} = \mathscr{H}_i, \tag{5.4}$$

where $\mathscr{H}_i$ is the range of H_i. Since the ith control v_i is to leave the outputs y_j, $j \neq i$, unaffected, we require also

$$H_j\{A + BC|\{BK_i\}\} = 0, \qquad j \neq i. \tag{5.5}$$

Recalling the equivalence of (4.1) and (4.2), we can express conditions (5.4) and (5.5) more neatly as follows. Write $\mathscr{E}^n = \mathscr{E}$ and

$$\mathscr{N}(H_i) = \mathscr{N}_i, \qquad i = 1, \cdots, k. \tag{5.6}$$

Then our problem is: *Given A, B and $\mathscr{N}_1, \cdots, \mathscr{N}_k$, find a matrix C and controllability subspaces $\mathscr{R}_1, \cdots, \mathscr{R}_k$, with the properties*:

$$\mathscr{R}_i = \{A + BC|\mathscr{B} \cap \mathscr{R}_i\}, \qquad i = 1, \cdots, k, \tag{5.7}$$

$$\mathscr{R}_i + \mathscr{N}_i = \mathscr{E}, \qquad i = 1, \cdots, k, \tag{5.8}$$

$$\mathscr{R}_i \subset \bigcap_{j \neq i} \mathscr{N}_j, \qquad i = 1, \cdots, k. \tag{5.9}$$

Here (5.8) and (5.9) are equivalent, respectively, to (5.4) and (5.5).

The relations (5.7)–(5.9) provide a geometric formulation of the problem of simultaneous decoupling and complete control of the output vectors $y_i, \cdots, y_k$. Thus stated, the problem definition is both natural and intuitively transparent.

We observe that the output matrices H_i play no role beyond specification of the subspaces $\mathscr{N}_i$. Since the H_i need have no special structure, the $\mathscr{N}_i$ are similarly

unrestricted. Nevertheless, we shall rule out trivialities by tacitly assuming:

(i) $\mathcal{N}_i \neq \mathcal{E}$, $i = 1, \cdots, k$.

(ii) The subspaces $\mathcal{N}_i^{\perp}$ are mutually independent.[1] In particular, the $\mathcal{N}_i$ are distinct and

$$\mathcal{N}_i \neq 0, \qquad i = 1, \cdots, k. \tag{5.10}$$

(iii) The pair (A, B) is completely controllable, i.e., $\{A|\mathcal{B}\} = \mathcal{E}$.

For if (i) fails, then for some i, $\mathcal{N}_i = \mathcal{E}$; that is, $H_i = 0$ and $y_i \equiv 0$. If (ii) fails, then for some i,

$$\mathcal{N}_i^{\perp} \cap \sum_{j \neq i} \mathcal{N}_j^{\perp} \neq 0$$

or, by taking orthogonal complements,

$$\mathcal{N}_i + \bigcap_{j \neq i} \mathcal{N}_j \neq \mathcal{E}$$

and (5.8) must fail. For (iii), if $\{A|\mathcal{B}\} = \mathcal{E}_1 \neq \mathcal{E}$ we can write $\mathcal{E} = \mathcal{E}_1 \oplus \mathcal{E}_2$ and (2.1) as

$$\dot{x}_1 = A_1 x_1 + A_3 x_2 + B_1 u,$$

$$\dot{x}_2 = A_2 x_2,$$

where $x_i \in \mathcal{E}_i$, $i = 1, 2$, and $\{A_1|\mathcal{B}_1\} = \mathcal{E}_1$. The problem is unrealistic unless A_2 is stable (i.e., the pair (A, B) is stabilizable [4]). Hence, we may assume $x_2(t) \equiv 0$ and take as starting point

$$\dot{x}_1 = A_1 x_1 + B_1 u.$$

The problem can then be reformulated with $\mathcal{E}_1$ in place of $\mathcal{E}$.

We turn now to the determination of necessary and sufficient conditions for the existence of a solution to (5.7)–(5.9) in two special, but interesting, cases.

In the following sections, $\bar{\mathcal{R}}_i$ denotes the maximal controllability subspace such that

$$\bar{\mathcal{R}}_i \subset \bigcap_{j \neq i} \mathcal{N}_j, \qquad i = 1, \cdots, k. \tag{5.11}$$

The $\bar{\mathcal{R}}_i$ are constructed according to Theorem 4.3.

6. Decoupling when rank $(H) = n$. Our assumption is equivalent to

$$\bigcap_{i=1}^{k} \mathcal{N}_i = 0. \tag{6.1}$$

That is, there is a one-to-one mapping of state variables into output variables.

THEOREM 6.1. *If* (6.1) *holds, then the problem* (5.7)–(5.9) *has a solution if and only if*

$$\bar{\mathcal{R}}_i + \mathcal{N}_i = \mathcal{E}, \qquad i = 1, \cdots, k. \tag{6.2}$$

[1] Equivalently, the row spaces of the H_i are mutually independent.

Proof. If the problem has a solution $\mathscr{R}_i$, $i = 1, \cdots, k$, then by maximality of the $\bar{\mathscr{R}}_i$, $i = 1, \cdots, k$, there follows $\mathscr{R}_i \subset \bar{\mathscr{R}}_i$, and (6.2) follows from (5.8).

Conversely, suppose (6.2) holds. The $\bar{\mathscr{R}}_i$ are mutually independent; for, by (5.11) and (6.1),

$$\bar{\mathscr{R}}_i \cap \sum_{\mu \neq i} \bar{\mathscr{R}}_\mu \subset \left[\bigcap_{j \neq i} \mathscr{N}_j\right] \cap \left[\sum_{\mu \neq i} \bigcap_{\nu \neq \mu} \mathscr{N}_\nu\right] \subset \bigcap_{j \neq i} \mathscr{N}_j \cap \mathscr{N}_i = 0.$$

Let C_i be chosen such that

$$\bar{\mathscr{R}}_i = \{A + BC_i | \mathscr{B} \cap \bar{\mathscr{R}}_i\}, \qquad i = 1, \cdots, k.$$

Since the $\bar{\mathscr{R}}_i$ are independent there exists, by Lemma 3.1, a matrix C such that $Cr = C_i r$ ($r \in \bar{\mathscr{R}}_i$, $i = 1, \cdots, k$), i.e.,

$$(A + BC)r = (A + BC_i)r, \qquad r \in \bar{\mathscr{R}}_i, \quad i = 1, \cdots, k.$$

Then

$$\bar{\mathscr{R}}_i = \{A + BC | \mathscr{B} \cap \bar{\mathscr{R}}_i\}, \qquad i = 1, \cdots, k;$$

and C, together with the $\bar{\mathscr{R}}_i$, satisfy (5.7)–(5.9).

Remark 5. By Theorem 4.2, the C_i can be chosen so that $A + BC_i$, restricted to $\bar{\mathscr{R}}_i$, has any desired spectrum. Hence, the same is true for $A + BC$. Furthermore, there exists $b_i \in \mathscr{B} \cap \bar{\mathscr{R}}_i$ such that

$$\bar{\mathscr{R}}_i = \{A + BC | \{b_i\}\}, \qquad i = 1, \cdots, k.$$

7. Decoupling when rank $(B) = k$. Our assumption is equivalent to

$$\dim \mathscr{B} = k. \tag{7.1}$$

Here the situation has been simplified by narrowing the choice of generating subspaces $\mathscr{B} \cap \mathscr{R}_i$. The same assumption was made in [1], [2] and [3], with the additional restriction that the outputs y_i be scalars.

THEOREM 7.1. *If* (7.1) *holds, then the problem* (5.7)–(5.9) *has a solution if and only if*

$$\bar{\mathscr{R}}_i + \mathscr{N}_i = \mathscr{E}, \qquad i = 1, \cdots, k, \tag{7.2}$$

and

$$\mathscr{B} = \sum_{i=1}^{k} \mathscr{B} \cap \bar{\mathscr{R}}_i. \tag{7.3}$$

Furthermore, if $C, \mathscr{R}_1, \cdots, \mathscr{R}_k$ *is any solution, then*

$$\mathscr{R}_i = \bar{\mathscr{R}}_i, \qquad i = 1, \cdots, k. \tag{7.4}$$

Proof. Part 1. Suppose $C, \mathscr{R}_1, \cdots, \mathscr{R}_k$ is a solution. The necessity of (7.2) follows, as in the proof of Theorem 6.1. To verify (7.3), write

$$\mathscr{B} \cap \mathscr{R}_i = \mathscr{B}_i \oplus \left[\mathscr{B} \cap \mathscr{R}_i \cap \sum_{j \neq i} \mathscr{R}_j\right], \qquad i = 1, \cdots, k.$$

The $\mathscr{B}_i$ are mutually independent; in fact,

$$\mathscr{B}_i \cap \sum_{j \neq i} \mathscr{B}_j \subset \mathscr{B}_i \cap \sum_{j \neq i} \mathscr{B} \cap \mathscr{R}_j \subset \mathscr{B}_i \cap (\mathscr{B} \cap \mathscr{R}_i) \cap \sum_{j \neq i} \mathscr{R}_j = 0.$$

Recall that the $\mathscr{R}_i$ are $(A + BC)$-invariant. Then

$$\mathscr{R}_i = \{A + BC|\mathscr{B}_i\} + \tilde{\mathscr{R}}_i,$$

where

$$\tilde{\mathscr{R}}_i \subset \left\{ A + BC \middle| \sum_{j \neq i} \mathscr{R}_j \right\} \subset \sum_{j \neq i} \mathscr{R}_j \subset \sum_{j \neq i} \bigcap_{\mu \neq j} \mathscr{N}_\mu \subset \mathscr{N}_i.$$

Therefore, by (5.8),

$$\{A + BC|\mathscr{B}_i\} + \mathscr{N}_i = \mathscr{E},$$

and since $\mathscr{N}_i \neq \mathscr{E}$ there follows $\mathscr{B}_i \neq 0$, $i = 1, \cdots, k$. Therefore

$$\dim \sum_{i=1}^{k} \mathscr{B}_i = \sum_{i=1}^{k} \dim \mathscr{B}_i = k;$$

so

$$\mathscr{B} = \mathscr{B}_1 \oplus \cdots \oplus \mathscr{B}_k \tag{7.5}$$

and

$$\dim \mathscr{B}_i = 1, \qquad i = 1, \cdots, k.$$

Since $\mathscr{B}_i \subset \mathscr{B} \cap \mathscr{R}_i \subset \mathscr{B} \cap \bar{\mathscr{R}}_i$, it follows that (7.3) is true.

Proof. Part 2. To verify (7.4), it is enough to show that the subspaces $\mathscr{B} \cap \bar{\mathscr{R}}_i$ are independent. For then,

$$\dim (\mathscr{B} \cap \bar{\mathscr{R}}_i) = 1, \qquad i = 1, \cdots, k,$$

and so

$$\mathscr{B} \cap \mathscr{R}_i = \mathscr{B} \cap \bar{\mathscr{R}}_i, \qquad i = 1, \cdots, k. \tag{7.6}$$

Assuming (7.6) is true, let $\bar{\mathscr{R}}_i = \hat{\mathscr{R}}_i \oplus \mathscr{R}_i$ and choose C_i, by Lemma 3.1, such that

$$(A + BC_i)\hat{\mathscr{R}}_i \subset \mathscr{R}_i, \qquad C_i r = Cr, \quad r \in \mathscr{R}_i, \quad i = 1, \cdots, k.$$

Then $C_i \in \mathbf{C}(\mathscr{R}_i) \cap \mathbf{C}(\bar{\mathscr{R}}_i)$, so that

$$\begin{aligned} \mathscr{R}_i &= \{A + BC_i|\mathscr{B} \cap \mathscr{R}_i\} \\ &= \{A + BC_i|\mathscr{B} \cap \bar{\mathscr{R}}_i\} \\ &= \bar{\mathscr{R}}_i \end{aligned}$$

which proves (7.4).

We proceed to show that the $\mathscr{B} \cap \bar{\mathscr{R}}_i$ are independent. Write

$$\bar{\mathscr{R}}_i^* = \sum_{j \neq i} \bar{\mathscr{R}}_j.$$

It is even true that

$$\mathscr{B} \cap \bar{\mathscr{R}}_i \cap \bar{\mathscr{R}}_i^* = 0, \qquad i = 1, \cdots, k. \tag{7.7}$$

On the contrary, suppose (7.7) fails for, say, $i = 1$. If $\dim (\mathscr{B} \cap \bar{\mathscr{R}}_1) = 1$, then

$$\mathscr{B} \cap \bar{\mathscr{R}}_1 \subset \bar{\mathscr{R}}_1^*. \tag{7.8}$$

If $\dim (\mathscr{B} \cap \bar{\mathscr{R}}_1) \geqq 2$, and

$$\mathscr{B} \cap \bar{\mathscr{R}}_i \not\subset \bar{\mathscr{R}}_i^*, \qquad i = 1, \cdots, k, \tag{7.9}$$

then

$$\dim \left[\sum_{i=1}^{\mu+1} \mathscr{B} \cap \bar{\mathscr{R}}_i \right] \geqq \dim \left[\sum_{i=1}^{\mu} \mathscr{B} \cap \bar{\mathscr{R}}_i \right] + 1,$$

$\mu = 1, \cdots, k-1$, that is,

$$\dim \left[\sum_{i=1}^{2} \mathscr{B} \cap \bar{\mathscr{R}}_i \right] \geqq 3;$$

and by induction

$$\dim \left[\sum_{i=1}^{k} \mathscr{B} \cap \bar{\mathscr{R}}_i \right] \geqq k + 1,$$

a contradiction. Thus (7.9) is false; combining this result with (7.8) there follows

$$\mathscr{B} \cap \bar{\mathscr{R}}_\alpha \subset \bar{\mathscr{R}}_\alpha^* \tag{7.10}$$

for some $\alpha \in (1, \cdots, k)$. It will be shown below that there exists $\tilde{C}_\alpha$ such that

$$(A + B\tilde{C}_\alpha)\bar{\mathscr{R}}_\alpha \subset \bar{\mathscr{R}}_\alpha; \qquad (A + B\tilde{C}_\alpha)\bar{\mathscr{R}}_\alpha^* \subset \bar{\mathscr{R}}_\alpha^*. \tag{7.11}$$

Assuming (7.11) is true, we have

$$\bar{\mathscr{R}}_\alpha = \{A + B\tilde{C}_\alpha | \mathscr{B} \cap \bar{\mathscr{R}}_\alpha\} \subset \{A + B\tilde{C}_\alpha | \bar{\mathscr{R}}_\alpha^*\} \subset \bar{\mathscr{R}}_\alpha^* \subset \mathscr{N}_\alpha,$$

and therefore (7.2) fails for $i = \alpha$. With this contradiction, (7.7) is established. It remains to verify the existence of $\tilde{C}_\alpha$. For this we need the following result.

LEMMA 7.1. *Let $\mathscr{V}$, $\mathscr{W}$ be arbitrary. There exists C such that*

$$(A + BC)\mathscr{V} \subset \mathscr{V}, \qquad (A + BC)\mathscr{W} \subset \mathscr{W}$$

if and only if

$$A\mathscr{V} \subset \mathscr{B} + \mathscr{V},$$

$$A\mathscr{W} \subset \mathscr{B} + \mathscr{W},$$

$$A(\mathscr{V} \cap \mathscr{W}) \subset \mathscr{B} + \mathscr{V} \cap \mathscr{W}.$$

Proof. Necessity is obvious. For sufficiency, write

$$\mathscr{V} + \mathscr{W} = \hat{\mathscr{V}} \oplus (\mathscr{V} \cap \mathscr{W}) \oplus \hat{\mathscr{W}},$$

where $\hat{\mathscr{V}} \subset \mathscr{V}$, $\hat{\mathscr{W}} \subset \mathscr{W}$. By the construction of Lemma 3.2, C can be chosen such that

$$(A + BC)(\mathscr{V} \cap \mathscr{W}) \subset \mathscr{V} \cap \mathscr{W},$$
$$(A + BC)\hat{\mathscr{V}} \subset \mathscr{V},$$
$$(A + BC)\hat{\mathscr{W}} \subset \mathscr{W}.$$

This completes the proof of the lemma.

Consider now $\bar{\mathscr{R}}_\alpha$, $\bar{\mathscr{R}}_\alpha^*$. Clearly $A\bar{\mathscr{R}}_\alpha \subset \mathscr{B} + \bar{\mathscr{R}}_\alpha$; $A\bar{\mathscr{R}}_\alpha^* \subset \mathscr{B} + \bar{\mathscr{R}}_\alpha^*$. By (7.3) $\mathscr{B} = \mathscr{B} \cap \bar{\mathscr{R}}_\alpha + \mathscr{B} \cap \bar{\mathscr{R}}_\alpha^*$, and so

$$\begin{aligned} A(\bar{\mathscr{R}}_\alpha \cap \bar{\mathscr{R}}_\alpha^*) &\subset (\mathscr{B} + \bar{\mathscr{R}}_\alpha) \cap (\mathscr{B} + \bar{\mathscr{R}}_\alpha^*) \\ &= \mathscr{B} + (\mathscr{B} + \bar{\mathscr{R}}_\alpha) \cap \bar{\mathscr{R}}_\alpha^* \\ &= \mathscr{B} + (\mathscr{B} \cap \bar{\mathscr{R}}_\alpha^* + \bar{\mathscr{R}}_\alpha) \cap \bar{\mathscr{R}}_\alpha^* \\ &= \mathscr{B} + \bar{\mathscr{R}}_\alpha \cap \bar{\mathscr{R}}_\alpha^*. \end{aligned}$$

By applying Lemma 7.1, the existence of $\tilde{C}_\alpha$ is finally established.

Proof. Part 3. We now prove that (7.2) and (7.3) are sufficient conditions for existence of a solution. Let $\bar{\mathscr{V}}_i$ be the maximal subspace such that

$$A\bar{\mathscr{V}}_i \subset \mathscr{B} + \bar{\mathscr{V}}_i, \qquad \bar{\mathscr{V}}_i \subset \bigcap_{j \neq i} \mathscr{N}_j, \qquad i = 1, \cdots, k. \tag{7.12}$$

It is enough to check that the $\bar{\mathscr{V}}_i$ are *compatible*, in the sense that there exists C such that

$$(A + BC)\bar{\mathscr{V}}_i \subset \bar{\mathscr{V}}_i, \qquad i = 1, \cdots, k.$$

We show first that the subspaces

$$\bar{\mathscr{V}}_i^* = \sum_{j \neq i} \bar{\mathscr{V}}_i$$

are compatible. From (7.12) there follows

$$\begin{aligned} A\bar{\mathscr{V}}_i^* &\subset \mathscr{B} + \bar{\mathscr{V}}_i^* \\ &= \mathscr{B} \cap \bar{\mathscr{V}}_i + \bar{\mathscr{V}}_i^* \qquad \text{(by (7.3))} \\ &= \mathscr{B}_i + \bar{\mathscr{V}}_i^*, \qquad i = 1, \cdots, k, \end{aligned}$$

where $\mathscr{B}_i = \mathscr{B} \cap \bar{\mathscr{V}}_i$. By Lemma 3.2, there exist B_i with $\{B_i\} = \mathscr{B}_i$, and C_i, such that

$$(A + B_iC_i)\bar{\mathscr{V}}_i^* \subset \bar{\mathscr{V}}_i^*, \qquad i = 1, \cdots, k.$$

Choosing a basis $\{v_1, \cdots, v_\mu\}$ for $\bar{\mathscr{V}}_1 + \cdots + \bar{\mathscr{V}}_k$, we define C such that

$$BCv_\nu = \sum_{i=1}^{k} B_iC_iv_\nu, \qquad \nu = 1, \cdots, \mu.$$

Then

$$
\begin{aligned}
(A + BC)\bar{\mathscr{V}}_i^* &= \left(A + B_iC_i + \sum_{j\neq i} B_jC_j\right)\bar{\mathscr{V}}_i^* \\
&\subset (A + B_iC_i)\bar{\mathscr{V}}_i^* + \sum_{j\neq i} \mathscr{B}_j \\
&\subset \bar{\mathscr{V}}_i^* + \sum_{j\neq i} \bar{\mathscr{V}}_j \\
&= \bar{\mathscr{V}}_i^*, \qquad i = 1, \cdots, k.
\end{aligned} \tag{7.13}
$$

This proves compatibility of the $\bar{\mathscr{V}}_i^*$. Now define

$$
\mathscr{V}_i = \bigcap \bar{\mathscr{V}}_j^*, \qquad i = 1, \cdots, k.
$$

Clearly, $\mathscr{V}_i \supset \bar{\mathscr{V}}_i$, $i = 1, \cdots, k$. By (7.13),

$$
(A + BC)\mathscr{V}_i \subset \mathscr{V}_i, \qquad i = 1, \cdots, k, \tag{7.14}
$$

and, furthermore by the second condition of (7.12),

$$
\mathscr{V}_i \subset \bigcap_{j\neq i} \sum_{\alpha\neq j} \bigcap_{m\neq\alpha} \mathscr{N}_m = \bigcap_{j\neq i} \mathscr{N}_j.^2 \tag{7.15}
$$

By (7.14) and (7.15), the $\mathscr{V}_i$ satisfy the conditions imposed on the $\bar{\mathscr{V}}_i$ in (7.12). Since the $\bar{\mathscr{V}}_i$ are maximal, there results $\mathscr{V}_i \subset \bar{\mathscr{V}}_i$, and, therefore, $\mathscr{V}_i = \bar{\mathscr{V}}_i$, $i = 1, \cdots, k$.

Remark 6. If the conditions of Theorem 7.1 are satisfied, then

$$
\begin{aligned}
\sum_{i=1}^{k} \bar{\mathscr{R}}_i &= \sum_{i=1}^{k} \{A + BC|\mathscr{B} \cap \bar{\mathscr{R}}_i\} = \left\{A + BC \middle| \sum_{i=1}^{k} \mathscr{B} \cap \bar{\mathscr{R}}_i\right\} \\
&= \{A + BC|\mathscr{B}\} = \{A|\mathscr{B}\} = \mathscr{E}.
\end{aligned} \tag{7.16}
$$

We turn now to the problem of pole assignment. In contrast to the situation of § 6, it is no longer possible, in general, to vary the spectrum of $A + BC$ on each $\bar{\mathscr{R}}_i$ independently. The following example shows that certain eigenvalues of $A + BC$ may even be fixed for all admissible C.

Let $\mathscr{E} = \mathscr{E}^3$, $k = 2$ and

$$
A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix},
$$

$$
\mathscr{N}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}^{\perp}, \qquad \mathscr{N}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}^{\perp}.
$$

[2] This identity and its dual, $\sum \cap \sum = \sum$, are readily established by using the (modular) distributive rule for subspaces.

It is easily checked that (5.7)–(5.9) have the (unique) solution

$$\mathscr{R}_1 = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}, \qquad \mathscr{R}_2 = \left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

and that C must have the form

$$C = \begin{bmatrix} c_1 & 0 & 0 \\ 0 & 0 & c_2 \end{bmatrix}$$

with arbitrary c_1, c_2. Then

$$\det (A + BC - \lambda I) = (1 + c_1 - \lambda)(1 - \lambda)(1 + c_2 - \lambda).$$

Observe that the eigenvalue $\lambda = 1$, belonging to the eigenvector $(0, 1, 0)'$ of $A + BC$, is fixed.

To discuss the present case in general, we introduce a suitable decomposition of $\mathscr{E}$. Assume that the problem of (5.7)–(5.9) has a solution $C, \bar{\mathscr{R}}_1, \cdots, \bar{\mathscr{R}}_k$, and let $\mathbf{C}$ denote the class of matrices C for which $(A + BC)\bar{\mathscr{R}}_i \subset \bar{\mathscr{R}}_i$, $i = 1, \cdots, k$. We know that the spaces $\bar{\mathscr{R}}_i$ are the unique solutions: for simplicity of notation, write $\mathscr{R}_i$ for $\bar{\mathscr{R}}_i$. Define

$$\mathscr{E}_0 = \bigcap_{i=1}^{k} \mathscr{R}_i^*, \tag{7.17}$$

and let $\mathscr{E}_i$ be any subspace such that

$$\mathscr{R}_i = \mathscr{E}_i \oplus (\mathscr{R}_i \cap \mathscr{E}_0), \qquad i = 1, \cdots, k. \tag{7.18}$$

In the following, J denotes the set of indices $(1, \cdots, k)$, J_0 the set $(0, 1, \cdots, k)$. In intersections and summations involving $\mathscr{R}$'s, the index ranges over J; in those involving $\mathscr{E}$'s, the index ranges over J_0.

LEMMA 7.2. *The subspaces $\mathscr{E}_i$ have the properties*

$$\mathscr{E}_0 \oplus \mathscr{E}_1 \oplus \cdots \oplus \mathscr{E}_k = \mathscr{E}, \tag{7.19}$$

$$(A + BC)\mathscr{E}_i \subset \mathscr{E}_i + \mathscr{E}_0, \qquad i \in J_0, \quad C \in \mathbf{C}. \tag{7.20}$$

Proof. Assertion (7.20) is obvious by the fact that the $\mathscr{R}_i$ are $(A + BC)$-invariant. For (7.19), observe first that

$$\mathscr{R}_i \cap \mathscr{E}_0 = \mathscr{R}_i \cap \bigcap_{j \neq i} \mathscr{R}_j^* \cap \mathscr{R}_i^* = \mathscr{R}_i \cap \mathscr{R}_i^*$$

and so, if $i \in J$,

$$\begin{aligned} \mathscr{E}_i \cap \left(\mathscr{E}_0 + \sum_{\substack{j \neq 0 \\ j \neq i}} \mathscr{E}_j \right) &\subset \mathscr{E}_i \cap (\mathscr{E}_0 + \mathscr{R}_i^*) \\ &= \mathscr{E}_i \cap \mathscr{R}_i^* \\ &= \mathscr{E}_i \cap \mathscr{R}_i \cap \mathscr{R}_i^* \\ &= 0. \end{aligned} \tag{7.21}$$

Now for arbitrary subspaces $\mathscr{S}_i$, $i = 1, 2, 3$, if

$$\mathscr{S}_1 \cap (\mathscr{S}_2 + \mathscr{S}_3) = \mathscr{S}_1 \cap \mathscr{S}_2 + \mathscr{S}_1 \cap \mathscr{S}_3,$$

then

$$\mathscr{S}_2 \cap (\mathscr{S}_1 + \mathscr{S}_3) = \mathscr{S}_1 \cap \mathscr{S}_2 + \mathscr{S}_2 \cap \mathscr{S}_3.$$

Applying this fact and using (7.21) we have

$$\mathscr{E}_1 \cap \left(\mathscr{E}_0 + \sum_{j=2}^{k} \mathscr{E}_j\right) = 0 = \mathscr{E}_1 \cap \mathscr{E}_0 + \mathscr{E}_1 \cap \sum_{j=2}^{k} \mathscr{E}_j,$$

and therefore

$$\begin{aligned}\mathscr{E}_0 \cap \left(\mathscr{E}_1 + \sum_{j=2}^{k} \mathscr{E}_j\right) &= \mathscr{E}_0 \cap \mathscr{E}_1 + \mathscr{E}_0 \cap \sum_{j=2}^{k} \mathscr{E}_j \\ &= \mathscr{E}_0 \cap \sum_{j=2}^{k} \mathscr{E}_j.\end{aligned}$$

Repetition of this argument yields, after $k - 2$ steps,

$$\mathscr{E}_0 \cap \left(\mathscr{E}_1 + \sum_{j=2}^{k} \mathscr{E}_j\right) = \mathscr{E}_0 \cap \mathscr{E}_k = 0. \tag{7.22}$$

Equations (7.21), (7.22) state that the $\mathscr{E}_i$, $i \in J_0$, are independent. Finally, by (7.16),

$$\sum_{i=0}^{k} \mathscr{E}_i = \sum_{i=1}^{k} (\mathscr{E}_i + \mathscr{E}_0) \supset \sum_{i=1}^{k} \mathscr{R}_i = \mathscr{E}.$$

Remark 7. If the $\mathscr{R}_i$ are independent, then $\mathscr{E}_0 = 0$ and $\mathscr{E}_i = \mathscr{R}_i$, $i \in J$.

For $i \in J_0$ let P_i be the projection on $\mathscr{E}_i$ along $\sum_{j \neq i} \mathscr{E}_j$, and now let $C \in \mathbf{C}$ be fixed.

LEMMA 7.3. *Let* $\mathscr{B} \cap \mathscr{R}_i = \{b_i\}$, $i \in J$. *Then*

$$\mathscr{E}_i = \{P_i(A + BC) | \{P_i b_i\}\}, \qquad i \in J. \tag{7.23}$$

Proof. By (7.18) and (7.19), $\mathscr{E}_i = P_i \mathscr{R}_i$. By (7.18) and (7.20),

$$\begin{aligned}P_i \mathscr{R}_i &= P_i \sum_{j=1}^{n} (A + BC)^{j-1} \{b_i\} \\ &= \sum_{j=1}^{n} [P_i(A + BC)]^{j-1} \{P_i b_i\} \\ &= \{P_i(A + BC) | \{P_i b_i\}\}.\end{aligned}$$

LEMMA 7.4.

$$\mathscr{B} \cap \mathscr{E}_0 = 0. \tag{7.24}$$

Proof. By (7.3) and (7.7),

$$\begin{aligned}\mathscr{B} \cap \mathscr{E}_0 &= \left(\sum_{i=1}^{k} \mathscr{B} \cap \mathscr{R}_i\right) \cap \bigcap_{j=1}^{k} \mathscr{R}_j^* \\ &= \left(\mathscr{B} \cap \mathscr{R}_1 + \sum_{i=2}^{k} \mathscr{B} \cap \mathscr{R}_i\right) \cap \mathscr{R}_1^* \cap \bigcap_{j=2}^{k} \mathscr{R}_j^* \\ &= \left(\sum_{i=2}^{k} \mathscr{B} \cap \mathscr{R}_i\right) \cap \bigcap_{j=2}^{k} \mathscr{R}_j^* \\ &\vdots \\ &= \mathscr{B} \cap \mathscr{R}_k \cap \mathscr{R}_k^* = 0.\end{aligned}$$

This completes the proof of Lemma 7.4.

Next let $C = C_1$ be a fixed member of $\mathbf{C}$, let $C_2 \in \mathbf{C}$, and write $D = C_2 - C_1$; thus $A + BC_2 = A + BC_1 + BD$. Now $b_i \in \mathscr{R}_i \subset \mathscr{E}_i + \mathscr{E}_0$ $(i \in J)$; and (7.20) yields $BD\mathscr{E}_i \subset \mathscr{E}_i + \mathscr{E}_0$ $(i \in J_0)$; therefore

$$P_i b_j = 0, \qquad P_i BD\mathscr{E}_j = 0, \qquad i,j \in J, \quad i \neq j. \tag{7.25}$$

Also, using (7.24)

$$BD\mathscr{E}_0 \subset \mathscr{B} \cap \mathscr{E}_0 = 0. \tag{7.26}$$

Write

$$BD = \sum_{j=1}^{k} b_j d_j', \tag{7.27}$$

where, as before, $\{b_j\} = \mathscr{B} \cap \mathscr{R}_j$. Then

$$P_i BD\mathscr{E}_j = P_i b_i d_i' \mathscr{E}_j = 0, \qquad i \neq j, \quad i \in J, \quad j \in J_0. \tag{7.28}$$

We can now compute the spectrum Λ of $A + BC_2$. Define

$$A_i = P_i(A + BC_1), \qquad i \in J_0. \tag{7.29}$$

By (7.26) and (7.29),

$$P_0(A + BC_2) = P_0(A + BC_2)(I - P_0) + A_0 P_0, \tag{7.30}$$

and by (7.28) and (7.29),

$$\begin{aligned} P_i(A + BC_2) &= A_i + P_i BD \sum_{j=0}^{k} P_j \\ &= A_i + P_i b_i d_i' P_i, \qquad i \in J. \end{aligned} \tag{7.31}$$

Suppose $\lambda \in \Lambda$, with corresponding (complex) eigenvector ξ. A brief calculation from (7.30), (7.31) shows that either (i) for some $i \in J$, $P_i\xi \neq 0$ and $(A_i + P_i b_i d_i')P_i\xi = \lambda P_i\xi$, or (ii) $\xi = P_0\xi$ and $A_0\xi = \lambda\xi$. Conversely, if $A_0\xi = \lambda\xi$ for $0 \neq \xi \in \mathscr{E}_0$, or $(A_i + P_i b_i d_i')\xi = \lambda\xi$ for $0 \neq \xi \in \mathscr{E}_i$ and some $i \in J$, then $\lambda \in \Lambda$. Therefore

$$\Lambda = \bigcup_{i=0}^{k} \Lambda_i,$$

where Λ_i, $i \in J_0$, is the spectrum of the restriction of $P_i(A + BC_2)$ to $\mathscr{E}_i$. By (7.30), Λ_0 is independent of the choice of C_2, i.e., is fixed uniquely by the requirement $C \in \mathbf{C}$. On the other hand, for $i \in J$ Lemma 7.3 states that $\mathscr{E}_i$ is the controllability space of the pair $(A_i, P_i b_i)$. Hence, any choice of Λ_i can be realized by appropriate choice of d_i: indeed, for any $w \in \mathscr{E}$ there exists d_i such that

$$d_i' x = \begin{cases} w'x & \text{for } x \in \mathscr{E}_i, \\ 0 & \text{for } x \in \sum_{j \neq i} \mathscr{E}_j. \end{cases}$$

These results are summarized in the following theorem.

THEOREM 7.2. *Let the conditions of Theorem* 7.1 *be satisfied. If* $C \in \mathbf{C}$, *the eigenvalues of* $A + BC$ *can be partitioned into* $k + 1$ *disjoint sets*

$$\Lambda_i = \{\lambda_{i1}, \cdots, \lambda_{in_i}\}, \qquad i \in J_0,$$

where

$$n_0 = \dim\left(\bigcap_{j=1}^{k} \bar{\mathscr{R}}_j^*\right),$$

$$n_i = \dim(\bar{\mathscr{R}}_i) - \dim(\bar{\mathscr{R}}_i \cap \bar{\mathscr{R}}_i^*), \qquad i \in J.$$

The set Λ_0 *and the integers* n_i $(i \in J_0)$ *are fixed for all* $C \in \mathbf{C}$. *The sets* Λ_i $(i \in J)$ *can be assigned freely (by suitable choice of* $C \in \mathbf{C}$*) subject only to the requirement that any* λ_{ij} *with* $\operatorname{Im} \lambda_{ij} \neq 0$ *occur in* Λ_i *in a conjugate pair.*

Remark 8. If basis vectors are chosen in the $\mathscr{E}_i$, then the system differential equation can be put in a simple "normal" form. Let

$$z_i = P_i x, \qquad i \in J_0,$$

and

$$\dot{x} = (A + BC_2)x + Bv.$$

Multiplying through by P_i and using (7.30), (7.31), we obtain

$$\begin{aligned} \dot{z}_i &= (A_i + P_i b_i d_i')z_i + P_i B v, \qquad i \in J, \\ \dot{z}_0 &= P_0(A + BC_2)(z_1 + \cdots + z_k) + A_0 z_0 + P_0 B v. \end{aligned} \tag{7.32}$$

Let K be an $m \times m\,(= k \times k)$ matrix such that $BK = [b_1 \cdots b_k]$ and put $v = Kw$, $w \equiv (w_1, \cdots, w_k)'$. Since $b_i \in \mathscr{E}_i \oplus \mathscr{E}_0$, we have

$$b_i = P_i b_i + P_0 b_i \equiv \hat{b}_i + \hat{b}_{i0}.$$

Adopting n_i-dimensional representations of the z_i, etc., we see that (7.32) can be written as

$$\begin{aligned} \dot{z}_i &= (\hat{A}_i + \hat{b}_i \hat{d}_i')z_i + \hat{b}_i w_i, \qquad i \in J, \\ \dot{z}_0 &= \sum_{j=1}^{k} \hat{A}_{0j} z_j + \hat{A}_0 z_0 + \hat{B}_0 w. \end{aligned} \tag{7.33}$$

Equation (7.33) exhibits the system (2.1) as an array of k decoupled subsystems, each completely controllable by an independent scalar input w_i, plus one additional subsystem which is driven by the others and by w. Finally, since $\mathscr{R}_i \cap \mathscr{E}_0 = \mathscr{R}_i \cap \mathscr{R}_i^* \subset \mathscr{N}_i$, it follows by (5.8) and (7.18) that $\mathscr{E}_i + \mathscr{N}_i = \mathscr{E}$, that is, $H_i \mathscr{E}_i = \mathscr{H}_i$.

Remark 9. The decoupled system is acceptable in practice only if the eigenvalues in the fixed set Λ_0 are all stable. It is possible to check for stability of Λ_0 as follows. Recall that $\mathscr{R}_i \subset \mathscr{E}_i + \mathscr{E}_0$ $(i \in J)$ and note from (7.20) that $A(\mathscr{E}_i + \mathscr{E}_0) \subset \mathscr{E}_i + \mathscr{E}_0 + \mathscr{B}$ $(i \in J)$. Furthermore,

$$\mathscr{E}_i + \mathscr{E}_0 \subset \bigcap_{j \neq i} \mathscr{N}_j, \qquad i \in J.$$

It follows by Theorem 4.3 and the maximality of the $\mathscr{R}_i$ $(= \bar{\mathscr{R}}_i)$ that

$$\mathscr{R}_i = \{A + BC | \mathscr{B} \cap (\mathscr{E}_i + \mathscr{E}_0)\}$$

for any C with the property (7.20). That is, (7.20) is both necessary and sufficient that $C \in \mathbf{C}$. Thus, to compute Λ_0 it is necessary only to compute the spectrum of $A + BC_0$ (restricted to $\mathscr{E}_0$) where C_0 is any matrix such that $(A + BC_0)\mathscr{E}_0 \subset \mathscr{E}_0$.

Concluding remark. This article represents a preliminary investigation of the general decoupling problem formulated in § 5. The results for the special cases of § 6 and § 7 suggest the possibility of a complete and detailed geometric theory of linear multivariable control, in which the concept of controllability subspace would play a central role. Specific problems for future study include not only that of § 5 but also the problem of decoupling by adjunction of suitable dynamics (augmentation of the state space), and the problem of sensitivity. As formulated, decoupling represents a "hard" constraint, an all-or-nothing algebraic property. Of course, for applications a quantitative approach via "soft" constraints might also prove rewarding.

It is clear that an adequate qualitative theory of large linear multivariable systems is currently lacking; and equally clear that, with computers, such a theory would find wide application.

REFERENCES

[1] Z. V. REKASIUS, *Decoupling of multivariable systems by means of state variable feedback*, Proc. Third Allerton Conference on Circuit and System Theory, Urbana, Illinois, 1965, pp. 439–447.

[2] P. L. FALB AND W. A. WOLOVICH, *Decoupling in the design and synthesis of multivariable control systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 651–659.

[3] E. G. GILBERT, *The decoupling of multivariable systems by state feedback*, this Journal, 7 (1969), pp. 50–63.

[4] W. M. WONHAM, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 660–665.

[5] M. HEYMANN, *Pole assignment in multi-input linear systems*, Ibid., AC-13 (1968), pp. 748–749.

SIAM J. CONTROL
Vol. 8, No. 1, February 1970

TRANSFER EQUIVALENCE OF LINEAR DYNAMICAL SYSTEMS*

MICHAEL HEYMANN† AND JOHN A. THORPE‡

Abstract. The concepts of weak and strong transfer equivalence of constant (time-invariant) linear dynamical systems are defined and analyzed. The analysis leads to a simple new algorithm for constructing minimal realizations of transfer function matrices. In addition, it provides new information on the significance of the polynomial invariants which appear in the Smith–McMillan canonical form.

1. Introduction. During the past ten years, fundamental advances have been made in the structure theory of linear dynamical systems [1]–[11]. This basic work, stimulated in large measure by R. E. Kalman, has led to the development of a rigorous axiomatic theory of linear systems. In particular, the relationship between the differential equation (state variable) description and the transfer function (impulse-response) description of a constant (time-invariant) linear system is now well understood. Nevertheless, there are still important structural questions in constant linear systems theory which remain unanswered. In this paper we investigate some of these.

By a *constant linear dynamical system* we shall mean a triple (F, G, H), where F is a real square matrix and G and H are real rectangular matrices of appropriate sizes so that the matrix product HFG is defined. Thus (F, G, H) is the basic data required to describe a system of constant coefficient linear differential equations of the form

$$\begin{aligned} \dot{x} &= Fx + Gu, \\ y &= Hx \end{aligned} \tag{1}$$

relating an input vector $u = u(t)$ to an output vector $y = y(t)$ through a state vector $x = x(t)$.

Assuming that the system (1) starts at rest at time $t = 0$, the Laplace transforms $Y = Y(s)$ and $U = U(s)$ of y and u are related by

$$Y = ZU,$$

where $Z = Z(s)$ is the *transfer function matrix* of the system (F, G, H) and is given by

$$Z(s) = H(Is - F)^{-1}G. \tag{2}$$

The matrix Z is a proper rational matrix; that is, each entry in Z is a quotient of polynomials in s with the degree of the numerator lower than that of the denominator. The matrix Z exhibits the transfer (input–output) behavior of the system but suppresses the internal (state) behavior.

* Received by the editors January 23, 1969. This research was done at Mobil Research and Development Corporation, Central Research Division Laboratory, Princeton, New Jersey.

† Department of Chemical Engineering, University of the Negev, Beer Sheba, Israel.

‡ Department of Mathematics, State University of New York at Stony Brook, Stony Brook, New York 11790.

Thus, to each system (F, G, H) is associated a unique proper rational matrix Z, its transfer function matrix. However, to each proper rational matrix Z there is associated a whole class of systems, called *realizations* of Z, each having Z as transfer function matrix. These systems share the same input–output behavior but can differ internally. In particular, they can differ in the dimension of the state space (size of F). The realizations of Z having the smallest possible state space dimension are of particular interest; these are the *minimal realizations* of Z.

Various algorithms for constructing minimal realizations of transfer function matrices have been given [4], [5], [8]. One of these, described by Kalman, requires the reduction of the transfer function matrix by means of elementary row and column operations to Smith–McMillan form [12], a diagonal form in which certain divisibility conditions hold. We shall describe a related algorithm for constructing minimal realizations which requires only the reduction of the transfer function matrix to diagonal form.

If, in our algorithm, we reduce the transfer function matrix to the Smith–McMillan form, we obtain the same realization as the one obtained by Kalman [5]. However, our algorithm does not coincide with Kalman's even in that case. In fact, our algorithm then yields new information about the significance of the polynomial invariants ε_i which appear as numerators in the Smith–McMillan form. (Indeed, it was the problem of interpreting these invariants which motivated our research.) We are able to exhibit directly the role played in the *output structure* by the proper parts of the polynomials ε_i; that is, by the remainders obtained from the numerators ε_i after division by the corresponding denominators ψ_i in the Smith–McMillan form. Our results seem to indicate that these proper parts are more basic to linear systems theory than are the ε_i themselves.

The basic tool which we shall use is *transfer equivalence*. Systems are called strongly transfer equivalent if their transfer function matrices have the same Smith–McMillan form. A more fundamental concept, called weak transfer equivalence, is also defined. An important property of weak transfer equivalence is that in each weak transfer equivalence class there are systems which are completely uncoupled.

This paper is organized as follows. We begin (§ 2) by developing the basic properties of weak and strong transfer equivalence. In § 3 we derive the complete analytic relationship between any pair of controllable and observable systems which are weakly transfer equivalent. These results are applied in § 4 to obtain our algorithm for constructing minimal realizations. We conclude (§ 5) with a discussion of various related topics. In particular, we discuss two methods for "realizing" improper rational matrices. We also discuss an interpretation of the invariants ε_i which exhibits the polynomials ε_i themselves and not just their proper parts.

2. Transfer equivalence. Let $Z = Z(s)$ be a rational matrix; that is, Z is a matrix whose entries are quotients of polynomials in s with real coefficients. Associated with Z is a diagonal matrix $\Lambda = \Lambda(s)$, called the Smith–McMillan canonical form [12] of Z, obtained as follows. By letting $\psi = \psi(s)$ denote the monic polynomial (leading coefficient 1) which is the least common denominator of the entries of Z, the matrix ψZ is a polynomial matrix. By applying a sequence

of elementary row and column operations (that is, operations which: (i) interchange two rows or columns, (ii) multiply a row or column by a nonzero real number, or (iii) add a polynomial multiple of one row or column to another) to ψZ we can obtain a unique (independent of the row and column operations used) diagonal matrix.

$$\Gamma = \operatorname{diag}[\gamma_1, \gamma_2, \cdots, \gamma_R, 0, \cdots, 0] \tag{3}$$

such that each diagonal element $\gamma_i = \gamma_i(s)$ is a monic polynomial which divides its successor γ_{i+1}, $i = 1, \cdots, R-1$. The matrices ψZ and Γ are related by

$$\psi Z = A\Gamma B, \tag{4}$$

where $A = A(s)$ and $B = B(s)$ are polynomial matrices with constant nonzero determinants [13]. Dividing both sides of this equation by ψ and reducing each polynomial fraction γ_i/ψ by cancellation of common factors, we obtain

$$Z = A\Lambda B, \tag{5}$$

where

$$\Lambda = \operatorname{diag}[\varepsilon_1/\psi_1, \cdots, \varepsilon_R/\psi_R, 0, \cdots, 0], \tag{6}$$

and where the $\varepsilon_i = \varepsilon_i(s)$ and the $\psi_i = \psi_i(s)$ are monic polynomials, with ε_i dividing ε_{i+1} and ψ_{i+1} dividing ψ_i for each i, $1 \leqq i \leqq R-1$, such that each pair (ε_i, ψ_i) is relatively prime. The matrix Λ is the Smith–McMillan form of Z.

DEFINITION. Two rational matrices Z and $\tilde{Z}$ are called *strongly equivalent* if there exist polynomial matrices A and B with constant nonzero determinants such that $\tilde{Z} = AZB$. Two constant linear dynamical systems are said to be *strongly transfer equivalent* if their transfer function matrices are strongly equivalent.

Note that, since products and inverses of polynomial matrices with constant nonzero determinants are again polynomial matrices with constant nonzero determinants, strong equivalence and strong transfer equivalence are equivalence relations.

Clearly, each rational matrix is strongly equivalent to a unique Smith–McMillan canonical matrix; that is, each strong equivalence class contains exactly one Smith–McMillan form. It follows that the rank R together with the $2R$ polynomials $\psi_1, \cdots, \psi_R, \varepsilon_1, \cdots, \varepsilon_R$ form a complete set of invariants for strong equivalence of rational matrices.

A basic handicap, from the system theoretic point of view, of the notion of strong transfer equivalence is that, although each rational *matrix* is strongly equivalent to a Smith–McMillan form, it is not true that each linear dynamical *system* is strongly *transfer* equivalent to a *system* whose transfer function matrix is in Smith–McMillan form. This is a consequence of the fact that the Smith–McMillan form is not, in general, proper; that is, the degrees of the numerators in the Smith–McMillan form need not be lower than the degrees of the corresponding denominators. However, this drawback can be eliminated by weakening the notion of equivalence.

First recall that two polynomials $\alpha = \alpha(s)$ and $\beta = \beta(s)$ are said to be congruent modulo the polynomial $\psi = \psi(s)$, written $\alpha \equiv \beta \pmod{\psi}$, provided that α and β have the same remainder after division by ψ. Similarly, two polynomial matrices

$A = A(s)$ and $B = B(s)$ are said to be congruent modulo ψ, written $A \equiv B \pmod{\psi}$, provided that corresponding entries of A and B have the same remainders after division by ψ. Thus $A \equiv B \pmod{\psi}$ if and only if there exists a polynomial matrix $C = C(s)$ such that $A = B + \psi C$.

DEFINITION. Let Z and $\tilde{Z}$ be rational matrices and let ψ and $\tilde{\psi}$ be, respectively, the least common denominators of the entries of Z and $\tilde{Z}$. Z and $\tilde{Z}$ are called *weakly equivalent* if

(i) $\psi = \tilde{\psi}$ and

(ii) there exist polynomial matrices A and B with constant nonzero determinants such that $\psi\tilde{Z} \equiv A(\psi Z)B \pmod{\psi}$.

Two constant linear dynamical systems are said to be *weakly transfer equivalent* if their transfer function matrices are weakly equivalent.

It is clear that weak equivalence and weak transfer equivalence are equivalence relations. It is also clear that strong equivalence implies weak equivalence (the equality of the least common denominators ψ and $\tilde{\psi}$ for strongly equivalent Z and $\tilde{Z}$ follows from the uniqueness of the Smith–McMillan form (6) and from the fact that, in (6), $\psi_1 = \psi$).

Note that, although a given constant linear dynamical system will not in general be *strongly* transfer equivalent to a system which is completely uncoupled, that is, one whose transfer function matrix is diagonal, each constant linear dynamical system will be *weakly* transfer equivalent to a system which is completely uncoupled. Indeed, one need only take any (possibly improper) diagonal matrix D which is strongly equivalent to the given transfer function matrix Z and reduce it by replacing each entry in ψD by its remainder after division by ψ to obtain $\psi D'$, where D' is a proper rational diagonal matrix which is weakly equivalent to Z. Any system with D' as transfer function matrix will then be weakly transfer equivalent to the given system.

The matrix D' will be called the *proper part* of the matrix D. Note that D' can also be obtained by replacing the numerator of each entry in D by its remainder after division by the corresponding denominator. Indeed, if α/β is any quotient of polynomials with β dividing ψ, then the remainder after dividing $\psi(\alpha/\beta)$ by ψ is just $\psi(\alpha'/\beta)$, where α' is the remainder after dividing α by β.

The proper part Λ' of the Smith–McMillan form Λ of a transfer function matrix Z will be of special importance. If Λ is of the form (6), then Λ' is of the form

$$\Lambda' = \operatorname{diag}[\varepsilon_1'/\psi_1, \cdots, \varepsilon_r'/\psi_r, 0, \cdots, 0], \tag{7}$$

where $r = \max\{i | 1 \leqq i \leqq R, \psi_i \neq 1\}$ and where ε_i' is the remainder after dividing ε_i by ψ_i. We shall call Λ' the *reduced Smith–McMillan form* of Z.

3. External equivalence. In this section we investigate the relationship between linear dynamical systems which are weakly transfer equivalent. For this we shall need two facts about the rational matrix $(Is - F)^{-1}$, where F is a square matrix.

(i) Let $\psi(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_0$ denote the minimal polynomial of F; that is, ψ is the monic polynomial of least degree such that $\psi(F) = 0$. Then $(Is - F)^{-1}$ is given by the formula

$$(Is - F)^{-1} = \frac{1}{\psi(s)} \sum_{k=0}^{n-1} \varphi_k(F)s^k, \tag{8}$$

where the φ_k are the polynomials

$$\varphi_k(x) = x^{n-k-1} + a_{n-1}x^{n-k-2} + \cdots + a_{k+1}, \qquad k = 0, 1, \cdots, n-1. \tag{9}$$

Indeed, an elementary computation shows that

$$\psi(s)I = \psi(s)I - \psi(F) = \left[\sum_{k=0}^{n-1} \varphi_k(F)s^k\right](Is - F),$$

and then right multiplication by $(Is - F)^{-1}$ yields (8).

(ii) For each nonnegative integer k we have

$$s^k\psi(s)(Is - F)^{-1} \equiv F^k\psi(s)(Is - F)^{-1} \pmod{\psi} \tag{10}$$

and

$$\psi(s)(Is - F)^{-1}s^k \equiv \psi(s)(Is - F)^{-1}F^k \pmod{\psi}. \tag{11}$$

Indeed,

$$s\psi(s)(Is - F)^{-1} - F\psi(s)(Is - F)^{-1} = (Is - F)\psi(s)(Is - F)^{-1} = \psi(s)I \equiv 0 \pmod{\psi},$$

so (10) is valid for $k = 1$. An elementary induction argument establishes (10) in general. Formula (11) is a consequence of (10) and the fact that F commutes with $(Is - F)^{-1}$.

THEOREM 1. *Let $(\tilde{F}, \tilde{G}, \tilde{H})$ be a constant linear dynamical system. Let A and B be polynomial matrices (not necessarily square) of appropriate sizes so that the matrix products $A\tilde{H}$ and $\tilde{G}B$ are defined. Let $A = \sum A_i s^i$ and $B = \sum B_j s^j$ express A and B as matrix polynomials. Define (F, G, H) to be the constant linear dynamical system given by*

$$F = \tilde{F}, \qquad G = \sum \tilde{F}^j\tilde{G}B_j, \qquad H = \sum A_i\tilde{H}\tilde{F}^i. \tag{12}$$

Then the transfer function matrix Z of the system (F, G, H) is related to the transfer function matrix $\tilde{Z}$ of $(\tilde{F}, \tilde{G}, \tilde{H})$ by

$$\psi Z \equiv A(\psi\tilde{Z})B \pmod{\psi}, \tag{13}$$

where ψ is the minimal polynomial of F.

Proof. First note that, by (8), ψZ and $A(\psi\tilde{Z})B$ are polynomial matrices so that it makes sense to ask if they are congruent (mod ψ). By (10) and (11) we have

$$\begin{aligned} Z &= H(Is - F)^{-1}G = (\textstyle\sum A_i\tilde{H}\tilde{F}^i)(Is - \tilde{F})^{-1}(\sum \tilde{F}^j\tilde{G}B_j) \\ &\equiv (\textstyle\sum A_i\tilde{H}s^i)(Is - \tilde{F})^{-1}(\sum s^j\tilde{G}B_j) \pmod{\psi} \\ &= (\textstyle\sum A_i s^i)\tilde{H}(Is - \tilde{F})^{-1}\tilde{G}(\sum B_j s^j) \\ &= A\tilde{Z}B, \end{aligned}$$

as claimed.

DEFINITION. Two constant linear dynamical systems (F, G, H) and $(\tilde{F}, \tilde{G}, \tilde{H})$ are called *externally equivalent* if

$$\tilde{F} = F, \qquad \tilde{G} = \sum F^jGB_j, \qquad \tilde{H} = \sum A_iHF^i, \tag{14}$$

where $A = \sum A_i s^i$ and $B = \sum B_j s^j$ are (square) polynomial matrices with constant nonzero determinants.

An elementary computation shows that external equivalence is an equivalence relation. In particular, if $(\tilde{F}, \tilde{G}, \tilde{H})$ is related to (F, G, H) by equations of the form (14), then (F, G, H) is related to $(\tilde{F}, \tilde{G}, \tilde{H})$ by equations of the same form.

One consequence of Theorem 1 is that systems which are externally equivalent are also weakly transfer equivalent. Thus weak transfer equivalence is a weaker notion of system equivalence than external equivalence.

Note that the polynomial matrices A and B in the definition of external equivalence may always be taken to be of degree less than n, the degree of the minimal polynomial of F. This is because each power of F is expressible as a polynomial (with scalar coefficients) in F of degree less than n. Thus, given a system (F, G, H), every system $(\tilde{F}, \tilde{G}, \tilde{H})$ externally equivalent to (F, G, H) is given by

$$\tilde{F} = F, \qquad \tilde{G} = \sum_{j=0}^{n-1} F^j G B_j, \qquad \tilde{H} = \sum_{i=0}^{n-1} A_i H F^i$$

for appropriate polynomial matrices

$$A = \sum_{i=0}^{n-1} A_i s^i \quad \text{and} \quad B = \sum_{j=0}^{n-1} B_j s^j$$

with constant determinants. These expressions for $\tilde{G}$ and $\tilde{H}$ can be rewritten as

$$\tilde{G} = [G, FG, \cdots, F^{n-1}G] \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_{n-1} \end{bmatrix} \quad \text{and} \quad \tilde{H} = [A_0, \cdots, A_{n-1}] \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{n-1} \end{bmatrix}.$$

The matrix

$$[G, FG, \cdots, F^{n-1}G] \tag{15}$$

is known [7] as the controllability matrix of the system (F, G, H), and the matrix

$$\begin{bmatrix} H \\ HF \\ \vdots \\ HF^{n-1} \end{bmatrix} \tag{16}$$

is known as the observability matrix of the system. Thus externally equivalent systems are related to one another by certain matrix operations on the controllability and observability matrices.

An important property of external equivalence is that externally equivalent systems have the same controllability and observability properties. Recall [7] that a constant linear dynamical system (F, G, H) is *controllable* if the controllability matrix (15) is of maximal rank; that is, if the matrix (15) has n linearly independent rows. Similarly, (F, G, H) is *observable* if the observability matrix (16) is of maximal rank (has n linearly independent columns).

THEOREM 2. *Let (F, G, H) and $(\tilde{F}, \tilde{G}, \tilde{H})$ be externally equivalent constant linear dynamical systems. Then their controllability matrices have equal rank and their observability matrices have equal rank. In particular, $(\tilde{F}, \tilde{G}, \tilde{H})$ is controllable if and only if (F, G, H) is controllable and $(\tilde{F}, \tilde{G}, \tilde{H})$ is observable if and only if (F, G, H) is observable.*

Proof. Since (F, G, H) and $(\tilde{F}, \tilde{G}, \tilde{H})$ are externally equivalent, we have $\tilde{F} = F$, $\tilde{G} = \sum F^j G B_j$, and $\tilde{H} = \sum A_i H F^i$ for some polynomial matrices $A = \sum A_i s^i$ and $B = \sum B_j s^j$ with constant determinants. Suppose c is a row vector such that

$$c[G, FG, \cdots, F^{n-1}G] = 0.$$

Then, since each power of F can be expressed as a polynomial in F of degree less than n, it follows that $cF^kG = 0$ for all nonnegative integers k. Hence

$$cF^k\tilde{G} = \sum cF^{k+j}GB_j = 0$$

for all $k \geqq 0$ and so

$$c[\tilde{G}, F\tilde{G}, \cdots, F^{n-1}\tilde{G}] = 0.$$

Thus, viewing $K = [G, FG, \cdots, F^{n-1}G]$ and $\tilde{K} = [\tilde{G}, F\tilde{G}, \cdots, F^{n-1}\tilde{G}]$ as linear operators acting on row vectors, we see that the null space of K is contained in the null space of $\tilde{K}$. But, since external equivalence is a symmetric relation, it follows that the null space of $\tilde{K}$ is also contained in the null space of K; that is, these null spaces must be equal. By the rank and nullity theorem of linear algebra, we conclude that the controllability matrices K and $\tilde{K}$ have the same rank. The proof for observability is similar.

We shall call two constant linear dynamical systems (F, G, H) and $(\tilde{F}, \tilde{G}, \tilde{H})$ *internally isomorphic* if there exists a nonsingular matrix T such that $\tilde{F} = TFT^{-1}$, $\tilde{G} = TG$, and $\tilde{H} = HT^{-1}$. Clearly, internally isomorphic systems have the same transfer function matrix. Conversely, it is well known [6] that any pair of controllable and observable systems which have the same transfer function matrix are internally isomorphic.

THEOREM 3 (Basic equivalence theorem). *Two controllable and observable constant linear dynamical systems are weakly transfer equivalent if and only if they differ (at most) by an external equivalence and an internal isomorphism.*

Proof. The sufficiency part is clear from Theorem 1. To prove necessity, suppose $(\hat{F}, \hat{G}, \hat{H})$ and $(\tilde{F}, \tilde{G}, \tilde{H})$ are weakly transfer equivalent controllable and observable systems. Then their transfer function matrices $\hat{Z}$ and $\tilde{Z}$ are related by $\psi\hat{Z} \equiv A(\psi\tilde{Z})B \pmod{\psi}$, where $A = \sum A_i s^i$ and $B = \sum B_j s^j$ are polynomial matrices with constant nonzero determinants and where ψ is the least common denominator of the entries both of $\hat{Z}$ and of $\tilde{Z}$. But it is known [5] that the least

common denominator of the entries in the transfer function matrix of any controllable and observable system (F, G, H) is the minimal polynomial of the state matrix F. (This fact will also follow from our Theorem 4 below.) Thus, ψ is the minimal polynomial of both $\hat{F}$ and $\tilde{F}$. Let (F, G, H) be constructed from $\tilde{F}, \tilde{G}, \tilde{H}, A$ and B as in formula (12), and let Z denote its transfer function matrix. By Theorem 1, $\psi Z \equiv A(\psi\tilde{Z})B \equiv \psi\hat{Z} \pmod{\psi}$. But, since Z and $\hat{Z}$ are both proper, this implies that $\psi Z = \psi\hat{Z}$ and hence $Z = \hat{Z}$. Furthermore, by Theorem 2, both $(\hat{F}, \hat{G}, \hat{H})$ and (F, G, H) are controllable and observable. It follows that $(\hat{F}, \hat{G}, \hat{H})$ and (F, G, H) are internally isomorphic. Since (F, G, H) and $(\tilde{F}, \tilde{G}, \tilde{H})$ are externally equivalent, this completes the proof.

Explicitly, the systems $(\hat{F}, \hat{G}, \hat{H})$ and $(\tilde{F}, \tilde{G}, \tilde{H})$ are related by

$$\hat{F} = T\tilde{F}T^{-1}, \qquad \hat{G} = T\sum \tilde{F}^j\tilde{G}B_j, \qquad \hat{H} = \sum A_i\tilde{H}\tilde{F}^iT^{-1}$$

for some nonsingular matrix T.

4. Realization theory. It is well known [3] that a constant linear dynamical system is controllable and observable if and only if it is a minimal realization of its transfer function matrix. Thus, in order to construct a minimal realization of a given proper rational matrix, one must construct a system which is controllable and observable and has the given matrix as its transfer function matrix. Since each proper rational matrix is weakly equivalent to a proper rational diagonal matrix, it suffices (in view of Theorems 1 and 2) to minimally realize proper diagonal matrices. But this is easily accomplished by taking direct sums of systems realizing the diagonal entries.

We shall say that a system (F, G, H) is the *direct sum* of the systems (F_i, G_i, H_i), $i = 1, \cdots, k$, if

$$F = \begin{bmatrix} F_1 & & & \\ & F_2 & & \\ & & \ddots & \\ & & & F_k \end{bmatrix}, \quad G = \begin{bmatrix} G_1 & & & \\ & G_2 & & \\ & & \ddots & \\ & & & G_k \end{bmatrix}, \quad H = \begin{bmatrix} H_1 & & & \\ & H_2 & & \\ & & \ddots & \\ & & & H_k \end{bmatrix}.$$

Clearly the direct sum (F, G, H) of the systems (F_i, G_i, H_i) is controllable if and only if each (F_i, G_i, H_i) is controllable, and is observable if and only if each (F_i, G_i, H_i) is observable. (Note that our concept of direct sum is different from that adopted, e.g., by Kalman [5].)

THEOREM 4 (Realization theorem). *Given a proper rational matrix Z, let D' be any proper rational diagonal matrix which is weakly equivalent to Z, say $D' = \operatorname{diag}[\varepsilon_1'/\psi_1, \cdots, \varepsilon_r'/\psi_r, 0, \cdots, 0]$ where, for each i, ε_i' and ψ_i are relatively prime, $\varepsilon_i' \neq 0$ and ψ_i is monic. Let A and B be polynomial matrices with constant determinants such that $\psi Z \equiv A(\psi D')B \pmod{\psi}$, where ψ is the least common denominator of the entries in Z. Construct a system (F, G, H) as follows.*

(a) *For each* $i = 1, \cdots, r$, *let* (F_i, G_i, H_i) *be the system given by*

$$F_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -a_{i0} & -a_{i1} & -a_{i2} & \cdots & -a_{i,n_i-1} \end{bmatrix},$$

$$G_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \qquad H_i = [b_{i0}, b_{i1}, \cdots, b_{i,m_i}, 0, \cdots, 0], \tag{17}$$

where the a_{ij} *and the* b_{ij} *are the coefficients of* ψ_i *and of* ε_i' *respectively (i.e.,* $\psi_i = \sum_{j=0}^{n_i} a_{ij}s^j$ *where* $a_{i,n_i} = 1$ *and* $\varepsilon_i' = \sum_{j=0}^{m_i} b_{ij}s^j$*). (These systems will be minimal realizations of the* 1×1 *matrices* $[\varepsilon_i'/\psi_i]$.)

(b) *Define* $(\tilde{F}, \tilde{G}, \tilde{H})$ *to be the system obtained by taking the direct sum of the systems* (F_i, G_i, H_i) *and then augmenting, if necessary, by adding columns of zeros to* $\tilde{G}$ *and/or rows of zeros to* $\tilde{H}$ *to make* $\tilde{H}(Is - \tilde{F})^{-1}\tilde{G}$ *of the same size as* Z*. (This system will be a minimal realization of* D'.)

(c) *Let* (F, G, H) *be obtained from* $\tilde{F}$, $\tilde{G}$, $\tilde{H}$, A *and* B *as in formula* (12).

Then (F, G, H) *is a minimal realization of* Z.

Remark 1. The diagonal matrix D' can be taken to be the proper part of any diagonal matrix D which is strongly equivalent to Z. In particular, D' can be taken to be the reduced Smith–McMillan form (7) of Z. However, we do not require that D' be this canonical form (for example, we do not require any divisibility relations among the ψ_i). In fact, the canonical form (3) is not in general the first diagonal matrix encountered in the standard algorithm [13] for reducing a polynomial matrix to canonical form and so, in general, the reduced Smith–McMillan form may not be the most convenient diagonal matrix to use in the realization procedure described above.

Remark 2. It is clear from Theorem 4 and the fact that two controllable and observable systems realizing the same Z can differ at most by an internal isomorphism, that the dimension of the state space (size of F) for any minimal realization of Z is equal to $\sum_{i=1}^{r} n_i$, the sum of the degrees of the denominators ψ_i appearing in D'. This number $\sum_{i=1}^{r} n_i$ is known as the (McMillan) degree of the rational matrix Z (see [4], [12]).

Remark 3. Note that the minimal polynomial of the matrix F constructed in Theorem 4 is the least common multiple of the polynomials ψ_i appearing in D', since each F_i has minimal polynomial ψ_i. Thus the minimal polynomial of F is equal to the least common denominator of the entries in D' and, by weak equivalence, also of the entries in Z. Since minimal realizations of Z differ only by internal isomorphisms, it follows that the least common denominator of the entries in Z is equal to the minimal polynomial of the state matrix of any minimal realization of Z. (This known fact was used in our proof of Theorem 3. Note that Theorem 3 is not used in the proof given below for Theorem 4.)

In order to prove Theorem 4, we shall need the following lemma.

LEMMA. *Let F be the companion matrix*

$$F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}$$

associated with the polynomial $\psi(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_0$. Then

$$\psi(s)(Is - F)^{-1} = \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{n-1} \end{bmatrix} [\varphi_0(s), \cdots, \varphi_{n-1}(s)] - \psi(s) \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ s & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ s^{n-2} & s^{n-3} & \cdots & 1 & 0 \end{bmatrix}, \tag{18}$$

where the φ_k, $k = 0, 1, \cdots, n-1$, are as in (9).

Proof. Multiply both sides of (18) on the left by $Is - F$ and compute.

Proof of Theorem 4. First, using (18) with $F = F_i$ and $\psi = \psi_i$, we see that

$$\psi_i(s)H_i(Is - F_i)^{-1}G_i = b_{i0} + b_{i1}s + \cdots + b_{i,m_i}s^{m_i}$$

and hence the system (F_i, G_i, H_i) of (17) is a realization of the 1×1 proper rational matrix $[\varepsilon_i'/\psi_i]$. Moreover, it is a minimal realization because, given any realization $(\hat{F}_i, \hat{G}_i, \hat{H}_i)$ of $[\varepsilon_i'/\psi_i]$, the matrix $\hat{F}_i$ must be of size $\hat{n}_i \times \hat{n}_i$ for some $\hat{n}_i \geqq n_i$. Indeed, for $\hat{F}_i$ of size $\hat{n}_i \times \hat{n}_i$, its minimal polynomial $\hat{\psi}_i$ is of degree $\leqq \hat{n}_i$ and, by (8) (or (18)),

$$\hat{\psi}_i\hat{H}_i(Is - \hat{F}_i)^{-1}\hat{G}_i = \hat{\psi}_i(\varepsilon_i'/\psi_i)$$

is a polynomial matrix, so ψ_i divides $\hat{\psi}_i$ and hence

$$\hat{n}_i = \deg \hat{\psi}_i \geqq \deg \psi_i = n_i.$$

Thus each (F_i, G_i, H_i) is controllable and observable. Since the direct sum of controllable and observable systems is controllable and observable, it follows that the system $(\tilde{F}, \tilde{G}, \tilde{H})$ constructed in (b) is controllable and observable and is a minimal realization of D'.

Finally, by Theorem 1 (with $\tilde{Z}$ replaced by D'), we see that the system (F, G, H) constructed in (c) is a realization of Z. By Theorem 2, it is controllable and observable; that is, it is a minimal realization of Z.

Remark 4. It may be of interest to decompose step (a) in the realization procedure into two substeps as follows. First realize the rational matrix $[1/\psi_i]$ with the system $(F_i, G_i, \hat{H}_i)$, where F_i and G_i are as in (17) and where $\hat{H}_i = [1, 0, \cdots, 0]$. Then construct H_i from $\hat{H}_i$ by the formula $H_i = \hat{H}_i\varepsilon_i'(F_i)$ to achieve the realization of $[\varepsilon_i'/\psi_i]$.

More generally, given a system (F, G, H) realizing a proper rational matrix Z and given a polynomial ε such that εZ is still proper, a system $(\tilde{F}, \tilde{G}, \tilde{H})$ realizing εZ is obtained by taking $\tilde{F} = F$, $\tilde{G} = G$ and $\tilde{H} = H\varepsilon(F)$ or, alternatively, by taking $\tilde{F} = F$, $\tilde{G} = \varepsilon(F)G$, and $\tilde{H} = H$. That these systems do realize εZ is an immediate consequence of Theorem 1.

To illustrate the realization procedure, we consider two examples.

Example 1. Let

$$Z = \frac{1}{\psi}\begin{bmatrix} s+1 & 2s^2+s-1 & s^2-1 \\ -s^2-s & -s^2+s & s \end{bmatrix},$$

where $\psi(s) = s^3 + 3s^2 + 2s = s(s+1)(s+2)$. Then we have $Z = ADB$, where

$$A = \begin{bmatrix} 1 & 0 \\ -s & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} 1/(s^2+2s) & 0 & 0 \\ 0 & s^2/(s^2+3s+2) & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} 1 & 2s-1 & s-1 \\ 0 & 2 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Denoting by D' the proper part of the matrix D we have that D' is weakly equivalent to Z and, in fact, $\psi Z \equiv A(\psi D')B$, where A and B are as above. The matrix D' is given by

$$D' = \begin{bmatrix} 1/(s^2+2s) & 0 & 0 \\ 0 & (-3s-2)/(s^2+3s+2) & 0 \end{bmatrix}.$$

According to parts (a) and (b) of Theorem 4, a minimal realization $(\tilde{F}, \tilde{G}, \tilde{H})$ of D' is given by

$$\tilde{F} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -2 & -3 \end{bmatrix}, \qquad \tilde{G} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \tilde{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & -2 & -3 \end{bmatrix}.$$

Since

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix} s = A_0 + A_1 s$$

and

$$B = \begin{bmatrix} 1 & -1 & -1 \\ 0 & 2 & 1 \\ 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} s = B_0 + B_1 s,$$

part (c) of Theorem 4 yields the following minimal realization of Z:

$$F = \tilde{F} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -2 & -3 \end{bmatrix},$$

$$G = \tilde{G}\begin{bmatrix} 1 & -1 & -1 \\ 0 & 2 & 1 \\ 0 & 1 & 0 \end{bmatrix} + \tilde{F}\tilde{G}\begin{bmatrix} 0 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 1 \\ 1 & -5 & -3 \\ 0 & 0 & 0 \\ 0 & 2 & 1 \end{bmatrix},$$

$$H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\tilde{H} + \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}\tilde{H}\tilde{F} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & -2 & -3 \end{bmatrix}.$$

Example 2. Let

$$Z = \frac{s-2}{s^4-1}\begin{bmatrix} 3s^2 - 5s - 2 & 2s^2 - 5s - 1 \\ 4s^2 - 2s + 4 & 3s^2 - 2s + 5 \end{bmatrix}.$$

Then $Z = ADB$, where

$$A = \begin{bmatrix} -24 & 0 \\ -18 & -5/24 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 33 & 0 \end{bmatrix}s + \begin{bmatrix} 0 & 0 \\ 16 & 0 \end{bmatrix}s^2 + \begin{bmatrix} 0 & 0 \\ 5 & 0 \end{bmatrix}s^3,$$

$$D = \begin{bmatrix} (s-2)/(s^4-1) & 0 \\ 0 & (s^3 + s^2 - 12)/(s^2 + 1) \end{bmatrix},$$

$$B = \begin{bmatrix} 1/12 & 1/24 \\ 22/5 & 23/5 \end{bmatrix} + \begin{bmatrix} 5/24 & 5/24 \\ -3 & -2 \end{bmatrix}s + \begin{bmatrix} -1/8 & -1/12 \\ 0 & 0 \end{bmatrix}s^2.$$

In this case, the matrix D is the Smith–McMillan form (6) of Z. The proper part D' of D is then the reduced Smith–McMillan form (7) of Z and is given by

$$D' = \begin{bmatrix} (s-2)/(s^4-1) & 0 \\ 0 & (-s - 13)/(s^2 + 1) \end{bmatrix}.$$

Following the realization procedure yields

$$\tilde{H} = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -13 & -1 \end{bmatrix},$$

$$\tilde{G} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix},$$

$$F = \tilde{F} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix},$$

$$G = \begin{bmatrix} 0 & 0 \\ -1/8 & -1/12 \\ 5/24 & 5/24 \\ 1/12 & 1/24 \\ -3 & -2 \\ 22/5 & 23/5 \end{bmatrix},$$

$$H = \begin{bmatrix} 48 & -24 & 0 & 0 & 0 & 0 \\ 41 & -84 & 1 & 6 & 65/24 & 5/24 \end{bmatrix}.$$

Remark 5. As mentioned in § 3, the polynomial matrices which are used in this realization procedure can always be chosen to be of degree less than the degree of ψ. In practice, however, it may be as easy to make the extra matrix computations required as to reduce modulo ψ the polynomials involved.

Remark 6. Kalman [5] has described another realization procedure which will give the same result as ours when D' is taken to be the reduced Smith–McMillan form of Z. His procedure is based on the formula

$$\psi(s)(Is - F)^{-1} \equiv \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{n-1} \end{bmatrix} [\varphi_0(s), \cdots, \varphi_{n-1}(s)] \pmod{\psi}, \tag{19}$$

where F is the companion matrix with minimal polynomial ψ and $\varphi_0, \cdots, \varphi_{n-1}$ are as in (9). This formula follows immediately from (18). From (19) it follows that the transfer function matrix Z of a system (F, G, H) (F a companion matrix) satisfies

$$\psi(s)Z(s) \equiv H \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{n-1} \end{bmatrix} [\varphi_0(s), \cdots, \varphi_{n-1}(s)]G \pmod{\psi}.$$

Hence a minimal realization of $Z = A(\varepsilon/\psi)B$, where A is a 1-column polynomial matrix and B is a 1-row polynomial matrix, is obtained by taking F to be the

companion matrix associated with ψ and solving the equations

$$(20)\quad H\begin{bmatrix}1\\ s\\ \vdots\\ s^{n-1}\end{bmatrix} \equiv A(s)\varepsilon(s) \pmod{\psi} \quad \text{and} \quad [\varphi_0(s), \cdots, \varphi_{n-1}(s)]G \equiv B(s) \pmod{\psi}$$

for H and G. (Kalman actually puts ε with B instead of with A in these equations, but points out that it can go either place.) A minimal realization of a general Z is obtained expressing $Z = A\Lambda B$, where Λ is the Smith–McMillan form of Z as in (6), rewriting this equation as

$$Z = \sum_{i=1}^{r} A^{(i)}\frac{\varepsilon_i}{\psi_i}B^{(i)} + \sum_{i=r+1}^{R} A^{(i)}\varepsilon_i B^{(i)}$$

where the $A^{(i)}$ are the columns of A (*not* the coefficients of A as a matrix polynomial) and the $B^{(i)}$ are the rows of B, and observing that the sum (in Kalman's sense) of the realizations of the $A^{(i)}(\varepsilon_i/\psi_i)B^{(i)}$, $i = 1, \cdots, r$, then gives a minimal realization of Z. That Kalman's procedure gives the same result as ours is a consequence of the fact that, for transfer functions of the form $Z = A(\varepsilon/\psi)B$, our realization satisfies Kalman's equations (20). Indeed, taking F to be the companion matrix of ψ, we see that our H satisfies (see (12), (17) and (19))

$$\begin{aligned}
H\begin{bmatrix}1\\ s\\ \vdots\\ s^{n-1}\end{bmatrix}[\varphi_0(s), \cdots, \varphi_{n-1}(s)] &\equiv \sum A_i[b_0, \cdots, b_m, 0, \cdots, 0]F^i\psi(s)(Is - F)^{-1} \\
&\equiv \sum A_i s^i[b_0, \cdots, b_m, 0, \cdots, 0]\psi(s)(Is - F)^{-1} \qquad \text{(by (10))} \\
&\equiv A(s)[b_0, \cdots, b_m, 0, \cdots, 0]\begin{bmatrix}1\\ s\\ \vdots\\ s^{n-1}\end{bmatrix}[\varphi_0(s), \cdots, \varphi_{n-1}(s)] \\
&\equiv A(s)\varepsilon'(s)[\varphi_0(s), \cdots, \varphi_{n-1}(s)] \\
&\equiv A(s)\varepsilon(s)[\varphi_0(s), \cdots, \varphi_{n-1}(s)] \pmod{\psi}.
\end{aligned}$$

Since $\varphi_{n-1}(s) = 1$, equality (mod ψ) of the last columns of the matrices on the left and the right gives the first of equations (20). The second is derived similarly.

In addition to the computational saving resulting from the option of using an arbitrary diagonal form of the transfer function matrix, our procedure has a theoretical advantage over Kalman's in that it more clearly exhibits the role of the polynomial invariants which appear as numerators in the Smith–McMillan canonical form. Indeed, if we take D' in Theorem 4 to be the reduced Smith–McMillan form (7) of Z, then the entries in the output matrices H_i of (17) are just the coefficients of the polynomials obtained as remainders after dividing the

numerators in the Smith–McMillan form (6) of Z by the corresponding denominators.

Our procedure has the further advantage that it exhibits directly the relationship between the system realizing a given transfer function matrix and the completely uncoupled system realizing the associated (reduced) Smith–McMillan form.

5. Further remarks. (a) As was pointed out in § 2, a complete set of invariants for strong equivalence of rational matrices is provided by the rank R together with the $2R$ polynomials, $\psi_1, \cdots, \psi_R$ and $\varepsilon_1, \cdots, \varepsilon_R$, which occur in the Smith–McMillan canonical form. There still remains the problem of finding a complete set of invariants for weak equivalence. A step toward the solution of this problem is provided by the following theorem which suggests a possible candidate for a canonical form for weak equivalence.

THEOREM 5. *Let Z be a rational matrix. Then Z is weakly equivalent to a proper diagonal matrix Ω of the form*

$$\Omega = \operatorname{diag}[1/\psi_1, \cdots, 1/\psi_{n-1}, \omega/\psi_r, 0, \cdots, 0],$$

where $\psi_1, \cdots, \psi_r$ are the polynomials which appear as denominators in the Smith–McMillan form of Z and where ω is a (nonzero) polynomial which is relatively prime to ψ_r.

Proof. We shall show that, for each i with $1 \leqq i \leqq r$, Z is weakly equivalent to a proper diagonal matrix of the form

$$\Omega_i = \operatorname{diag}[1/\psi_1, \cdots, 1/\psi_{i-1}, \omega_{ii}/\psi_i, \cdots, \omega_{ir}/\psi_r, 0, \cdots, 0],$$

where each ω_{ij}, $j = i, \cdots, r$, is a (nonzero) polynomial which is relatively prime to the corresponding ψ_j. Then, taking $i = r$ and setting $\omega = \omega_{rr}$ will complete the proof.

Note that each rational matrix is weakly equivalent to a matrix of the form Ω_1 since each matrix is weakly equivalent to its reduced Smith–McMillan form (7). We now show that each matrix of the form Ω_i for $1 \leqq i < r$ is weakly equivalent to one of the form Ω_{i+1}; transitivity of weak equivalence will then imply the theorem.

We shall alter Ω_i into the form Ω_{i+1} by means of weak equivalence operations, working only with the ith and $(i+1)$st rows and columns. By denoting ω_{ii} by θ and $\omega_{i,i+1}$ by η, the effect of these operations on the 2×2 submatrix obtained from the ith and $(i+1)$st rows and columns will be as follows (with $\psi = \psi_1$):

$$\psi\begin{bmatrix} \theta/\psi_i & 0 \\ 0 & \eta/\psi_{i+1} \end{bmatrix} \to \begin{bmatrix} \theta\psi/\psi_i & \alpha\theta\psi/\psi_i \\ 0 & \eta\psi/\psi_{i+1} \end{bmatrix} \equiv \begin{bmatrix} \theta\psi/\psi_i & \psi/\psi_i \\ 0 & \eta\psi/\psi_{i+1} \end{bmatrix}$$

$$\to \begin{bmatrix} \theta\psi/\psi_i & \psi/\psi_i \\ -\eta\theta\psi/\psi_{i+1} & 0 \end{bmatrix} \to \begin{bmatrix} 0 & \psi/\psi_i \\ -\eta\theta\psi/\psi_{i+1} & 0 \end{bmatrix} \to \psi\begin{bmatrix} 1/\psi_i & 0 \\ 0 & -\eta\theta/\psi_{i+1} \end{bmatrix}.$$

Each of these operations consists of the addition of some polynomial multiple of one row or column to another, a congruence modulo ψ, or an interchange of two columns. The polynomial α is chosen so that $\alpha\theta + \beta\psi_i = 1$ for some β; it exists because θ and ψ_i are relatively prime. The congruence mod ψ is valid since

$\alpha\theta\psi/\psi_i + \beta\psi = \psi/\psi_i$ and hence $\alpha\theta\psi/\psi_i \equiv \psi/\psi_i \pmod{\psi}$. Note that $-\eta\theta$ is relatively prime to ψ_{i+1} since both η and θ are (θ is prime to ψ_1 and ψ_{i+1} divides ψ_i).

Clearly, the proper part of the matrix obtained after this sequence of operations is of the form Ω_{i+1}, it is weakly equivalent to Ω_i, and the proof is complete.

It is clear that the number r together with the r polynomials $\psi_1, \cdots, \psi_r$ are invariants of weak equivalence, since any pair of weakly equivalent proper rational matrices has externally equivalent minimal realizations and the polynomials ψ_i are the invariant polynomials of the common state matrix F. Hence, if it could be shown that ω is also an invariant of weak equivalence, then Ω would be a canonical form for weak equivalence (i.e., there would be one and only one matrix of the form Ω in each equivalence class) and $\{r, \psi_1, \cdots, \psi_r, \omega\}$ would be a complete set of invariants for weak equivalence.

There is a polynomial ζ closely related to ω which *is* an invariant of weak equivalence. Given a rational matrix A, this invariant is obtained as the mod ψ reduction (the remainder after division by ψ) of the greatest common divisor of the $r \times r$ minors of ψZ, where ψ and r are as in the reduced Smith–McMillan form (7) of Z. For the matrix Ω of Theorem 5, this invariant ζ is just the mod ψ reduction of $\psi/\psi_1, \cdots, \psi\omega/\psi_r$. The polynomial ω cannot be determined from a knowledge of $\psi_1, \cdots, \psi_r$, and ζ; this may be seen by considering the example $\Omega = \operatorname{diag}[1/s^4, 1/s^2, (as + b)/s^2]$.

It may be of interest to note that the polynomial ω obtained in the proof of Theorem 5 is just the mod ψ_r reduction of the polynomial $(-1)^r\varepsilon_1, \cdots, \varepsilon_r$, where $\varepsilon_1, \cdots, \varepsilon_r$, and ψ_r are as in the Smith–McMillan canonical form of Z. This may be seen upon close examination of the proof.

(b) The realization procedure described in § 4 suggests a way of "realizing" any (not necessarily proper) rational matrix as follows. Given a rational matrix Z, let ψ denote the least common denominator of the entries in Z and let $\hat{Z}$ denote the (proper) rational matrix with the property that

$$\psi\hat{Z} \equiv \psi Z \pmod{\psi};$$

that is, $\psi\hat{Z}$ is the proper part of ψZ. Then we can define a *realization of the rational matrix* Z to be any realization of $\hat{Z}$.

In terms of this definition of realization of improper rational matrices, our realization technique (§ 4) can be described as follows: first realize any diagonalization of Z in the obvious way (parts (a) and (b) of Theorem 4); then construct the realization of Z itself using formula (12).

(c) Another way of "realizing" an improper rational matrix is by enlarging the class of linear dynamical systems. As is well known, a regular rational matrix (one in which the degree of each numerator is less than or equal to the degree of the corresponding denominator) can be realized by a system (F, G, H, K), where F, G, H and K are matrices corresponding to equations of the form

$$\dot{x} = Fx + Gu,$$

$$y = Hx + Ku.$$

Similarly, any improper rational matrix Z can be realized by a system (F, G, H, K), where F, G and H are matrices and K is a matrix polynomial corresponding to

equations of the form

$$\begin{aligned} \dot{x} &= Fx + Gu, \\ y &= Hx + K(D)u, \end{aligned} \tag{21}$$

where D is the differentiation operator. The system (F, G, H) corresponds to the "proper part" (the mod ψ reduction $\hat{Z}$) of Z and the polynomial K corresponds to the "improper part" (the polynomial matrix $Z - \hat{Z}$) of Z. Physically, the presence of an improper part alters the system by introducing the input and its derivatives directly into the output (bypassing the state space).

(d) We consider now the question of when is a Smith–McMillan form (6) equal to the Smith–McMillan form of some linear dynamical system (F, G, H). More generally we ask: when is an improper rational matrix *strongly* equivalent to a proper one? In view of remark (b) above, this is equivalent to asking: when is a generalized system of the type (21) strongly transfer equivalent to an ordinary system of the type (1)? The answer is provided by a degree condition.

THEOREM 6. *Let Z be a rational matrix. Then Z is strongly equivalent to a proper rational matrix if and only if the degree of $\Delta_R(\psi Z)$ is less than or equal to $R(n-1)$, where R is the rank of Z, ψ is the least common denominator of the entries in Z, $\Delta_R(\psi Z)$ is the greatest common divisor of the $R \times R$ minors of ψZ, and n is the degree of ψ.*

Proof. First recall that ψ, R and $\Delta_R(\psi Z)$ are invariant under strong equivalence. Hence the necessity of the condition is clear, since for a proper rational Z the maximal degree of the entries in ψZ is $n - 1$ and hence the maximal degree of any $R \times R$ minor is $R(n-1)$.

For sufficiency we may assume that Z is in Smith–McMillan form, since every rational Z is strongly equivalent to such a form. Then ψZ is of the form (3) where the degrees $d(\gamma_i)$ of the γ_i satisfy

$$\sum_{i=1}^{R} d(\gamma_i) = d[\Delta_R(\psi Z)] \leqq R(n-1). \tag{22}$$

Let p denote the largest integer in $\{1, \cdots, R\}$ such that γ_p is proper (i.e., such that $d(\gamma_p) \leqq n - 1$). (Note that γ_1 is proper since, by the divisibility relations among the γ_i, $d(\gamma_i) \geqq d(\gamma_1)$ for all i and hence $d(\gamma_1) \leqq (\sum d(\gamma_i))/R \leqq n - 1$.) We shall show that for each $k = p, \cdots, R$, ψZ is strongly equivalent to a matrix M_k of the form

$$M_k = \begin{bmatrix} \gamma_i & & & & & & \\ & \ddots & & & & & \\ & & \gamma_{j-1} & & & & \\ & & & \boxed{\alpha} & & & \\ & & & & \gamma_{k+1} & & \\ & & & & & \ddots & \\ & & & & & & \gamma_R \\ & & & & & & & 0 \\ & & & & & & & & \ddots \end{bmatrix}, \tag{23}$$

where each element in the box is proper, j is some positive integer less than p (depending on k), and α is a (nonzero) polynomial which is of maximal degree among all the elements in the kth row and all the elements in the kth column and which satisfies

$$d(\alpha) = n - 1 + \sum_{i=j}^{k} [d(\gamma_i) - (n-1)]. \tag{24}$$

Taking $k = R$ will then complete the proof.

The proof now proceeds by induction on k. Taking $k = p$ we see that $M_p = \psi Z$ satisfies (23) and (24) with $j = p$ and $\alpha = \gamma_p$. So now we assume we have found M_k $(p \leqq k < R)$ and we proceed to construct M_{k+1}.

Let q_1 be a polynomial of degree $n - 1 - d(\alpha)$ ($q_1(s) = s^{n-1-d(\alpha)}$ will do). Dividing γ_{k+1} by $q_1\alpha$ we obtain

$$\gamma_{k+1} = p_1 q_1 \alpha + r_1 \tag{25}$$

for some polynomials p_1 and r_1 with $d(r_1) < d(q_1\alpha) = d(q_1) + d(\alpha) = n - 1$. Multiplying the kth row of M_k by p_1 and adding it to the $(k+1)$st row, then multiplying the jth column of this new matrix by q_1 and subtracting it from the $(k+1)$st column, we obtain a matrix $M_k^{(1)}$ which is strongly equivalent to M_k (and hence to ψZ) and which is of the form

$$M_k^{(1)} = \begin{bmatrix} \ddots & & & \\ & \gamma_{j-1} & & \\ & & \boxed{\alpha \quad \square} & \\ & & \boxed{p_1\alpha \cdots \quad \square} & \\ & & & \gamma_{k+2} \\ & & & & \ddots \end{bmatrix}, \tag{26}$$

where all entries in the two boxes are proper and where $p_1\alpha$ is an element of maximal degree in its row and column. Moreover, by (24) and (25),

$$\begin{aligned} d(p_1\alpha) &= d(\gamma_{k+1}) - d(q_1) = d(\gamma_{k+1}) - (n-1) + d(\alpha) \\ &= n - 1 + \sum_{i=j}^{k+1} [d(\gamma_i) - (n-1)]. \end{aligned} \tag{27}$$

If $p_1\alpha$ is proper, we may take $M_{k+1} = M_k^{(1)}$ and we are done. Furthermore, if $j = 1$ in (26) (i.e., if the square box extends all the way to the upper left-hand corner of $M_k^{(1)}$), then $p_1\alpha$ is necessarily proper since then, using (22),

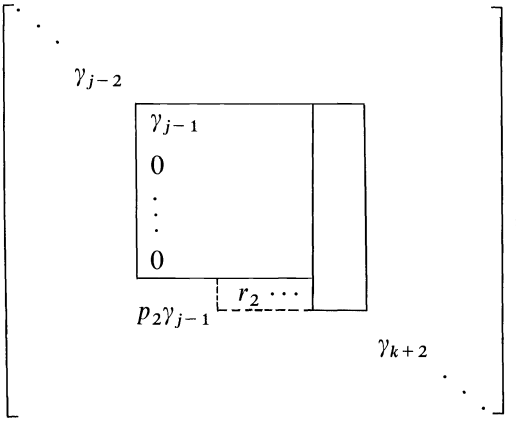
$$\begin{aligned} d(p_1\alpha) &= n - 1 + \sum_{i=1}^{R} d(\gamma_i) - (k+1)(n-1) - \sum_{i=k+2}^{R} d(\gamma_i) \\ &\leqq n - 1 + R(n-1) - (k+1)(n-1) - [R - (k+1)](n-1) \\ &= n - 1. \end{aligned}$$

If $p_1\alpha$ is improper (i.e., $d(p_1\alpha) > n - 1$), then we can use γ_{j-1} to reduce its degree. Let q_2 be a polynomial of degree $n - 1 - d(\gamma_{j-1})$. Then

$$p_1\alpha = p_2q_2\gamma_{j-1} + r_2, \tag{28}$$

where $d(r_2) < d(q_2\gamma_{j-1}) = n - 1$. Multiplying the $(j-1)$st row of $M_k^{(1)}$ by p_2 and adding it to the $(k+1)$st row, then multiplying the $(j-1)$st column of this new matrix by q_2 and subtracting it from the jth column, we obtain a matrix $M_k^{(2)}$ which is strongly equivalent to $M_k^{(1)}$ (and hence to ψZ) and which is of the form

(29) $M_k^{(2)} =$



where all the entries in the solid boxes are proper and where $p_2\gamma_{j-1}$ is an element of maximal degree in its column. Furthermore, we may assume that $p_2\gamma_{j-1}$ is an element of maximal degree in its row; otherwise we can accomplish this by multiplying the $(j-1)$st column of (29) by appropriate polynomials q_i and subtracting from the other columns. (This process will not introduce improper entries into the solid boxes in (29) since: (i) $p_1\alpha$ was of maximal degree in its row in (26), and (ii) $d(p_1\alpha) > n - 1 > d(r_2)$; hence for each element β in the dotted box in (29) we have, using (28),

$$d(\beta) \leqq d(p_1\alpha) = d(p_2q_2\gamma_{j-1}) = d(q_2) + d(p_2\gamma_{j-1}),$$

so the polynomial q_i required to reduce β will have $d(q_i) \leqq d(q_2)$ and therefore $d(q_i\gamma_{j-1}) \leqq d(q_2\gamma_{j-1}) = n - 1$.) Also we have, by (27) and (28),

$$\begin{aligned} d(p_2\gamma_{j-1}) &= d(p_1\alpha) - d(q_2) = d(p_1\alpha) - (n-1) + d(\gamma_{j-1}) \\ &= n - 1 + \sum_{i=j-1}^{k+1} [d(\gamma_i) - (n-1)]. \end{aligned}$$

Hence, if $p_2\gamma_{j-1}$ is proper, we may take $M_{k+1} = M_k^{(2)}$ and we are done. If $p_2\gamma_{j-1}$ is improper, we are again in the same situation as (26) but with a box one row and one column larger and we may repeat the above process to obtain $M_k^{(3)}$. Clearly this process must stop and we obtain $M_{k+1} = M_k^{(l)}$ for some $l \leqq j$.

(e) As mentioned in § 4, our realization procedure provides new insight into the significance of the polynomial invariants ε_k of a system which occur in the Smith–McMillan form of the transfer function matrix Z. Another interpretation of these polynomials is obtained as follows.

Instead of the polynomial ε_k which occurs in the Smith–McMillan form of Z, we consider the polynomial invariant $\Delta_k(\psi Z)$ obtained as the greatest common divisor of all $k \times k$ minors of ψZ, $k = 1, \cdots, R$. As is well known, these polynomials $\Delta_k(\psi Z)$ are related to the polynomials γ_k appearing in the canonical form (3) of ψZ by

$$\Delta_k(\psi Z) = \gamma_1 \cdots \gamma_k. \tag{30}$$

In particular, the polynomials $\Delta_k(\psi Z)$ contain the same information as the polynomials ε_k.

The polynomial $\Delta_1(\psi Z) = \gamma_1$ has an immediate interpretation, as a consequence of its definition as the greatest common divisor of the entries in ψZ: it is the polynomial whose roots form the set of zeros of the given system (i.e., the set of frequencies s to which the system is completely unresponsive). A similar interpretation of the $\Delta_k(\psi Z)$ $(k > 1)$ can be given by constructing a system whose transfer function matrix has as entries the $k \times k$ minors of ψZ. We shall carry out this construction here only for $k = 2$. The generalization to arbitrary k is straightforward.

We recall first a few facts from multilinear algebra [14]. Recall that to each vector V is attached another space $\Lambda^2(V)$, the spaces of bivectors of V. Formally, $\Lambda^2(V)$ is the vector space generated by objects of the form $u \wedge v$ $(u, v \in V)$ and subject to the relations

$$(u_1 + u_2) \wedge v = u_1 \wedge v + u_2 \wedge v,$$

$$(cu) \wedge v = c(u \wedge v),$$

$$u \wedge v = -v \wedge u,$$

where $u, v \in V$ and c is a scalar. Given a basis $\{v_1, \cdots, v_n\}$ for V, the set $\{v_i \wedge v_j | i < j\}$ is a basis for $\Lambda^2(V)$. In particular, $\Lambda^2(V)$ has dimension $n(n-1)/2$, where n is the dimension of V.

Given a pair A, B of linear operators on V, there is induced a linear operator $A \wedge B$ on $\Lambda^2(V)$, defined by

$$A \wedge B(u \wedge v) = \tfrac{1}{2}[A(u) \wedge B(v) - A(v) \wedge B(u)].$$

The following properties of this "wedge product" of operators are easily checked:

$$(A + B) \wedge C = A \wedge C + B \wedge C,$$

$$(cA) \wedge B = c(A \wedge B),$$

$$A \wedge B = B \wedge A,$$

$$(A \wedge B)(C \wedge D) = (AC) \wedge (BD).$$

It is also easily verified that, given a basis $\{v_1, \cdots, v_n\}$ for V, the matrix for the operator $A \wedge A$ relative to the basis $\{v_i \wedge v_j | i < j\}$ for $\Lambda^2(V)$ is a matrix whose entries are the 2×2 minors of the matrix for A relative to $\{v_1, \cdots, v_n\}$.

With slight modifications, the above discussion is also valid in the more general situation where V is a module over a commutative ring.

Now, given a constant linear dynamical system (F, G, H) with transfer function matrix Z, we can exhibit a related system whose transfer function matrix

is $Z \wedge Z$. Since

$$\begin{aligned} Z \wedge Z &= [H(Is - F)^{-1}G] \wedge [H(Is - F)^{-1}G] \\ &= H \wedge H[(Is - F)^{-1} \wedge (Is - F)^{-1}]G \wedge G \\ &= H \wedge H[(Is - F) \wedge (Is - F)]^{-1}G \wedge G \\ &= H \wedge H[I \wedge Is^2 - 2F \wedge Is + F \wedge F]^{-1}G \wedge G, \end{aligned}$$

it is clear that $Z \wedge Z$ is the transfer function matrix of the second order system

$$\ddot{x} - 2F \wedge I\dot{x} + F \wedge Fx = G \wedge Gu,$$

$$y = H \wedge Hx.$$

Transforming this system in the standard way to a first order system, we conclude that $Z \wedge Z$ is the transfer function matrix of the linear dynamical system $(\tilde{F}, \tilde{G}, \tilde{H})$ given by

$$\tilde{F} = \begin{bmatrix} 0 & I \wedge I \\ -F \wedge F & 2F \wedge I \end{bmatrix}, \quad \tilde{G} = \begin{bmatrix} 0 \\ G \wedge G \end{bmatrix}, \quad \tilde{H} = [H \wedge H, \quad 0].$$

Furthermore, the invariant $\Delta_2(\psi Z)$ of the original system (F, G, H) is the polynomial whose roots are the zeros of the associated system $(\tilde{F}, \tilde{G}, \tilde{H})$.

The blocks appearing in $\tilde{F}$ each admit simple interpretations: $I \wedge I$ is the identity, $F \wedge F$ represents the induced operator on $\Lambda^2(V)$ (V the state space of (F, G, H)) given by

$$F \wedge F(u \wedge v) = F(u) \wedge F(v)$$

and $2F \wedge I$ represents the extension of F to a derivation on $\Lambda^2(V)$, since

$$2F \wedge I(u \wedge v) = F(u) \wedge v + u \wedge F(v).$$

REFERENCES

[1] E. G. GILBERT, *Controllability and observability in multivariable control systems*, this Journal, 1 (1963), pp. 128–151.

[2] R. E. KALMAN, *On the general theory of control systems*, Proc. 1st International Congress on Automatic Control (Moscow, 1960), vol. 1, Butterworths, London, 1961, pp. 481–492.

[3] ———, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

[4] ———, *Irreducible realizations and the degree of a rational matrix*, J. Soc. Indust. Appl. Math., 13 (1965), pp. 520–544.

[5] ———, *On structural properties of linear, constant multivariable systems*, Proc. 3rd IFAC Congress, London, 1966, pp. 6A.1–6A.9.

[6] ———, *Algebraic aspects of the theory of dynamical systems*, Differential Equations and Dynamical Systems, J. K. Hale and J. P. LaSalle, eds., Academic Press, New York, 1967.

[7] R. E. KALMAN, Y. C. HO AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1963), pp. 189–213.

[8] B. L. HO AND R. E. KALMAN, *Effective construction of state-variable models from input/output relations*, Regelungstechnik, 14 (1966), pp. 545–548.

[9] H. H. ROSENBROCK, *Transformation of linear constant system equations*, Proc. Inst. Elec. Engrs., 114 (1967), pp. 541–544.

[10] ———, *On linear system theory*, Ibid., 114 (1967), pp. 1353–1359.

[11] L. WEISS AND R. E. KALMAN, *Contributions to linear system theory*, Internat. J. Engrg. Sci., 3 (1965), pp. 141–171.

[12] B. McMillan, *Introduction to formal realizability theory*, Bell System Tech. J., 31 (1952), pp. 541–600.
[13] R. F. Gantmacher, *The Theory of Matrices*, vol. 1, Chelsea, New York, 1959.
[14] N. Bourbaki, *Algèbre multilinéaire*, Elements de mathématique, Livre II (Algèbre), Hermann, Paris, 1958, Chap. 3.

SIAM J. CONTROL
Vol. 8, No. 1, February 1970

THE VALIDITY OF A FAMILY OF OPTIMIZATION METHODS*

ROBERT MEYER†

Abstract. A family of iterative optimization methods, which includes most of the well-known algorithms of mathematical programming, is described and analyzed with respect to the properties of its accumulation points. It is shown that these accumulation points have desirable properties under appropriate assumptions on a relevant point-to-set mapping. The conditions under which these assumptions hold are then discussed for a number of algorithms, including steepest descent, the Frank–Wolfe method, feasible direction methods, and some second order methods. Five algorithms for a special class of nonconvex problems are also analyzed in the same manner. Finally, it is shown that the results can be extended to the case in which the subproblems constructed are only approximately solved and to algorithms which are composites of two or more algorithms.

1. Semicontinuity and mathematical programming. The concepts of upper and lower semicontinuity for point-to-set mappings have been studied by a number of prominent mathematicians, including Hausdorff [1], Berge [2] and Dantzig [3]. Several similar definitions of the two concepts have been formulated, and some comparisons may be found in a recent paper by Jacobs [4]. The following definitions, which are essentially the same as those given in Debreu [5], will be used in this paper: a point-to-set mapping Ω with domain G and range consisting of subsets of a set R is said to be (i) *upper semicontinuous* (u.s.c.) *at a point* y belonging to G if $y_i \to y$, $\{y_i\} \subset G$, and $z_i \to z$ with $z_i \in \Omega(y_i)$ for each i imply $z \in \Omega(y)$; (ii) *lower semicontinuous* (l.s.c.) *at a point* y belonging to G if $z \in \Omega(y)$, $y_i \to y$, $\{y_i\} \subset G$ imply the existence of an integer m and a sequence $\{z_m, z_{m+1}, \cdots\}$ with the properties that (a) $z_i \in \Omega(y_i)$ for $i \geqq m$ and (b) $z_i \to z$; and (iii) *continuous at a point* y if it is both upper and lower semicontinuous at y. Note that these definitions are meaningful whenever the notion of convergence is defined in both G and R. In particular, they are valid if G and R are subsets of topological or metric spaces. (If $\Omega(y)$ is a single point for every $y \in G$, i.e., a function, then it is easily seen that l.s.c. at a point implies u.s.c. and hence continuity at that point. Similarly, if Ω is single-valued and R is a sequentially compact subset of a topological space, then it is true that u.s.c. at a point implies l.s.c. and hence continuity at that point. However, it is easy to construct set-valued mappings that are only u.s.c. or only l.s.c. even when R is a compact subset of E^n. Examples displaying this behavior appear below. These notations should not be confused with *numerical upper and lower semicontinuity* for real-valued functions, which have quite different definitions.)

An important class of point-to-set mappings consists of those mappings that involve the linearization of some or all of the constraints defining a set about a point. Let M denote the set $S \cap \{z|u(z) \geqq 0\} \cap \{z|v(z) = 0\}$, where S is a closed subset of a Banach space B and u and v are continuously Fréchet differentiable vector-valued functions. For a point $y \in B$, we say that the "linearization" of M about y is the set

$$\Gamma m(y) \equiv S \cap \{z|u(y) + u'(y)(z - y) \geqq 0, v(y) + v'(y)(z - y) = 0\}.$$

* Received by the editors August 16, 1968, and in revised form June 17, 1969.

† Computer Sciences Department, University of Wisconsin, Madison, Wisconsin. Now at Shell Development Company, Emeryville, California 94608. This research was sponsored in part by the National Science Foundation under Grant NSF-GP-6070.

Note that if $y \in M$, then $\Gamma m(y)$ is nonempty, since $y \in \Gamma m(y)$. We shall now show that the point-to-set mapping Γm is u.s.c. at every point of B. For, let $y \in B$, and let $y_i \to y$. If $z_i \in \Gamma m(y_i)$ for each i and $z_i \to z$, then it follows from the closure of S that $z \in S$, and it follows from the continuity of u' and v' at y that

$$u(y) + u'(y)(z - y) = \lim (u(y_i) + u'(y_i)(z_i - y_i)) \geqq 0$$

and

$$v(y) + v'(y)(z - y) = \lim (v(y_i) + v'(y_i)(z_i - y_i)) = 0,$$

so that $z \in \Gamma m(y)$, proving u.s.c. Without additional hypotheses, however, it is not true that Γm is l.s.c. This fact was demonstrated by Rosen [6], and it can also be deduced from the following very simple example where B is taken to be the real line.

Example. Take $S = E^1$, $v \equiv 0$, and $u(z) = z^3$. Let $y_i = 1/i$, so that $y_i \to 0$ as $i \to \infty$ and $\Gamma m(y_i) = \{z | y_i^3 + 3y_i^2(z - y_i) \geqq 0\} = \{z | z \geqq 2/3i\}$. However, $\Gamma m(0) = E^1$, and it is clear that Γm is u.s.c. but not l.s.c. at the point 0.

In the case that $B = E^n$ and S is a convex set consisting of the points satisfying $f(z) \geqq 0$, where f is continuous and vector-valued, the next theorem gives sufficient conditions for l.s.c. of Γm in the neighborhood of a point. We adopt the convention of calling the inequality constraint $f_i(z) \geqq 0$ $(u_i(y) + u_i'(y)(z - y) \geqq 0)$ active at the point $\bar{z} \in \Gamma m(y)$ if $f_i(\bar{z}) = 0$ $(u_i(y) + u_i'(y)(\bar{z} - y) = 0)$. The (possibly vector-valued) function consisting of active constraint functions at $\bar{z} \in \Gamma m(y)$ is understood to consist of those functions $f_i(z)$ and $u_i(y) + u_i'(y)(z - y)$ which correspond to active inequality constraints at $\bar{z}$ as well as the function $v(y) + v'(y)(z - y)$.

THEOREM 1.1. *Under the preceding assumptions on B and S, the point-to-set mapping Γm is l.s.c. in a neighborhood of a point y^* if the set $\Gamma m(y^*)$ contains a point z^* at which the gradients to the active constraint functions at z^* are linearly independent.*

Proof. See Appendix.

We shall now obtain three basic results relating semicontinuity of point-to-set mappings to mathematical programming. Similar results may be found in Berge [2] and Debreu [5]. It will be assumed that f is a real-valued function defined and continuous on $R \times G$ and that the optimal value function $\mu(y) = \min_{z \in \Omega(y)} f(z, y)$ is well-defined for every $y \in G$.

LEMMA 1.2. *If Ω is u.s.c. at a point $y^* \in G$ and R is sequentially compact, then μ is (numerically) lower semicontinuous at y^*.*

Proof. Let $y_i \to y^*$, $\{y_i\} \subset G$. Then there exist sequences $\{y_{n_i}\}$ and $\{z_{n_i}\}$ such that $\mu(y_{n_i}) = f(z_{n_i}, y_{n_i})$, $z_{n_i} \to z^*$, and $\mu(y_{n_i}) \to \liminf \mu(y_i)$ as $i \to \infty$. It follows from u.s.c. that $z^* \in \Omega(y^*)$, and thus $\liminf \mu(y_i) = \lim \mu(y_{n_i}) = f(z^*, y^*) \geqq \mu(y^*)$.

If the compactness hypothesis is deleted, the conclusion is no longer valid. Examples illustrating this are easily constructed. However, compactness is not required in the following complementary result.

LEMMA 1.3. *If Ω is l.s.c. at a point $y^* \in G$, then μ is (numerically) upper semicontinuous at y^*.*

Proof. Let $z^* \in \Omega(y^*)$ be such that $\mu(y^*) = f(z^*, y^*)$, and let $\{y_i\}$ be an arbitrary sequence in G converging to y. Choose $\{y_{n_i}\}$ and $\{z_{n_i}\}$ such that $\mu(y_{n_i})$

$\to \limsup \mu(y_i)$ and $z_{n_i} \to z^*$, with $z_{n_i} \in \Omega(y_{n_i})$. We then have $\mu(y^*) = f(z^*, y^*) = \lim f(z_{n_i}, y_{n_i}) \geqq \lim \mu(y_{n_i}) = \limsup \mu(y_i)$.

Combining the previous two lemmas, we obtain the following theorem.

THEOREM 1.4. *If Ω is continuous at a point $y^* \in G$ and R is sequentially compact, then μ is (numerically) continuous at y^*.*

The next theorem reflects a slightly different viewpoint. It shows that continuity of Ω is a sufficient condition for the limit of a set of solutions to solve the limiting problem. Note that compactness does not enter directly into the statement of the result.

THEOREM 1.5. *Let $M(y)$ denote the subset of $\Omega(y)$ consisting of all points z such that $f(z, y) = \mu(y)$. If Ω is continuous at $y^* \in G$, then the point-to-set mapping M is u.s.c. at y^*.*

Proof. Let $\{y_i\} \subset G$ converge to y^*, and let $z_i \in M(y_i)$ for each i, with $z_i \to z^*$. Since Ω is u.s.c. at y^*, it follows that $z^* \in \Omega(y^*)$, and thus $\mu(y^*) \leqq f(z^*, y^*)$. On the other hand, by Lemma 1.3, $f(z^*, y^*) = \lim \mu(y_i) \leqq \mu(y^*)$.

(The previous theorem and also Theorem 1.1 are similar to results published recently by Dantzig, Folkman and Shapiro [3]. Theorem 1.1 differs from the corresponding result [3, Corollary II.3.5] in that the former: (i) allows for the intersection of the linearized constraints with the convex set S, and also (ii) has a stronger conclusion. As with Theorem 1.5, the method of proof is quite different than that used by Dantzig et al. in obtaining similar results. Finally, as noted previously, Theorem 1.5 is proved assuming only that the notion of convergence is defined in the spaces dealt with, whereas the analogous Corollary I.2.3 is stated for a pair of metric spaces.)

Let us now suppose that $R \subset G$ and that we have a continuous function φ defined on G with the property that $y' \in M(y)$ implies $\varphi(y') < \varphi(y)$ unless $y \in M(y)$. (This will sometimes be referred to as the *strict monotonic property.*) Consider the algorithm defined as follows:

(a) Choose an arbitrary $y_0 \in G$.

(b) Let $y_{i+1} = y_i$ if $y_i \in M(y_i)$; otherwise let $y_{i+1} \in M(y_i)$.

THEOREM 1.6. *If $\{y_i\}$ is contained in a sequentially compact set and y^* is an accumulation point of $\{y_i\}$ at which Ω is continuous, then $y^* \in M(y^*)$.*

Proof. If the conclusion were false, we would have $\varphi(\bar{y}) < \varphi(y^*)$ for all $\bar{y} \in M(y^*)$ by the assumption on φ. We shall show that this leads to a contradiction. Let subsequences $\{y_{n_i}\}$ and $\{y_{n_i+1}\}$ be chosen so that $y_{n_i} \to y^*$ and $y_{n_i+1} \to y'$. It follows from the previous theorem that $y' \in M(y^*)$, so that $\varphi(y') < \varphi(y^*)$. However, since $\{\varphi(y_i)\}$ is a monotone decreasing sequence we have $\varphi(y') = \lim \varphi(y_{n_i+1}) = \lim \varphi(y_i) = \lim \varphi(y_{n_i}) = \varphi(y^*)$, a contradiction.

The preceding theorem can be looked upon as a special case of a result of Zangwill [7], who assumes u.s.c. of M instead of the continuity of Ω. From Theorem 1.5 it follows that Zangwill's result is more general than Theorem 1.6; but, from the standpoint of application, the latter appears to be a more useful formulation in many cases. In addition by considering the continuity properties of Ω rather than M, a sharper result can be obtained in the important case $f(z, y) \equiv \varphi(z)$. This result, given in the following theorem, states that in such cases the conclusion of Theorem 1.6 continues to hold when the assumptions of sequential compactness and u.s.c. of Ω are dropped. (As shown by an example in the Appendix, sequential

compactness cannot be deleted in the more general case. Simple examples can also be constructed to show the need for u.s.c. of Ω in both Theorems 1.5 and 1.6.)

THEOREM 1.7. *If $\varphi(z) = f(z, y)$ and y^* is an accumulation point of $\{y_i\}$ at which Ω is l.s.c., then $y^* \in M(y^*)$.*

Proof. Again suppose that the conclusion is false, and let $\bar{y} \in M(y^*)$, so that $\varphi(\bar{y}) < \varphi(y^*)$. Since Ω is assumed l.s.c. at y^*, there exists a sequence of points $\{z_{n_i}\}$ with $z_{n_i} \in \Omega(y_{n_i})$ for each n_i and such that $z_{n_i} \to \bar{y}$. Thus we conclude that $\varphi(y^*) > \varphi(\bar{y}) = \lim \varphi(z_{n_i}) \geqq \lim \varphi(y_{n_i+1}) = \varphi(y^*)$, which cannot hold.

2. An application: Reverse–convex programming. Consider the following problem:

I. Minimize

$$\varphi(z)$$

subject to

$$z \in F \equiv S \cap \{z | u(z) \geqq 0\},$$

where S is a closed, convex subset of E^n, u is a vector-valued, convex and continuously differentiable function, and φ is continuous and real-valued on F. We shall further assume that F is bounded, which is easily seen to imply that F is compact. The curious feature of problems of the form I is the nonconvexity of the feasible region F. The convexity of u implies that the region $\{z | u(z) \leqq 0\}$ given by the *reverse* inequalities is convex (see Fig. 1). For this reason, the sets $U \equiv \{z | u(z) \geqq 0\}$ and F will be called *reverse–convex* and the problem I a *reverse–convex minimization problem.*

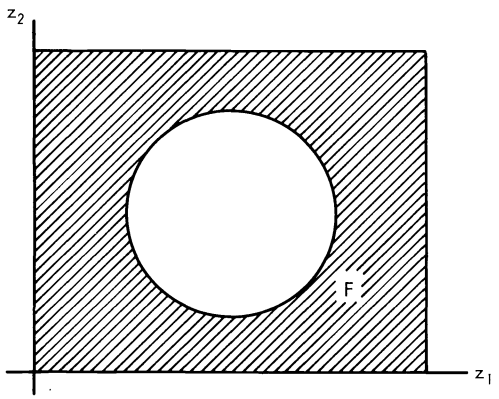


FIG. 1. *A reverse–convex set*

Such problems arise, for example, when we wish to determine the minimum of a function in a region from which an open sphere about a point has been removed.

It is easy to show that even with a linear objective function such a problem may have a local minimum that is not a global minimum. The numerical methods

proposed below, like all numerical methods based on local searches for solutions, can at best be expected to yield local minima for problems of the form I. Global minimality could be assured only by exhaustive searches over successively finer grids. Grid search techniques, however, usually require more function evaluations than would be computationally feasible for most practical problems.

In order to establish iterative procedures for problems of type I, we first consider a technique for generating convex subsets of F. This is conveniently done by "linearization." That is, we define $W(y) \equiv \{z | u(y) + u'(y)(z - y) \geqq 0\}$, the "linearization" of the set U about the point y (here $u'(y)$ is the Jacobian of u evaluated at y); and we let $\Gamma(y) \equiv S \cap W(y)$. For every $y \in F$ the set $\Gamma(y)$ has the three important properties: (i) $\Gamma(y)$ is convex and compact, (ii) $\Gamma(y) \subset F$, and (iii) $y \in \Gamma(y)$. (This was first pointed out by Rosen [6].) Property (ii) is an immediate consequence of the convexity of u, and properties (i) and (iii) are obvious. Note that by the results of § 1, the point-to-set mapping Γ is everywhere u.s.c.

Consider now the following subproblem, $R(y)$, derived from the problem I:

$R(y)$. Minimize

$$\varphi(z)$$

subject to

$$z \in \Gamma(y).$$

If $y \in F$, by property (i) above, $R(y)$ has a solution; by property (ii), every point at which the minimum value is attained must be in F; and by property (iii), if z^* solves $R(y)$, $\varphi(z^*) \leqq \varphi(y)$. Of course, it is not likely that a solution of the subproblem $R(y)$ can be obtained by numerical means unless the objective function has some convexity property. (For example, if φ is strictly quasi-convex [8], a local minimum for $R(y)$ will be a global minimum.) In many problems of interest the objective function will be linear, so that if S is a polytope (the intersection of a finite number of half-spaces), the problem $R(y)$ can be solved by linear programming (LP). In any event, the following iterative scheme proposed by Rosen [6] is mathematically well-defined:

METHOD A.

(a) Choose an arbitrary $y_0 \in F$.

(b) Given y_i, let y_{i+1} be a solution of $R(y_i)$.

By the above discussion, Method A yields a sequence of feasible points satisfying $\varphi(y_{i+1}) \leqq \varphi(y_i)$, with strict inequality holding if y_i does not solve $R(y_i)$. Because F is compact, $\{y_i\}$ must have at least one accumulation point, and every accumulation point must lie in F. Note that in Method A, unlike the iterative procedure in § 1, we do not require that $y_{i+1} = y_i$ if y_i solves $R(y_i)$. This restriction was included in § 1 merely to assure that $\{\varphi(y_i)\}$ satisfied $\varphi(y_{i+1}) \leqq \varphi(y_i)$. By an immediate application of Theorem 1.7, then, the following theorem holds.

THEOREM 2.1. *If Γ is l.s.c. at an accumulation point y^* of a sequence $\{y_i\}$ generated by Method* A, *then y^* solves $R(y^*)$.*

There are several aspects of the previous theorem that warrant further discussion. The first point to be noted is that the compactness of F is only used to guarantee that the subproblems $R(y)$ have solutions and that the sequence $\{y_i\}$ has at least one accumulation point. It follows that compactness can be

replaced by those two hypotheses. In this case we might as well consider φ to be a real functional defined on a reverse–convex subset F of a Banach space, since the proof of the previous theorem was based solely on Theorem 1.7. It should be pointed out that while Theorem 1.1 gives a sufficient condition for l.s.c. of Γ at y^* when $F \subset E^n$, sharper results have been obtained in [9]. In particular, if S is determined by constraints of the form $g(z) \geqq 0$, where g is vector-valued and differentiable, it is sufficient that there be a point $z^* \in \Gamma(y^*)$ such that the gradients to the active constraints at z^* form a *positively linearly independent* set (i.e., no nontrivial nonnegative combination of the vectors of the set vanishes). With regard to the conclusion of the theorem only one observation will be made here. (This topic is examined in some detail in [9].) If $\Gamma(y^*)$ satisfies some form of constraint qualification at y^* (which is the case, for example, if S is a polytope), then the fact that y^* solves $R(y^*)$ implies that the Kuhn–Tucker (K–T) first order necessary conditions for a solution of I are satisfied at y^*. This is obvious, for at y^* the K–T conditions for the two problems $R(y^*)$ and I are identical.

If the function φ is differentiable on some open set containing F, we can construct the following subproblem for each point $y \in F$:

$L(y)$. Minimize

$$\varphi'(y)z$$

subject to

$$z \in \Gamma(y).$$

Of course, if φ is linear affine, the solutions to $L(y)$ coincide with the solutions of $R(y)$. However, even for the class of quasi-concave functions (which includes all linear affine functions), we have the crucial property that if y does not solve $L(y)$, then every solution $\bar{y}$ of $L(y)$ satisfies $\varphi(\bar{y}) < \varphi(y)$. This follows from the (differential) definition of quasi-concavity [8], which requires that $\varphi'(y)(\bar{y} - y) < 0$ imply $\varphi(\bar{y}) < \varphi(y)$. Consider now the following iterative method.

METHOD B.

(a) Choose an arbitrary $y_0 \in F$.

(b) Let $y_{i+1} = y_i$ if y_i solves $L(y_i)$; otherwise, let y_{i+1} be any solution of $L(y_i)$.

The following is an immediate consequence of Theorem 1.6.

THEOREM 2.2. *Let φ be quasi-concave and continuously differentiable on some open set containing F. If y^* is an accumulation point of a sequence $\{y_i\}$ generated by Method* B *and Γ is continuous at y^*, then y^* solves $L(y^*)$.*

A comparison of Theorems 2.1 and 2.2 is in order. The former is valid for all continuous objective functions (although from a numerical standpoint we can apply Method A only to objective functions with certain convexity properties), whereas the latter holds only for continuously differentiable quasi-concave functions (although Method B is numerically feasible whenever φ is differentiable). Although the last theorem specifies that Γ be continuous at y^*, it follows from a result of § 1 that Γ is u.s.c. everywhere, so that only l.s.c. at y^* need be assumed or verified. In order to apply Theorem 1.6 to prove the previous theorem, it is necessary that F be compact. As noted above, the compactness of F does not play so crucial a role in the proof of Theorem 2.1. Finally, if we assume again that $\Gamma(y^*)$ satisfies some type of constraint qualification at y^*, we conclude that if y^* solves $L(y^*)$, then y^* satisfies the K–T conditions for problem I.

In the event that φ is continuously differentiable but not quasi-concave, it is still possible to obtain algorithms in which the objective is linearized and which have the required properties if we assume that φ is twice continuously differentiable. These are based upon the observation that the linear part of the objective function dominates in a region sufficiently close to the point linearized about. Let the constant $R > 0$ be chosen so that

$$\tfrac{1}{2}(z_2 - z_1)^T \varphi''(z_1)(z_2 - z_1) \leqq R\|z_2 - z_1\|^2 \qquad \text{for all } z_1, z_2 \in F$$

(the norm is arbitrary but fixed for the remainder of the section). Given a point $y \in F$, let $M(y)$ be the set of solutions of $L(y)$. Let α be a fixed element of $(0, 1)$ and for $z \neq y$ define the real-valued function

$$K(y, z) := \min \{\alpha \cdot \varphi'(y)(y - z) \cdot \|z - y\|^{-1} \cdot R^{-1}, \|z - y\|\}.$$

For the three algorithms below, it is to be understood that: (i) y_0 is chosen arbitrarily from the feasible set F, and (ii) given y_i, that y_{i+1} is selected according to the rule of the algorithm, unless y_i itself solves $L(y_i)$. In the latter case, the rules are not used, but instead y_{i+1} is taken to be y_i. Finally, z_i^* is used below to denote an arbitrary element of $M(y_i)$.

ALGORITHM C. Choose y_{i+1} to minimize $\varphi'(y_i)z$ over the set

$$\Gamma(y_i) \cap \{z | \|z - y_i\| \leqq K(y_i, z_i^*)\}.$$

ALGORITHM D. Let $y_{i+1} = y_i + K(y_i, z_i^*) \cdot (z_i^* - y_i) \cdot \|z_i^* - y_i\|^{-1}$.

ALGORITHM E. Choose an element θ from the fixed interval $[\beta, \gamma]$, where $0 < \beta \leqq \gamma < 1$, and let $y_{i+1} = y_i + \theta^j \cdot (z_i^* - y_i)$, where i is the smallest non-negative integral exponent for which the inequality

$$\varphi(y_i + \theta^j \cdot (z_i^* - y_i)) \leqq \varphi(y_i) + (1 - \alpha) \cdot \theta^j \cdot \varphi'(y_i)(z_i^* - y_i)$$

is satisfied. (It will be shown that the previous inequality is satisfied for all sufficiently large j, so that the algorithm is well-defined.)

THEOREM 2.3. *Let φ be twice continuously differentiable on some open set containing F, and let the sequence $\{y_i\}$ be generated by one of the three procedures above. If y^* is an accumulation point of $\{y_i\}$ at which Γ is continuous, then y^* solves $L(y^*)$.*

Proof. See Appendix.

In general, the three previous algorithms will yield three different points if applied to a given point. A typical situation is shown in Fig. 2, where the points C, D and E correspond to the application of Algorithms C, D and E respectively. The dotted lines represent level lines of the linearized objective function $\varphi'(y_i)z$, and the square in the interior of $\Gamma(y_i)$ represents the set $\{z | \|z - y_i\| \leqq K(y_i, z_i^*)\}$. The figure illustrates a case in which we have chosen to work with a norm whose level surfaces are the surfaces of similar polyhedra. When these types of norms are used, and, in addition, S is a polytope, it follows that in order to obtain y_{i+1} via Algorithm C from y_i and z_i^*, we need only solve an LP problem. Hence, in this case we would solve two LP problems in order to move from y_i to y_{i+1} using Algorithm C. On the other hand, regardless of the norm used, when S is a polytope, only one LP problem must be solved when Algorithm D is used to obtain a successor to y_i. However, for both Algorithms C and D an estimate on the upper

bound of the norm of the Hessian matrix $\varphi''(z)$ is needed, and this may not be easily obtained. For Algorithm E this estimate is unnecessary, and y_{i+1} is obtained from y_i by solving one LP problem (assuming again that S is a polytope) and performing a finite number of evaluations of φ.

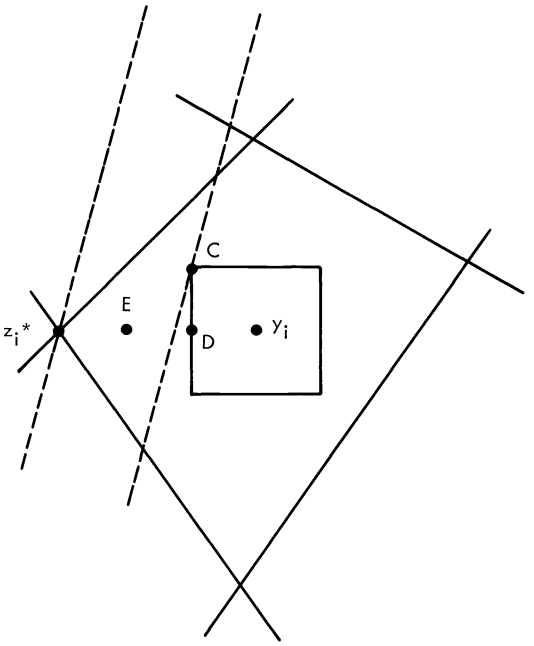


FIG. 2. *Successor points*

In the case that S is a polytope, Method B and the method corresponding to Algorithm C are special cases of the MAP method of Griffith and Stewart [10]. For the classes of minimization problems for which they are intended, the former two methods resolve the previously unsolved problem of step-size limits for the MAP method. Algorithms D and E can be contrasted with the well-known Frank–Wolfe algorithm in the special case when F is a polyhedron. (This will occur if S is a polytope and u is linear, and will mean that $\Gamma(y) = F$ for all y.) The Frank–Wolfe algorithm consists of choosing y_{i+1} to be a point on the line segment connecting y_i and z_i^* which satisfies $\varphi(y_{i+1}) \leqq \varphi(y_i) + (1 - \alpha)[(y_i^*) - \varphi(y_i)]$, where y_i^* is a point which minimizes φ on that line segment and $\alpha \in [0, 1)$. Algorithms D and E require no knowledge of the minimum of the function φ on line segments, and hence enjoy something of a theoretical advantage over the Frank–Wolfe scheme.

3. Applications to other mathematical programming algorithms. In this section we shall indicate how the results of § 1 may be applied to a number of well-known algorithms of mathematical programming.

3.1. Unconstrained minimization methods. In the notation of § 1, let $f(z, y) = \varphi(z)$ and $\Omega(y) = \{y + \lambda \cdot D(\varphi(y)) | \lambda \geqq 0\}$, where the composite function $D(\varphi)$ (which can be thought of as a direction-assigning function) is continuous for

all continuously differentiable φ and has the property that $\varphi'(y)D(\varphi(y)) \leqq 0$ with equality if and only if $\varphi'(y) = 0$. (When $D(\varphi(y))$ is chosen to be $(-\varphi'(y))^T$, the corresponding algorithm (see § 1) is the method of steepest descent. If φ is twice continuously differentiable and has a positive definite Hessian matrix at each point, then we may choose $D(\varphi(y)) = -[\varphi'(y)\varphi''^{-1}(y)]^T$. The corresponding algorithm is then a modification of the Newton–Raphson second order method.) It is clear that Ω is everywhere l.s.c. and that the iterates have the required monotonicity property. We thus conclude that an accumulation point y^* of such a method must solve the problem: Minimize $\varphi(z)$ subject to $z \in \Omega(y^*)$. This implies that $\varphi'(y^*) = 0$, and if φ is convex, y^* must be the global minimum of the unconstrained minimization problem.

3.2. Feasible direction methods. Topkis and Veinott [12] recently studied the properties of a general feasible direction algorithm which contains as special cases a feasible direction method of Zoutendijk [13], the Frank–Wolfe method [11], and second order feasible direction methods. We show below how the same general algorithm can be studied by the techniques of § 1. Again we consider a general minimization problem of the form I, but we shall assume here that the set $U = \{z | u(z) \geqq 0\}$ is convex (rather than reverse–convex as assumed in § 2). All other assumptions on the feasible set F, including compactness, are assumed to hold. We define the set $\Omega(y)$ to be those pairs (v, z) satisfying

$$v \geqq \varphi'(y)(z - y) + \tfrac{1}{2}(z - y)^T H(y)(z - y),$$

$$v \geqq -[u_i(y) + u_i'(y)(z - y)] \qquad \text{for all } i,$$

and $z \in S \cap (B + y)$, where H is a continuous mapping from E^n into the set of all positive semidefinite $n \times n$ matrices and B is a compact convex neighborhood of the origin. Letting $\mu(y) := \min \{v | (v, z) \in \Omega(y)\}$, the iterative procedure proposed by Topkis and Veinott is as follows:

(a) Choose an arbitrary $y_0 \in F$.

(b) Given y_i, let y_i^* be chosen so that $(v_i^*, y_i^*) \in \Omega(y_i)$ and $v_i^* = \mu(y_i)$; if $\mu(y_i) = 0$, let $y_{i+1} = y_i$, and if not, let y_{i+1} be a point in the intersection of F with the line segment connecting y_i and y_i^* such that $\varphi(y_{i+1}) \leqq \varphi(z)$ for all z in the intersection.

It is shown in the Appendix that the mapping Ω as defined above is continuous on S and that $\varphi(y_{i+1}) < \varphi(y_i)$ if $\mu(y_i) < 0$. By a slight modification of the proof of Theorem 1.6, it follows that a limit point y^* of the iterative procedure just described has the property that $\mu(y^*) = 0$. If some form of constraint qualification holds at y^* (for a particular case, see [12]), the relation $\mu(y^*) = 0$ implies that the Kuhn–Tucker necessary conditions for a solution of problem I must be satisfied at y^*. The Kuhn–Tucker conditions are also sufficient for optimality when φ is pseudoconvex and the constraint functions are quasi-concave (see Mangasarian [14]).

4. Generalizations. Because of such factors as finite arithmetic and rounding errors, there is little hope of obtaining exact analytic solutions to optimization problems on digital computers. One can expect at best very good approximations to the true solutions. In the theory developed in the preceding sections, however, the availability of exact solutions at each iteration was assumed. We shall now show how Theorems 1.6 and 1.7, upon which most of the results of this paper are

based, can be strengthened to provide for a certain type of approximate solution. (This type of approximation was considered by Dem'yanov and Rubinov [15] in a paper dealing with a convex programming method in Banach space. Other approximations, such as the class considered by Topkis and Veinott [12], can be handled in a similar manner.)

Let α be a fixed element of the open interval $(0, 1)$, and, using the notation and assumptions introduced for the statement of Theorem 1.6, let the sequence $\{\bar{y}_i\}$ be constructed in the following manner:

(a) Choose an arbitrary $y_0 \in G$.

(b) Let $\bar{y}_{i+1} = \bar{y}_i$ if $\bar{y}_i \in M(\bar{y}_i)$; otherwise let $\bar{y}_{i+1}$ be an element of $\Omega(\bar{y}_i)$ satisfying $\varphi(\bar{y}_i) - \varphi(\bar{y}_{i+1}) \geqq \alpha \cdot (\varphi(\bar{y}_i) - \varphi(y_i^*))$, where $y_i^* \in M(\bar{y}_i)$.

Roughly this means that at each iteration at least a fixed fraction of the theoretically possible decrease in φ is attained.

THEOREM 4.1. *If $\{\bar{y}_i\}$ and $\{y_i^*\}$ are contained in sequentially compact sets and $\bar{y}$ is an accumulation point of $\{\bar{y}_i\}$ at which Ω is continuous, then $\bar{y} \in M(\bar{y})$.*

Proof. As in the proof of Theorem 1.6, we assume that the conclusion is false, and show a contradiction. It follows from the assumptions preceding Theorem 1.6 that $\varphi(y') < \varphi(\bar{y})$ for all $y' \in M(\bar{y})$. Now let subsequences $\{\bar{y}_{n_i}\}$, $\{\bar{y}_{n_i+1}\}$, and $\{y_{n_i}^*\}$ be chosen so that $\bar{y}_{n_i} \to \bar{y}$, $\bar{y}_{n_i+1} \to \tilde{y}$, and $y_{n_i}^* \to y^*$. It follows that $y^* \in M(\bar{y})$ and that

$$0 < \varphi(\bar{y}) - \varphi(y^*) = \lim (\varphi(\bar{y}_{n_i}) - \varphi(y_{n_i}^*))$$

$$\leqq \alpha^{-1} \cdot \lim (\varphi(\bar{y}_{n_i}) - \varphi(\bar{y}_{n_i+1})) = 0,$$

which cannot hold.

By an analogous modification of the proof of Theorem 1.7, we obtain the following theorem.

THEOREM 4.2. *If $\varphi(z) = f(z, y)$ and $\bar{y}$ is an accumulation point of $\{\bar{y}_i\}$ at which Ω is l.s.c., then $\bar{y} \in M(\bar{y})$.*

Another computational aspect of algorithms that can be easily dealt with by the techniques of this paper is that of accelerating convergence by periodically taking a step in a direction other than that prescribed by the basic algorithm being used or taking slightly larger or slightly smaller steps than those prescribed. (The validity of procedures so modified has also been discussed by Topkis and Veinott [12].) It should be observed that the proofs of Theorems 1.6 and 1.7 depended only on the monotonicity of the sequence $\{\varphi(y_i)\}$ and the fact that y^* was the limit of a subsequence $\{y_{n_i}\}$ whose successor points were constructed by an algorithm with certain specified properties. Thus, if, with the goal of accelerating convergence, an algorithm without those properties is used periodically, we can conclude nevertheless that convergence of the iterates to a point y^* at which Ω is continuous (or l.s.c. in the case of Theorem 1.7) implies that $y^* \in M(y^*)$.

A further extension of Theorems 1.6 and 1.7 can be made if we note that the proofs still go through if we assume only that $\{\varphi(y_i)\}$ converges (i.e., it need not be monotonic) and that the strict monotonicity property holds at y^* (instead of everywhere). Such an extension of Theorem 1.7 can be used to prove the validity of Kelley's cutting-plane algorithm [16] in the following manner: (i) let $\{y_i\}$ be a set of points generated by Kelley's algorithm, (ii) let G be the union of $\{y_i\}$ and its accumulation points, (iii) define a point-to-set mapping Ω over G by letting

$\Omega(y_i)$, $i = 0, 1, 2, \cdots$, be the polyhedral set generated by the cutting-plane algorithm over which the objective function was minimized to obtain y_{i+1}, and $\Omega(y^*) \equiv \bigcap_{i=0}^{\infty} \Omega(y_i)$ for each accumulation point y^* of $\{y_i\}$, and (iv) show, making use of the fact that each accumulation point is feasible, that all of the assumptions of the generalized version of Theorem 1.7 are satisfied.

Appendix. The following property of sequences in normed spaces will be needed in the proof of Theorem 1.1.

LEMMA. *If* $z_i \to z$ *as* $i \to \infty$ *and* $z_{ij} \to z_i$ *as* $j \to \infty$, $i = 1, 2, \cdots$, *then there exist* $n_j, j = 1, 2, \cdots$, *such that* $z_{n_j j} \to z$ *as* $j \to \infty$.

Proof. Let $N(1)$ be chosen such that $\|z_i - z\| < 1$ for $i \geqq N(1)$, and let $N'(1)$ be chosen such that $\|z_{N(1)_j} - z_{N(1)}\| < 1$ for $j \geqq N'(1)$. Suppose now we have chosen $N(1), N(2), \cdots, N(k)$ and $N'(1), N'(2), \cdots, N'(k)$. Choose $N(k+1)$ and $N'(k+1)$ so that $\|z_i - z\| < 1/(k+1)$ for $i \geqq N(k+1)$, $\|z_{N(k+1)_j} - z_{N(k+1)}\| < 1/(k+1)$ for $j \geqq N'(k+1)$, and $N'(k+1) > N'(k)$. Let $N(0) = 1$ and define $n_j = N(l)$ when $N'(l) \leqq j < N'(l+1)$. It is easily verified that the sequence so defined satisfies $z_{n_j j} \to z$ as $j \to \infty$.

Proof of Theorem 1.1. We shall first show that the linear independence hypothesis is equivalent to assuming that there exists a point z' such that $f(z') > 0$, $u(y^*) + u'(y^*)(z' - y^*) > 0$, $v(y^*) + v'(y^*)(z' - y^*) = 0$, and that the Jacobian matrix $v'(y^*)$ has full row rank. For, we may choose a vector d such that $v'(y^*)d = 0$ and such that the inner product of d with each gradient to an active inequality constraint function at z^* is positive. It is now easily seen that a suitable choice of z' is $z^* + \theta d$, where θ is a sufficiently small positive scalar. (Since $v(y^*) + v'(y^*)(z^* - y^*) = 0$, the linear independence hypothesis implies that $v'(y^*)$ has full row rank.)

Now partition the variable z into the variables s and t (with values s' and t' at z') so that the function $\bar{v}$ defined by $\bar{v}(s, t, y) = v(y) + v'(y)(z - y)$ has a nonsingular Jacobian with respect to s at the point $(z', y^*) = (s', t', y^*)$. It follows from the implicit function theorem that there exists a neighborhood N of (t', y^*) and a differentiable function h defined on N with the properties that $h(t', y^*) = s'$ and $\bar{v}(h(t, y), t, y) = 0$ for $(t, y) \in N$. Without loss of generality we can assume that N was chosen small enough so that all of the inequality constraints involved in defining $\Gamma m(y)$ are satisfied by $(h(t, y), t)$ when $(t, y) \in N$. (This follows from elementary continuity arguments.) Hence if $\{y_i\}$ is any sequence converging to y^*, it follows that for i sufficiently large (say $i \geqq m$), we have $(t', y_i) \in N$, so that $\bar{v}(h(t', y_i), t', y_i) = 0$, and hence the equality constraints involved in defining $\Gamma m(y_i)$ are also satisfied at the point $(h(t', y_i), t') \equiv z_i$. The sequence $\{z_i\}$ so defined for $i \geqq m$ thus has the property that $z_i \in \Gamma m(y_i)$ and $z_i \to z'$. To complete the proof of l.s.c. at y^* we must prove the existence of a similar sequence for each $z \in \Gamma m(y^*)$. In order to do this, we first note that $\Gamma m(y^*)$ is a convex set, so that given any $z \in \Gamma m(y^*)$, the line segment connecting z and z' lies in $\Gamma m(y^*)$. Moreover, since

$$z \in S \cap \{z | u(y^*) + u'(y^*)(z - y) \geqq 0\}$$

and

$$z' \in \operatorname{int} S \cap \{z | u(y^*) + u'(y^*)(z - y) > 0\} =: \bar{S},$$

it follows from a well-known theorem on convex sets (see, for example, [2]) and

a simple computation that all points on that line segment with the possible exception of z also lie in $\bar{S}$. But at each point in $\bar{S} \cap \Gamma m(y^*)$ we can construct the sequence required in the definition of l.s.c. by exactly the same method used for z'. Letting $z_i' \equiv (1/i)z' + (1 - 1/i)z$ and performing such a construction for $i = 1, 2, \cdots$, we obtain a sequence of sequences from which, by the preceding lemma, we can construct a sequence converging to z and satisfying the requirements in the definition of l.s.c. This completes the proof of l.s.c. at y^*.

Now for y sufficiently close to y^* we have previously noted that the point $(h(t', y), t')$ lies in $\Gamma m(y)$ and satisfies all of the inequality constraints strictly. Since there also exists a neighborhood of y^* in which the Jacobian $v'(y)$ has full row rank, it follows that for all y in some neighborhood of y^* the point $(h(t', y), t')$ has the same properties with respect to $\Gamma m(y)$ that z' had with respect to $\Gamma m(y^*)$. Hence the proof of l.s.c. of Γm at such y may be carried out in the same manner.

The next example illustrates that the compactness hypothesis cannot be deleted in Theorem 1.6.

Example. Let $G = R = [-2, -1\frac{1}{2}] \cup [0, \frac{1}{2}] \cup [2, +\infty)$,

$$\Omega(y) = \begin{cases} \{y\} & \text{if } y \in [-2, -1\tfrac{1}{2}], \\ \{-2\} & \text{if } y = 0, \\ \{-2 + y, 1/y\} & \text{if } y \in (0, \tfrac{1}{2}], \\ \{1/(2y)\} & \text{if } y \in [2, +\infty); \end{cases}$$

$$f(z, y) = \begin{cases} z + 2 & \text{if } z \in [-2, -1\tfrac{1}{2}], \\ 0 & \text{if } z \in [0, \tfrac{1}{2}], \\ 1/z & \text{if } z \in [2, +\infty); \end{cases}$$

$$\varphi(y) = \begin{cases} y + 2 & \text{if } y \in [-2, -1\tfrac{1}{2}], \\ y + 1 & \text{if } y \in [0, \tfrac{1}{2}], \\ 1 + 2/(3y) & \text{if } y \in [2, +\infty). \end{cases}$$

It is easily verified that with the above definitions the conditions stated prior to Theorem 1.6 are satisfied, that Ω is continuous on G, and that f and φ may be extended to continuous functions on E^2 and E^1 respectively. Suppose that we choose $y_0 = \frac{1}{2}$. It may be verified that $M(y_0) = \{-1\frac{1}{2}, 2\}$, so that we can choose $y_1 = 2$. Since $\Omega(y_1) = \{\frac{1}{4}\}$, it follows that $y_2 = \frac{1}{4}$. Continuing in a similar fashion, we obtain the sequence of iterates $\{\frac{1}{2}, 2, \frac{1}{4}, 4, \frac{1}{8}, \cdots\}$. However, the accumulation point 0 does not belong to $M(0) = \{-2\}$.

Proof of Theorem 2.3. We shall show that all three algorithms have the strict monotonicity property. Using a second order Taylor expansion and the definition of R, we obtain for $z \in F$ the inequality $\varphi(z) \leqq \varphi(y_i) + \varphi'(y_i)(z - y_i) + R \cdot \|z - y_i\|^2$. If $\|z - y_i\| \leqq \delta \cdot K(y_i, z_i^*)$, this becomes

$$\begin{aligned} \varphi(z) &\leqq \varphi(y_i) + \varphi'(y_i)(z - y_i) + R \cdot \delta^2 \cdot K^2(y_i, z_i^*) \\ &\leqq \varphi(y_i) + \varphi'(y_i)(z - y_i) \\ &\quad - \delta^2 \cdot \alpha \cdot \varphi'(y_i)(z_i^* - y_i) \cdot \|z_i^* - y_i\|^{-1} \cdot K(y_i, z_i^*). \end{aligned}$$

If we denote by y' the point generated by applying Algorithm D to y_i, we have $y' - y_i = (z_i^* - y_i) \cdot \|z_i^* - y_i\|^{-1} \cdot K(y_i, z_i^*)$, and the inequality reduces to $\varphi(z) \leqq \varphi(y_i) + \varphi'(y_i)(z - y_i) - \delta^2 \cdot \alpha \cdot \varphi'(y_i)(y' - y_i)$. Three cases will now be considered: (i) if $z = y'$, choose $\delta = 1$, and the inequality becomes $\varphi(z) \leqq \varphi(y_i) + (1 - \alpha)\varphi'(y_i)(y' - y_i)$; (ii) if z is generated by Algorithm C, choose $\delta = 1$, and it follows from $\varphi'(y_i)z \leqq \varphi'(y_i)y'$ that $\varphi(z) \leqq \varphi(y_i) + \varphi'(y_i)(y' - y_i) - \alpha \cdot \varphi'(y_i) \cdot (y' - y_i) = \varphi(y_i) + (1 - \alpha)\varphi'(y_i)(y' - y_i)$; and (iii) if $z = y_i + \omega \cdot (y' - y_i)$, where $0 \leqq \omega \leqq 1$, choose $\delta = \omega$, yielding

$$\begin{aligned}\varphi(z) &\leqq \varphi(y_i) + \omega \cdot \varphi'(y_i)(y' - y_i) - \omega^2 \cdot \alpha \cdot \varphi'(y_i)(y' - y_i)\\ &= \varphi(y_i) + (1 - \omega\alpha) \cdot \omega \cdot \varphi'(y_i)(y' - y_i)\\ &\leqq \varphi(y_i) + (1 - \alpha) \cdot \omega \cdot \varphi'(y_i)(y' - y_i).\end{aligned}$$

By the analysis in case (iii), it is easily seen that in Algorithm E the relation

$$\varphi(y_{i+1}) \leqq \varphi(y_i) + (1 - \alpha) \cdot \theta_j \cdot \varphi'(y_i)(z_i^* - y_i),$$

where $y_{i+1} = y_i + \theta_j \cdot (z_i^* - y_i)$ is satisfied if $\theta_j \cdot \|z_i^* - y_i\| \leqq K(y_i, z_i^*)$, proving that the algorithm is well-defined. Moreover, since $\beta \leqq \theta$, the point z generated by Algorithm E must satisfy $\varphi(z) \leqq \varphi(y_i) + (1 - \alpha) \cdot \beta \cdot \varphi'(y_i)(y' - y_i)$. For all three algorithms, then, $\{\varphi(y_i)\}$ is a nonincreasing sequence.

Let y^* be an accumulation point of $\{y_i\}$ at which the point-to-set mapping Γ is continuous. Choose subsequences $\{y_{n_i}\}$, $\{y_{n_i+1}\}$ and $\{z_{n_i}^*\}$ such that $y_{n_i} \to y^*$ and the latter two are convergent with limit points $\bar{y}$ and z^* respectively. As a consequence of Theorem 1.5, z^* is a solution of $L(y^*)$. If we now suppose that y^* does not solve $L(y^*)$, then $K(y^*, z^*) > 0$. For Algorithms C and D we thus have

$$\begin{aligned}\varphi(\bar{y}) &= \lim \varphi(y_{n_i+1}) \leqq \lim [\varphi(y_{n_i}) + (1 - \alpha) \cdot \varphi'(y_{n_i})(y_{n_i+1} - y_{n_i})]\\ &= \varphi(y^*) + (1 - \alpha) \cdot \varphi'(y^*)(\bar{y} - y^*)\\ &= \varphi(y^*) + (1 - \alpha) \cdot \varphi'(y^*)(z^* - y^*) \cdot \|z^* - y^*\|^{-1} \cdot K(y^*, z^*)\\ &< \varphi(y^*).\end{aligned}$$

This is impossible, however, since $\{\varphi(y_i)\}$ is a nonincreasing sequence. By inserting the factor β in the appropriate places, we can prove the conclusion for Algorithm E. (Alternative proofs have been constructed (see Meyer [9]) for Methods C and D by establishing the strict monotonicity property and the continuity of certain point-to-set mappings. In this way the conclusion is obtained as a direct consequence of Theorem 1.6, but at the expense of increasing the complexity of the proof.)

Proof of assertions in § 3.2. By using the continuity of the terms involved, it is easily shown that Ω is everywhere u.s.c. To prove l.s.c. on S, we first observe that the point-to-set mapping defined by $\Omega'(y) = S \cap (B + y)$ is l.s.c. on S. This is seen by noting that interior points of $B + y$ that lie in S also lie in $B + \bar{y}$ for $\bar{y}$ sufficiently close to y, and that a boundary point of $B + y$ that lies in S is the limit of interior points of $B + y$ contained in S. Now let (v, z) be an arbitrary point of $\Omega(y)$ and let $\{y_n\}$ be a sequence of points in S converging to y. By the preceding argument, there exists a sequence of points $\{z_n\}$ with $z_n \in \Omega'(y_n)$ converging to z.

It is clear that a sequence $\{v_n\}$ converging to v can now be chosen so that $(v_n, z_n) \in \Omega(y_n)$, completing the proof of l.s.c.

To simplify notation, we shall drop the subscripts in the following proof of the monotonicity property asserted in § 3.2. Suppose that there exists a point $(v^*, z^*) \in \Omega(y)$ with $v^* < 0$. We show that for sufficiently small positive λ, the point $\bar{z} = y + \lambda(z^* - y)$ belongs to F and satisfies $\varphi(\bar{z}) < \varphi(y)$. It is clear that $\bar{z} \in S \cap (B + y)$ for $\lambda \in [0, 1]$, so to prove feasibility we need only show that $u(\bar{z}) \geqq 0$. If $u_i(y) > 0$, then clearly $u(\bar{z}) > 0$ for λ sufficiently small; and if $u_i(y) = 0$, then $0 > v^* \geqq -[u_i(y) + u_i'(y)(z^* - y)]$ implies $u_i'(y)(z^* - y) > 0$, and again it is true that $u_i(\bar{z}) > 0$ for sufficiently small positive λ. Since $H(y)$ is assumed positive semidefinite, $0 > v^* \geqq \varphi(y)(z^* - y) + \frac{1}{2}(z^* - y)H(y)(z^* - y)$ implies $0 > \varphi'(y) \cdot (z^* - y)$, and the required result follows.

REFERENCES

[1] F. HAUSDORFF, *Set Theory*, Chelsea, New York, 1962.

[2] C. BERGE, *Topological Spaces*, Macmillan, New York, 1963.

[3] GEORGE B. DANTZIG, JON FOLKMAN AND NORMAN SHAPIRO, *On the continuity of the minimum set of a continuous function*, J. Math. Anal. Appl., 17 (1967), pp. 519–548.

[4] M. Q. JACOBS, *Some existence theorems for linear optimal control problems*, this Journal, 5 (1967), pp. 418–437.

[5] G. DEBREU, *Theory of Value*, Cowles Foundation Monograph No. 17, John Wiley, New York, 1959.

[6] J. B. ROSEN, *Iterative solution of nonlinear optimal control problems*, this Journal, 4 (1966), pp. 223–244.

[7] W. I. ZANGWILL, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, New Jersey, 1969.

[8] J. PONSTEIN, *Seven kinds of convexity*, SIAM Rev., 9 (1967), pp. 115–119.

[9] R. R. MEYER, *The solution of non-convex optimization problems by iterative convex programming*, Doctoral thesis, The University of Wisconsin, Madison, 1968.

[10] R. E. GRIFFITH AND R. A. STEWART, *A nonlinear programming technique for the optimization of continuous processing systems*, Management Sci., 7 (1961), pp. 379–392.

[11] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[12] D. M. TOPKIS AND A. F. VEINOTT, JR., *On the convergence of some feasible direction algorithms for nonlinear programming*, this Journal, 5 (1967), pp. 280–294.

[13] G. ZOUTENDIJK, *Methods of Feasible Directions*, American Elsevier, New York, 1960.

[14] O. L. MANGASARIAN, *Pseudo-convex functions*, this Journal, 3 (1965), pp. 281–290.

[15] V. F. DEM'YANOV AND A. M. RUBINOV, *The minimization of a smooth convex functional on a convex set*, this Journal, 5 (1967), pp. 268–279.

[16] J. E. KELLEY, *The cutting plane method for solving convex programs*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 703–712.

SIAM J. CONTROL
Vol. 8, No. 1, February 1970

ON PERFORMANCE BOUNDS FOR UNCERTAIN SYSTEMS*

H. S. WITSENHAUSEN†

1. Generalities.

1.1. Introduction. A central problem with uncertain systems is to choose one element α out of a set A of possible decisions (designs, strategies, policies, controllers, estimators, coding schemes etc.). The performance is measured, at first, in terms of a function $K: A \times B \to [0, \infty]$ where $K(\alpha, \beta)$ is the cost incurred with decision α if the uncertain quantities affecting the system have the (system of) values denoted by β from set B (see [8]).

In principle the case of randomized decisions can be included in this framework by considering A to be the set of all possible randomizations and defining K accordingly, say by the expectation of an underlying cost function. But the intended interpretation for the sequel is the case where randomization is considered undesirable.

Two designs α_1, α_2 in A are said to be equivalent if and only if $K(\alpha_1, \beta) = K(\alpha_2, \beta)$ for all β in B. Equivalence classes of designs are partially ordered in the obvious way, writing $\alpha_1 \leqq \alpha_2$ when $K(\alpha_1, \beta) \leqq K(\alpha_2, \beta)$ for all β in B. A dominant choice α^* (an element of A such that $\alpha^* \leqq \alpha$ for all α in A) rarely exists. One way to proceed to a decision is to define a *supercriterion* $J: A \to [0, \infty]$ and seek designs which minimize J over A, either exactly or within ε. A priori, the minimal requirement on J is that it be compatible with the partial order, $\alpha_1 \leqq \alpha_2$ implying $J(\alpha_1) \leqq J(\alpha_2)$. This leaves room for a wealth of possibilities, such as the regret criterion of Savage [9]:

$$J(\alpha) = \sup_{\beta \in Q} (K(\alpha, \beta) - \inf_{\bar{\alpha} \in A} K(\bar{\alpha}, \beta)), \tag{1}$$

where $Q \subset B$ is given. With the latter definition the sign of $J(\alpha_1) - J(\alpha_2)$ can change when a third element α_3 is dropped from the set A. If such phenomena are not desired it becomes necessary to put further restrictions upon J. An important class of supercriteria is obtained by requiring that $J(\alpha)$ be defined to depend only on the function $K(\alpha, \cdot): B \to [0, \infty]$, by way of a functional V called an *evaluator*. That is

$$J(\alpha) = V(K(\alpha, \cdot)). \tag{2}$$

The problem (A, B, K, V) is then to determine the infimum J^* of J over A, with J defined by (2), and to determine an element of A for which this value is attained exactly or within ε.

Because of the difficulty of this task, suboptimal procedures are often used. One consists of selecting an element β_0 in B as being "typical" and seeking α_0 to minimize $K(\alpha, \beta_0)$ over A. In this paper bounds on $J(\alpha_0)/J^*$ are sought. Such bounds require of course more detailed assumptions about K, V and the notion of typical element. This motivates the following definitions.

* Received by the editors March 11, 1969.

† Bell Telephone Laboratories, Murray Hill, New Jersey, 07974.

1.2. Basic definitions.

DEFINITION 1. An *evaluator* for the problem (A, B, K) is a function

$$V: D \to [0, \infty], \tag{3}$$

where D is a subset of $[0, \infty]^B$, and which satisfies the following three conditions:

(i) $K(\alpha, \cdot) \in D$ for all $\alpha \in A$;

(ii) if $f_1, f_2 \in D$ and, for all β in B, $f_1(\beta) \leqq f_2(\beta)$ then $V(f_1) \leqq V(f_2)$;

(iii) if $f(\beta) = a \geqq 0$ for all β in B then $f \in D$ and $V(f) = a$.

DEFINITION 2. Evaluator V is *weakly subadditive* when $f \in D$, $a \geqq 0$ imply $f + a \in D$, $V(f + a) \leqq a + V(f)$. It is *subadditive* when $f, g \in D$ imply $f + g \in D$, $V(f + g) \leqq V(f) + V(g)$.

DEFINITION 3. An element β_0 of B is called a *representative element* for problem (A, B, K, V) when $K(\alpha, \beta_0) \leqq J(\alpha)$ for all α in A.

This generalizes beyond the stochastic realm a definition of Fréchet [5], itself a generalization of a definition by Doss [2] of one mean value concept for random elements of metric spaces.

1.3. Examples of evaluators. The most common evaluators are the following.

The stochastic mean of order $p \geqq 1$ is defined by specifying a σ-field on B and a probability measure on this field. Then, for $p \geqq 1$, f measurable,

$$V_p(f) = (E\{f^p(\beta)\})^{1/p} \tag{4}$$

and

$$V_\infty(f) = \operatorname{ess\,sup} f(\beta). \tag{5}$$

One nonstochastic evaluator is obtained by specifying a subset Ω of B with

$$V_\Omega(f) = \sup\{f(\beta) | \beta \in \Omega\}. \tag{6}$$

Then every element of Ω is representative. In general, evaluators can be composed by applying first an evaluator containing a parametric element and applying another evaluator to the parameter set, a process which can be continued through any number of stages. For example, if M is a set of probability measures on a common σ-field on B then an evaluator is defined by

$$V_M(f) = \sup_{\mu \in M} E_\mu\{f(\beta)\}.$$

All of the above examples are subadditive evaluators.

1.4. The zig-zag inequality.

DEFINITION 4. K satisfies the zig-zag inequality when for all α_1, α_2 in A and all β_1, β_2 in B

$$K(\alpha_1, \beta_1) \leqq K(\alpha_1, \beta_2) + K(\alpha_2, \beta_2) + K(\alpha_2, \beta_1). \tag{7}$$

The importance of this notion derives from the following fact.

THEOREM 1. *Let $K: A \times B \to [0, \infty)$. Then a necessary and sufficient condition for the existence of a real normed linear space L, $\|\cdot\|$ and of maps $m: A \to L$,*

$n: B \to L$ *such that*

$$K(\alpha, \beta) \equiv \|m(\alpha) - n(\beta)\| \tag{8}$$

is that K satisfy the zig-zag inequality.

Proof. First, one may assume that the sets A and B are disjoint (otherwise one lifts the definition of K to a pair of disjoint copies of A and B). On the union P of A and B define a pseudometric d in the following way:

If $\alpha \in A$, $\beta \in B$ then $d(\alpha, \beta) = d(\beta, \alpha) = K(\alpha, \beta)$.

If $\alpha \in A$, $\alpha' \in A$ then

$$d(\alpha, \alpha') = \sup_{\beta \in B} |K(\alpha, \beta) - K(\alpha', \beta)|. \tag{9}$$

If $\beta \in B$, $\beta' \in B$ then

$$d(\beta, \beta') = \sup_{\alpha \in A} |K(\alpha, \beta) - K(\alpha, \beta')|. \tag{10}$$

Note that (9) and (10) are the definitions of Wald's intrinsic pseudometric [10] on A and B. Because of the zig-zag inequality, d is a pseudometric on the union P. The equivalence relation $d(x, y) = 0$ on P defines a quotient space of equivalence classes on which d defines a metric. Let M be this metric space and q the quotient mapping. Let L be the linear space of bounded, continuous real functions on M with the supremum norm and let θ be an arbitrary fixed reference element of M. Define, according to a well-known technique [6], the mapping $\varphi: M \to L$ which sends an element of x of M into the element of L which, as a function $f: M \to R$, has the values: $f(z) = d(z, x) - d(z, \theta)$.

Then φ maps M isometrically into L. Denote by $i_A: A \to P$ and $i_B: B \to P$ the injection maps of A, B into their disjoint union. Then the compositions $m = \varphi \circ q \circ i_A$, $n = \varphi \circ q \circ i_B$ prove sufficiency. The triangle inequality implies the zig-zag inequality, proving necessity.

1.5. A general bound for suboptimal performance. For problem (A, B, K, V) let J be defined by (2) and $J^* = \inf\{J(\alpha) | \alpha \in A\}$. Assume β_0 is a representative element as per Definition 3, that is, for all α in A,

$$K(\alpha, \beta_0) \leqq V(K(\alpha, \cdot)); \tag{11}$$

and assume that α_0 is optimal versus β_0, that is, for all α in A

$$K(\alpha_0, \beta_0) \leqq K(\alpha, \beta_0); \tag{12}$$

and let

$$J_0 = J(\alpha_0) = V(K(\alpha_0, \cdot)). \tag{13}$$

Then one has the following result.

THEOREM 2. *If V is weakly subadditive and K satisfies the zig-zag inequality, then* $J_0 \leqq 3J^*$ *and this bound is sharp.*

Proof. By Definition 4, for all $\alpha \in A$, $\beta \in B$,

$$\begin{aligned} K(\alpha_0, \beta) &\leqq K(\alpha_0, \beta_0) + K(\alpha, \beta_0) + K(\alpha, \beta) \\ &\leqq 2K(\alpha, \beta_0) + K(\alpha, \beta) \qquad \text{by (12).} \end{aligned}$$

By Definition 1 (ii),

$$\begin{aligned} V(K(\alpha_0, \cdot)) &\leqq V(2K(\alpha, \beta_0) + K(\alpha, \cdot)) \\ &\leqq 2K(\alpha, \beta_0) + V(K(\alpha, \cdot)) \qquad \text{(by Definition 2)} \\ &\leqq 3V(K(\alpha, \cdot)) \qquad \text{by (11)}, \end{aligned}$$

or $J_0 = J(\alpha_0) \leqq 3J(\alpha)$; and, taking the infimum over all α in A, $J_0 \leqq 3J^*$ as claimed. The bound is attained already under much more special conditions, as was shown earlier [11].

1.6. Synopsis. In view of Theorems 1 and 2, most of the remainder is devoted to the case where $K(\alpha, \beta)$ is defined as $\|\alpha - \beta\|$ in a general normed space N. The stochastic mean of order p is taken as the evaluator. Theorem 3 below shows that the bounds for $p = \infty$ also apply to the nonstochastic minimax problems.

A direct geometric interpretation for the bounds is constructed, first in N and then in a space of random vectors with values in N. In this way a simple and general proof of the equality of the bounds for conjugate exponents p is obtained (Theorems 4 to 10). An alternative approach by Lagrange multipliers is described in § 4.2 for its computational value. But since the bound is the supremum of an expression which is neither concave nor (in general) differentiable and need not be attained, the direct approach is preferable.

The consequences of various assumptions are explored: The distribution of β may be symmetric about its mean; the constraint set for α may be convex. The normed space N may be a pre-Hilbert space; the dimension of N may have a given finite value. Lemmas 11 to 16 make explicit some of the simplifications upon which actual calculations are based.

A few of the bounds are then computed either in theorems or by machine. In the latter case the problem is analyzed to the point where one can guarantee that convergence of the computer program will be convergence to the correct solution. Some asymptotic formulas are derived and a logarithmic convexity conjecture is stated.

1.7. Relations between supremum and essential supremum evaluators. Let B be provided with Wald's intrinsic (pseudo-) metric d according to (10). By lifting the definition of K to the quotient space, if necessary, one can assume that d is a metric, (B, d) a metric space.

In view of Definition 3, two evaluators that produce the same supercriterion J, in (2), from a given K are equivalent for the purposes of this paper.

THEOREM 3. *If (B, d) is separable, then any supercriterion J produced from K by taking the essential supremum under a probability measure on the Borel sets of (B, d) can also be obtained by taking the supremum over some nonempty set Q in B, and vice versa.*

Proof. Any separable metric space is a second-countable Hausdorff space. For all α in A, $K(\alpha, \cdot)$ is Lipschitz-continuous on (B, d) with constant one, by virtue of (10). Hence Theorem 3 is a special case of the following lemma.

LEMMA 1. *Let T be a second-countable Hausdorff space. Then for every probability measure P on the Borel sets of T there is a set $S \subset T$ such that for every*

real function f continuous on T

$$\sup_{x\in S} f(x) = P\text{-ess}\sup_{x\in T} f(x). \tag{14}$$

Conversely, for every nonempty set S in T there is a P such that (14) *holds for all continuous f.*

Proof. Let P be given. By the Lindelöf property of second-countable spaces the union of all open null sets is equal to a countable subunion and is therefore a null set. Its complement S, the closed support of P, is the smallest closed set of probability one. Then $\sup f \geqq \text{ess sup } f$ because S has probability one. If $\sup f > \text{ess sup } f = a$, then the closed set $f^{-1}([-\infty, a]) \cap S$ would be a closed proper subset of S of probability one, a contradiction.

Now assume S given. Since continuous f are considered one may assume S closed. As a subspace of a second-countable space, S is separable. With $\{x_i \in S | i = 1, 2, \cdots\}$ dense in S let $P(\{x_i\}) = 2^{-i}$; then S is the closed support of P so that (14) holds.

1.8. Uncertainty cost and clairvoyance premium. If the assumption that $\beta = \beta_0$ were actually correct, then the cost with decision α would be just $K(\alpha, \beta_0)$ instead of $J(\alpha) = V(K(\alpha, \cdot))$. The difference $J(\alpha) - K(\alpha, \beta_0)$ may therefore be considered as the *cost of uncertainty.*

The interpretation of Definition 3 is that an element β_0 is representative when the cost of uncertainty is nonnegative for all possible decisions; that is, for every decision, one is better off with β fixed at β_0 rather than uncertain.

Now imagine that just before α must be selected, a medium (spy, instrument) reveals the actual value of β. Then the decision can be chosen to yield, exactly or within ε, the cost

$$m(\beta) = \inf_{\alpha\in A} K(\alpha, \beta).$$

Before the medium speaks out, though, $m(\beta)$ is an uncertain quantity. Thus the merits of this imaginary situation must still be evaluated. Assume m belongs to the domain of evaluator V. Then, since for all α and β, $m(\beta) \leqq K(\alpha, \beta)$,

$$\lambda = V(m(\cdot)) \leqq V(K(\alpha, \cdot)) = J(\alpha),$$

and taking the infimum over α,

$$\lambda \leqq J^*.$$

The difference $J^* - \lambda$, the *premium for clairvoyance*, can therefore not be negative but it can be zero. If it is positive, then part of this premium might be collected by partial clairvoyance, that is, by some increase in data gathering (feedback) during the decision process. Note that (A, B, K) is the reduced canonical form of the problem as opposed to the extensive form in which the time sequence of events is displayed.

In this paper, following [1], *certainty equivalence* is said to hold for a class of cases if $J_0 = J^*$ for all these cases. This does by no means imply that the premium for clairvoyance or the cost of uncertainty need be equal to zero.

2. Means of order p of norms.

2.1. Construction of spaces. The sequel will be devoted to the case where K

can be defined by a norm and the evaluator is the stochastic mean of order p, with $1 \leqq p \leqq \infty$.

Let N, $\|\cdot\|$ be a real normed linear space. N serves as the set B of § 1.1, while the set A is a nonempty subset of N.

Consider the cost function

$$K(\alpha, \beta) = \|\alpha - \beta\|. \tag{15}$$

Let $(\Omega, \mathscr{F}, P)$ be a probability space, i.e., Ω is a nonempty set, $\mathscr{F}$ a σ-algebra of subsets of Ω, and P a probability measure on $\mathscr{F}$.

For fixed p in $[1, \infty]$ consider the Bochner integrable [7] functions $q:\Omega \to N$ such that:

(i) $\bar{q} \equiv E\{q(\omega)\}$ belongs to N, not just to the completion of N.

(ii) $E\{\|q(\omega)\|^p\} < \infty$, or, for $p = \infty$, ess sup $\|q(\omega)\| < \infty$.

The set of all functions q satisfying the above requirements is a linear space under pointwise addition and scalar multiplication, and the same is true for the set of all equivalence classes of these functions, modulo almost sure equality. In the sequel, equivalence classes will, abusively, be referred to as functions. Their linear space will be normed by letting

$$|||q||| = (E\{\|q(\omega)\|^p\})^{1/p}, \tag{16}$$

or, for $p = \infty$,

$$|||q||| = \text{ess sup } \|q(\omega)\|, \tag{17}$$

where triple bars are used to avoid confusion with the norm of N.

This normed space will be denoted $\mathscr{N}(\Omega, \mathscr{F}, P, N, p)$ or more briefly by $\mathscr{N}$.

One may always consider N as a dense subset of its completion $\bar{N}$. The Bochner integrable functions with values in $\bar{N}$ satisfying requirement (ii) above form a Banach space. In this space, those functions which have an N-valued version form a subspace. Another subspace is the set of functions whose mean belongs to N. $\mathscr{N}$ is isometrically isomorphic to the intersection of those two subspaces, under the natural embedding.

It is important to note that $\mathscr{N}$ contains all simple functions, that is, all functions of which a version has finite range, each of the values being taken on a measurable set. Since P is a finite measure and all functions in $\mathscr{N}$ are Bochner integrable, the linear subspace of simple functions is dense in $\mathscr{N}$ in the sense of convergence in measure. It is norm dense in $\mathscr{N}$ for $1 \leqq p < \infty$.

2.2. Geometric interpretation in N. Any choice of a normed space N, an order $p \geqq 1$, an element q of $\mathscr{N}$ and a subset of A of N defines a decision problem with the cost (15) and the pth order mean evaluator. The resulting supercriterion $J:N \to [0, \infty)$ is given by

$$J(x) = (E\{\|x - q(\omega)\|^p\})^{1/p}, \tag{18}$$

or, for $p = \infty$, by

$$J(x) = \text{ess sup } \|x - q(\omega)\|. \tag{19}$$

That is, (4) and (5) are applied with the probability structure induced on $B = N$ by q. The optimal performance is $J^* = \inf_{x \in A} J(x)$.

LEMMA 2. *The mean* $\bar{q} = E\{q(\omega)\}$ *is a representative element in the sense of Definition* 3.

Proof. Let S be the set of all linear functionals φ on N with induced norm

$$\|\varphi\| = \sup_{x \in N} \frac{\langle \varphi, x \rangle}{\|x\|} \leqq 1,$$

where, as usual, the occurrence of 0/0 under the supremum sign is ignored because sup {num./den.|conditions} is interpreted as inf {k|conditions $\Rightarrow$ num. $\leqq k$ den.}. This interpretation is used throughout the sequel.

Then $\|x\| = \sup \{\langle \varphi, x \rangle | \varphi \in S\}$ (because the norm-dual is always norm determining). Since q is Bochner integrable, $\|q\|$ and $\langle \varphi, q \rangle$ are integrable. Thus one has, for $\beta_0 = \bar{q}$ and any $\alpha \in N$,

$$\begin{aligned} K(\alpha, \beta_0) &= \|\alpha - \bar{q}\| \\ &= \sup \{\langle \varphi, \alpha - \bar{q} \rangle | \varphi \in S\} \\ &= \sup \{E\{\langle \varphi, \alpha - q(\omega) \rangle\} | \varphi \in S\} \\ &\leqq E\{\sup \{\langle \varphi, \alpha - q(\omega) \rangle | \varphi \in S\}\} \\ &= E\{\|\alpha - q(\omega)\|\} \\ &\leqq (E\{\|\alpha - q(\omega)\|^p\})^{1/p} \\ &\leqq \operatorname{ess\,sup} \|\alpha - q(\omega)\|, \end{aligned}$$

where the last two inequalities hold by the monotonicity in p of the pth order mean.

Hence $K(\alpha, \beta_0) \leqq J(\alpha)$ for all α in N and, in particular, for α in A, satisfying Definition 3 as claimed.

With $\beta_0 = \bar{q}$, a suboptimal choice α_0 is one that satisfies (12), i.e.,

$$\alpha_0 \in A \quad \text{and} \quad \forall \alpha \in A : \|\alpha_0 - \bar{q}\| \leqq \|\alpha - \bar{q}\|. \tag{20}$$

The least value of k such that $J_0 = J(\alpha_0) \leqq kJ^*$, i.e., the supremum of J_0/J^*, is sought under various assumptions concerning Ω, $\mathscr{F}$, P, N, A. A bound k holds on a set A when the conditions (20) imply $J_0 \leqq kJ^*$. This means that on a set where no suboptimal α_0 exists any bound holds and on a set where several choices for α_0 exist k must be valid for every possible choice in order to qualify.

The parallelogram law. One important possible assumption is that the parallelogram law

$$\|x - y\|^2 + \|x + y\|^2 = 2\|x\|^2 + 2\|y\|^2 \tag{21}$$

holds in N. Then N is a real inner product (pre-Hilbert) space with the inner product given by

$$x \cdot y = \tfrac{1}{4}(\|x + y\|^2 - \|x - y\|^2). \tag{22}$$

In this connection, one has the so-called "certainty equivalence" property for quadratic criteria [1].

LEMMA 3. *If the parallelogram law holds in* N *and* $p = 2$, *then this law holds in* $\mathscr{N}$ *and* $J_0 = J^*$.

Proof. For x, y in $\mathcal{N}$, (21) holds pointwise, and by integration with

$$|||x|||^2 = E\{\|x(\omega)\|^2\} \tag{23}$$

it is seen to hold in $\mathcal{N}$. Because $\mathcal{N}$ is now an inner product space one has

$$\begin{aligned} J^2(\alpha) &= E\{\|\alpha - q(\omega)\|^2\} \\ &= \|\alpha - \bar{q}\|^2 + E\{\|q(\omega) - \bar{q}\|^2\}, \end{aligned} \tag{24}$$

where $\bar{q} = E\{q(\omega)\}$. If α_0 minimizes $\|\alpha - \bar{q}\|$ in A, then it minimizes the right side of (24) and thereby also $J(\alpha)$. Hence $J(\alpha_0) = J^*$.

Convexity. For a fixed element $q \in \mathcal{N}(\Omega, \mathcal{F}, P, N, p)$ the bound k, for all nonempty sets A in N, is just

$$k = \sup \left\{ \frac{J(u)}{J(v)} \middle| u, v \in N, \|u - \bar{q}\| \leqq \|v - \bar{q}\| \right\}. \tag{25}$$

Indeed if $u = \alpha_0$ satisfies (20) on some set A, then for any point $v \in A$, $J(u) \leqq kJ(v)$ by (25); whence $J_0 \leqq kJ^*$ showing that the bound holds. But for any pair of points u, v satisfying $\|u - \bar{q}\| \leqq \|v - \bar{q}\|$ one possible choice of A is $\{u, v\}$, with $\alpha_0 = u$; hence the bound is sharp. If the bound k_c holding over all *convex* sets A is sought, then

$$k_c = \sup \left\{ \frac{J(u)}{J(v)} \middle| u, v \in N, \forall\, \theta \in [0, 1] : \|u - \bar{q}\| \leqq \|\theta u + (1 - \theta)v - \bar{q}\| \right\} \tag{26}$$

because if $u = \alpha_0$ satisfies (20) on a convex set A and v is another point of A, the segment uv belongs to A, and by (26), $J(u) \leqq k_c J(v)$, implying that the bound holds. For u, v satisfying the condition in (26), the segment uv is a convex choice for A with $\alpha_0 = u$, and therefore the bound is sharp.

When the parallelogram law holds, the condition

$$\|u - \bar{q}\| \leqq \|\theta u + (1 - \theta)v - \bar{q}\| \quad \text{for all } \theta \in [0, 1] \tag{27}$$

is equivalent to

$$\left\| u - \frac{v + \bar{q}}{2} \right\| \leqq \frac{1}{2}\|v - \bar{q}\|. \tag{28}$$

For fixed $\bar{q}$ and v, equation (28) restricts u to a sphere, while in general the set C_v of all u satisfying (27) need not even be convex, though it is always star shaped with respect to both $\bar{q}$ and v.

Symmetry. The symmetry assumption is that the probability measure generated in N by q is symmetric. More precisely, the involution $x \to 2\bar{q} - x$ on N is assumed measure preserving. This implies $J(x) = J(2\bar{q} - x)$. The resulting simplification is great. First, since J is convex, $\bar{q}$ minimizes J on N, regardless of the value of p. Most important is the following.

LEMMA 4. *Let the space N, the class of subsets A and the order p be fixed. Let k be the supremum of J_0/J^* over all those choices of Ω, $\mathcal{F}$, P and of $q \in \mathcal{N}(\Omega, \mathcal{F}, P, N, p)$ that generate a symmetric distribution in N. Let k' be the supremum over all q with the fixed choice $\Omega = \{\omega_1, \omega_2\}$, $\mathcal{F} = 2^\Omega$, $P(\omega_1) = P(\omega_2)$. Then $k = k'$.*

Proof. With that choice of $(\Omega, \mathcal{F}, P)$, every q generates a two-point symmetric distribution and all such distributions can be so generated (all functions on Ω are simple functions, hence belong to $\mathcal{N}$). Therefore $k' \leqq k$ by inclusion. To show the reverse inequality let μ be any probability measure on the Borel sets of N, symmetric about some point $\bar{q}$, which is taken as the origin without loss of generality. Then J is finite everywhere if finite somewhere, and when α_0 satisfies (20) in A and $1 \leqq p < \infty$, one has

$$\begin{aligned} J_0^p = J^p(\alpha_0) &= E\{\|\alpha_0 - q(\omega)\|^p\} \\ &= E\{\|\alpha_0 + q(\omega)\|^p\} \qquad \text{(by symmetry)} \\ &= \tfrac{1}{2}E\{\|\alpha_0 - q(\omega)\|^p + \|\alpha_0 + q(\omega)\|^p\}. \end{aligned}$$

However, by definition of k' one has pointwise

$$\|\alpha_0 - q(\omega)\|^p + \|\alpha_0 + q(\omega)\|^p \leqq k'^p(\|\alpha - q(\omega)\|^p + \|\alpha + q(\omega)\|^p)$$

for all $\omega \in \Omega$ and $\alpha \in A$. Integration gives

$$\begin{aligned} J_0^p &\leqq k'^p \tfrac{1}{2}(E\{\|\alpha - q(\omega)\|^p + \|\alpha + q(\omega)\|^p\}) \\ &= k'^p E\{\|\alpha - q(\omega)\|^p\} \qquad \text{(by symmetry)} \\ &= k'^p J(\alpha)^p. \end{aligned}$$

Taking the infimum over α in A, then the power $1/p$ yields $J_0 \leqq k'J^*$ or $k \leqq k'$, establishing the claim for $p < \infty$. For $p = \infty$,

$$\begin{aligned} J_0 &= \operatorname{ess\,sup} \|\alpha_0 - q(\omega)\| \\ &= \operatorname{ess\,sup} \|\alpha_0 + q(\omega)\| \qquad \text{(by symmetry)} \\ &= \operatorname{ess\,sup} \max(\|\alpha_0 - q(\omega)\|, \|\alpha_0 + q(\omega)\|). \end{aligned}$$

By definition of k' one has for all α in A and ω in Ω,

$$\max(\|\alpha_0 - q(\omega)\|, \|\alpha_0 + q(\omega)\|) \leqq k' \max(\|\alpha - q(\omega)\|, \|\alpha + q(\omega)\|).$$

Taking the essential supremum gives

$$\begin{aligned} J_0 &\leqq k' \operatorname{ess\,sup} \max(\|\alpha - q(\omega)\|, \|\alpha + q(\omega)\|) \\ &= k' \operatorname{ess\,sup} \|\alpha - q(\omega)\| \qquad \text{(by symmetry)} \\ &= k'J(\alpha). \end{aligned}$$

Taking the infimum over α in A yields $J_0 \leqq k'J^*$ hence $k \leqq k'$, thus completing the proof of the lemma.

Since the simplification of the probability space is valid by Lemma 4 for any fixed choice of N it is automatically valid for bounds over any given class of normed spaces.

2.3. Geometric interpretation in $\mathcal{N}$. A natural embedding of N into $\mathcal{N}$ is obtained by assigning to α in N the function (i.e., equivalence class) on Ω almost surely equal to α. This is an isometric isomorphism between N and the subspace U of almost surely constant functions in $\mathcal{N}$.

Let M be the linear operator on $\mathcal{N}$ which maps q into the constant function with value $\bar{q} = E\{q(\omega)\}$. Then M is a projection operator with range U. The complementary projection $T = I - M$ translates a distribution in N to have zero mean. Let Q be the subspace of $\mathcal{N}$ determined by the condition $E\{q(\omega)\} = 0$, i.e., let Q be the kernel of M.

Then M and T are the canonical projections associated with the direct sum decomposition

$$\mathcal{N} = U + Q \tag{29}$$

so that $MT = TM = 0$, $M + T = I$, U = range M = kernel T, Q = range T = kernel M.

Since the criterion K as well as the classes of constraint sets A to be considered are translation invariant, one may, without loss of generality, restrict attention to the zero mean case, that is, consider only q in subspace Q. Now if u is the image in U of $\alpha \in N$, under the natural embedding, then $J(\alpha) = |||u - q|||$.

On the other hand, the distance in N of α to the mean is $\|\alpha\| = |||u|||$. Hence (20) becomes: u_0 is suboptimal in a subset A of subspace U if and only if for $u_0 \in A$

$$|||u_0||| \leqq |||v||| \quad \text{for all } v \in A. \tag{30}$$

Then

$$J_0 = |||q - u_0||| \tag{31}$$

while

$$J^* = \inf_{v \in A} |||q - v|||. \tag{32}$$

The fact that the mean, according to Lemma 2, is always a representative element can be expressed by the inequality (for all $q \in Q$ and for all $u \in U$)

$$|||u||| \leqq |||q + u||| \tag{33}$$

which is equivalent to, for all $x \in \mathcal{N}$, $|||Mx||| \leqq |||x|||$, that is, to

$$|||M||| = 1. \tag{34}$$

In the very special case where the parallelogram law holds in $\mathcal{N}$, relations (33) or (34) imply orthogonality of U to Q and therefore $|||T||| = 1$ as well.

In the symmetric case, with $\mathcal{N}$ the space of two-point symmetric distributions in N, as per Lemma 4, one has (for all $u \in U$ and all $q \in Q$)

$$|||u + q||| = |||u - q|||$$

from which

$$|||q||| = \tfrac{1}{2}|||(q + u) + (q - u)||| \leqq \tfrac{1}{2}(|||q + u||| + |||q - u|||) = |||q + u||| \tag{35}$$

so that $|||T||| = 1$ in that case also.

However, in general, from $|||M||| = 1$ and the triangle inequality (for operators) one only obtains

$$1 \leqq |||T||| \leqq 2, \tag{36}$$

and these bounds are sharp.

Finally since T and M are bounded, the subspaces U and Q are closed in $\mathcal{N}$.

2.4. Geometric expression for the sharp bounds. For fixed $(\Omega, \mathscr{F}, P, N, p)$ the sharp bound on J_0/J^* for all sets A in N and all random vectors in $\mathscr{N}$ can be expressed by combining the translation to zero of the mean, the geometric interpretation of § 2.3 and the expression (25) of the bound for fixed q. The result is

$$k = \sup_{q \in Q} \sup_{v \in U} \sup_{u \in S_v} \frac{|||q + u|||}{|||q + v|||}, \tag{37}$$

where

$$S_v = \{u \in U | \, |||u||| \leqq |||v|||\}. \tag{38}$$

When only convex sets A are considered one obtains likewise, from (20), an expression

$$k_c = \sup_{q \in Q} \sup_{v \in U} \sup_{u \in C_v} \frac{|||q + u|||}{|||q + v|||}, \tag{39}$$

where

$$C_v = \{u \in U | \text{ for all } \theta \in [0, 1], \, |||u||| \leqq |||\theta u + (1 - \theta)v|||\}. \tag{40}$$

2.5. Sharp bounds on classes. The most useful bounds for applications are those that hold under general, readily verifiable conditions and make the fullest use of these conditions. Therefore one seeks sharp bounds on classes of cases.

Let $\mathscr{P}$ denote a nonempty class of probability spaces, let Σ denote a nonempty class of normed spaces and let $1 \leqq p \leqq \infty$.

Then $k(\mathscr{P}, \Sigma, p)$ will denote the smallest number k such that $J_0 \leqq kJ^*$ for all cases where $N \in \Sigma$, $(\Omega, \mathscr{F}, P) \in \mathscr{P}$, $q \in \mathscr{N}(\Omega, \mathscr{F}, P, N, p)$, $A \subset N$ and J is defined by the pth order mean.

Similarly $k_c(\mathscr{P}, \Sigma, p)$ will denote the number defined as above but with the additional condition that only convex sets A are considered.

By inclusion and Theorem 2, one has

$$1 \leqq k_c(\mathscr{P}, \Sigma, p) \leqq k(\mathscr{P}, \Sigma, p) \leqq 3. \tag{41}$$

A class Σ will be called quadratic when each of its members satisfies the parallelogram law.

The two most important classes of probability spaces are:

$\mathscr{P}^\infty$: the class of all probability spaces,

and

$\mathscr{P}_s$: the class having as its only member the two-point symmetric probability space of Lemma 4.

Other classes of interest are those containing only probability spaces of finite cardinality. Among these is

$\mathscr{P}^n$: the class of all probability spaces with card $\Omega = n$, $\mathscr{F} = 2^\Omega$.

The most important classes of real normed spaces are:

Σ^∞: the class of all normed spaces,

Σ_p^∞: the (quadratic) class of all normed spaces in which the parallelogram law holds (the pre-Hilbert spaces),

Σ^d, Σ_p^d: the classes defined like Σ^∞, respectively Σ_p^∞, except that only spaces of dimension not exceeding the positive integer d are included.

3. Duality theory.

3.1. Simple duality. Assumptions concerning completeness, reflexivity and separability are unwise and unnecessary for most of the results on inequalities. For this reason a weaker form of duality than usual should be used here.

DEFINITION 5. Two real normed spaces N, N^* with a bilinear product $\langle \cdot, \cdot \rangle : N^* \times N \to R$ are said to be in *simple duality* if and only if the two following relations hold:

$$\|x\| = \sup_{x^* \in N^*} \frac{\langle x^*, x \rangle}{\|x^*\|} \quad \text{for all } x \in N, \tag{42}$$

$$\|x^*\| = \sup_{x \in N} \frac{\langle x^*, x \rangle}{\|x^*\|} \quad \text{for all } x^* \in N. \tag{43}$$

In other words, N^* is (isometrically isomorphic under the natural embedding to) a norm-determining subspace of the norm-dual of N and vice versa. A fortiori, each of the spaces "is" a total set of linear functionals on the other; that is, for x in N

$$(\text{for all } x^* \in N^*, \langle x^*, x \rangle = 0) \Leftrightarrow x = 0, \tag{44}$$

and for x^* in N^*

$$(\text{for all } x \in N, \langle x^*, x \rangle = 0) \Leftrightarrow x^* = 0. \tag{45}$$

When N is finite-dimensional (a "Minkowski space") then Definition 5 implies that N^* "is" the norm-dual of N and vice versa, but otherwise the natural embedding of N^* into the norm-dual of N need not even have an everywhere dense range.

It is crucial for the sequel that the simple duality between N and N^* transfers itself automatically to $\mathcal{N}$, $\mathcal{N}^*$ when the latter are constructed with a common probability space and with conjugate exponents. Because the present set-up is slightly weaker than the usual one it may not be redundant to give the proof in extenso.

LEMMA 5. *Let N, N^* be in simply duality, $(\Omega, \mathcal{F}, P)$ a probability space, $1 \leqq p$, $p^* \leqq \infty$ with $p^{-1} + p^{*-1} = 1$. Then $\mathcal{N}(\Omega, \mathcal{F}, P, N, p) = \mathcal{N}$ and $\mathcal{N}(\Omega, \mathcal{F}, P, N^*, p^*) = \mathcal{N}^*$ are in simple duality with the bilinear product defined, for $x^* \in \mathcal{N}^*$, $x \in \mathcal{N}$, by*

$$\langle\langle x^*, x \rangle\rangle = E\{\langle x^*(\omega), x(\omega) \rangle\}. \tag{46}$$

Proof. By symmetry only one of the two relations in Definition 5 need be proved, say

$$|||x||| = \sup_{x^* \in \mathcal{N}^*} \frac{\langle\langle x^*, x \rangle\rangle}{|||x^*|||}. \tag{47}$$

This holds for $x = 0$ so that one may assume $x \neq 0$ below.

By the simple duality of N and N^* one has pointwise

$$\langle x^*(\omega), x(\omega) \rangle \leqq \|x^*(\omega)\| \cdot \|x(\omega)\|. \tag{48}$$

Integrating this relation gives

$$\begin{aligned} \langle\!\langle x^*, x \rangle\!\rangle &= E\{\langle x^*(\omega), x(\omega)\rangle\} \\ &\leqq E\{\|x^*(\omega)\| \cdot \|x(\omega)\|\} \\ &\leqq |\!|\!| x^* |\!|\!| \cdot |\!|\!| x |\!|\!| , \end{aligned} \tag{49}$$

where the last step holds by Hölder's inequality for real functions.

Thus (47) holds with the $\geqq$ sign. The reverse inequality will first be proved for $x = s$, a simple function taking values s_i in N on sets Ω_i with probabilities w_i, $i = 1, \cdots, n$.

By the simple duality of N and N^* there exists, for all $\varepsilon > 0$, and $s_i \in N$, a ray in N^* such that for any s_i^* on that ray

$$\langle s_i^*, s_i \rangle \geqq (1 - \varepsilon)\|s_i^*\| \cdot \|s_i\|,$$

where the length $\|s_i^*\|$ can be chosen as desired.

Let $s^* \in \mathcal{N}^*$ be the simple function taking the values s_i^* on the set Ω_i, $i = 1, \cdots, n$; then

$$\begin{aligned} \langle\!\langle s^*, s \rangle\!\rangle &= \sum_{i=1}^{n} w_i \langle s_i^*, s_i \rangle \\ &\geqq (1 - \varepsilon) \sum_i w_i \|s_i^*\| \cdot \|s_i\| . \end{aligned}$$

For $p = 1$, choose $\|s_i^*\| = 1$, giving $|\!|\!| s^* |\!|\!| = 1$ and

$$\begin{aligned} \langle\!\langle s^*, s \rangle\!\rangle &\geqq (1 - \varepsilon) \sum_i w_i \|s_i\| \\ &= (1 - \varepsilon) |\!|\!| s^* |\!|\!| \cdot |\!|\!| s |\!|\!| , \end{aligned}$$

proving the assertion.

For $1 < p < \infty$, choose $\|s_i^*\| = \|s_i\|^{p/p^*}$; then

$$\begin{aligned} \sum_i w_i \|s_i\| \cdot \|s_i^*\| &= \left(\sum_i w_i \|s_i\|^p \right)^{1/p} \left(\sum_i w_i \|s_i^*\|^{p^*} \right)^{1/p^*} \\ &= |\!|\!| s^* |\!|\!| \cdot |\!|\!| s |\!|\!| \end{aligned}$$

from which

$$\langle\!\langle s^*, s \rangle\!\rangle \geqq (1 - \varepsilon) |\!|\!| s^* |\!|\!| \cdot |\!|\!| s |\!|\!|$$

as before.

For $p = \infty$ let j be one of the indices for which the maximum in $|\!|\!| s |\!|\!| = \max_{w_i > 0} \|s_i\|$ is attained, and choose $\|s_i^*\| = \delta_{ij}$. Then $|\!|\!| s^* |\!|\!| = \sum_i w_i \|s_i^*\| = w_j$ and

$$\begin{aligned} \langle\!\langle s^*, s \rangle\!\rangle &\geqq (1 - \varepsilon) \sum_i w_i \|s_i^*\| \cdot \|s_i\| \\ &= (1 - \varepsilon) w_j \|s_j\| \\ &= (1 - \varepsilon) |\!|\!| s^* |\!|\!| \cdot |\!|\!| s |\!|\!| , \end{aligned}$$

completing the duality proof for simple functions.

Now for $1 \leqq p < \infty$ the simple functions are norm-dense in $\mathscr{N}$, and for $x \neq 0$, $\varepsilon > 0$, one can choose s such that $|||s - x||| \leqq \varepsilon|||x|||$. Then $|||s||| \geqq |||x||| - |||s - x||| \geqq (1 - \varepsilon)|||x|||$ and constructing s^* as above, one has $s^* \neq 0$ and

$$\begin{aligned}\langle\!\langle s^*, x\rangle\!\rangle &= \langle\!\langle s^*, s\rangle\!\rangle + \langle\!\langle s^*, x - s\rangle\!\rangle \\ &\geqq \langle\!\langle s^*, s\rangle\!\rangle - |\langle\!\langle s^*, x - s\rangle\!\rangle| \\ &\geqq (1 - \varepsilon)|||s^*||| \cdot |||s||| - |||s^*||| \cdot |||s - x||| \\ &\geqq (1 - 3\varepsilon + \varepsilon^2)|||s^*||| \cdot |||x|||,\end{aligned}$$

proving that equality holds in (47).

Finally for $p = \infty$, $x \in \mathscr{N}$, $\varepsilon > 0$ the probability that

$$(1 - \varepsilon)|||x||| \leqq \|x(\omega)\| \leqq |||x|||$$

is positive by definition of $|||x||| = \operatorname{ess\,sup} \|x(\omega)\|$.

Since x is Bochner integrable it is almost separably-valued. Then the intersection of the above shell with the range of a version of x is separable so that, for all $\varepsilon > 0$, it can be covered by a countable collection of spheres of radius ε centered at a separating set. By countable additivity at least one of these spheres has positive probability.

Hence there exists $a \in N$ such that $(1 - \varepsilon)|||x||| \leqq \|a\| \leqq |||x|||$ and such that the set $F \subset \Omega$ on which $\|x(\omega) - a\| \leqq \varepsilon|||x|||$ has positive probability $P\{F\} = w$.

Choose $b \neq 0$ in N^* such that $\langle b, a\rangle \geqq (1 - \varepsilon)\|b\| \cdot \|a\|$ from which $\langle b, a\rangle \geqq (1 - \varepsilon)^2|||x||| \cdot \|b\|$. Let $x^*(\omega) = b$ for ω in F, 0 otherwise. Then $x^* \neq 0, x^* \in \mathscr{N}^*$ and

$$\begin{aligned}\langle\!\langle x^*, x\rangle\!\rangle &= E\{\langle x^*(\omega), x(\omega)\rangle\} \\ &= \int_F P(d\omega)\langle b, x(\omega)\rangle.\end{aligned}$$

However, on F

$$\begin{aligned}\langle b, x(\omega)\rangle &= \langle b, a\rangle + \langle b, x(\omega) - a\rangle \\ &\geqq (1 - \varepsilon)^2|||x||| \cdot \|b\| - \|x(\omega) - a\| \cdot \|b\| \\ &\geqq (1 - 3\varepsilon + \varepsilon^2)|||x||| \cdot \|b\|\end{aligned}$$

so that

$$\begin{aligned}\langle\!\langle x^*, x\rangle\!\rangle &\geqq (1 - 3\varepsilon + \varepsilon^2)|||x||| \cdot \|b\| \cdot w \\ &= (1 - 3\varepsilon + \varepsilon^2)|||x||| \cdot |||x^*|||,\end{aligned}$$

establishing (47) in the case $p = \infty$ and completing the proof of Lemma 5.

The definitions of the projections M, T and subspaces U, Q in $\mathscr{N}$ as per § 2.3, when applied in $\mathscr{N}^*$, yield entities denoted by M^*, T^*, U^*, Q^*. That is, M^* is the mean value projection, $T^* = I - M^*$; U^* consists of the almost constant functions Q^* of the functions of zero mean.

The notation is justified by the following lemma.

LEMMA 6. *M^* is the operator adjoint to M; T^* is the adjoint of T; U^* is the annihilator of Q; Q^* is the annihilator of U and vice versa.*

Proof. First, by Lemma 5, $\mathcal{N}^*$ is norm determining and, a fortiori, total on $\mathcal{N}$ and vice versa. Hence the adjoint of an operator is uniquely defined.

Since Bochner integration commutes with bounded linear operations:

$$\begin{aligned} \langle\!\langle M^*x^*, x\rangle\!\rangle &= E\{\langle \overline{x^*}, x(\omega)\rangle\} \\ &= \langle \overline{x^*}, \bar{x}\rangle \\ &= E\{\langle x^*(\omega), \bar{x}\rangle\} \\ &= \langle\!\langle x^*, Mx\rangle\!\rangle, \end{aligned} \tag{50}$$

and of course

$$\langle\!\langle T^*x^*, x\rangle\!\rangle = \langle\!\langle x^*, x\rangle\!\rangle - \langle\!\langle M^*x^*, x\rangle\!\rangle = \langle\!\langle x^*, x\rangle\!\rangle - \langle\!\langle x^*, Mx\rangle\!\rangle = \langle\!\langle x^*, Tx\rangle\!\rangle, \tag{51}$$

and for $u^* \in U^* = \text{range } M^*$, $q \in Q = \text{kernel } M$,

$$\langle\!\langle u^*, q\rangle\!\rangle = \langle\!\langle M^*u^*, q\rangle\!\rangle = \langle\!\langle u^*, Mq\rangle\!\rangle = \langle\!\langle u^*, 0\rangle\!\rangle = 0, \tag{52}$$

and likewise $\langle\!\langle q^*, u\rangle\!\rangle = 0$.

3.2. Duality lemmas for support functions. Now for u in U, u^* in U^*, the dual product $\langle\!\langle u^*, u\rangle\!\rangle$ is just $\langle \alpha^*, \alpha\rangle$, where α, α^* are the preimages of u, μ^* in N, N^* under the natural embedding.

More precisely, we have the following lemma.

LEMMA 7. *When $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is restricted to $U^* \times U$ it establishes a simple duality between these subspaces.*

Indeed

$$|||u||| = \|\alpha\| = \sup_{\alpha^* \in N^*} \frac{\langle \alpha^*, \alpha\rangle}{\|\alpha^*\|} = \sup_{u^* \in U^*} \frac{\langle\!\langle u^*, u\rangle\!\rangle}{|||u^*|||} \tag{53}$$

and vice versa. For this reason one can use the same symbols for elements of N, N^* and U, U^* without inconsistency. From Lemmas 5 and 7 follows the next obvious but important lemma.

LEMMA 8. *For $v \in U$, $v^* \in U^*$ and S_v, S_{v^*} defined as in* (38), *one has*

$$\sup_{u \in S_v} \langle\!\langle v^*, u\rangle\!\rangle = |||v^*||| \cdot |||v||| = \sup_{u^* \in S_{v^*}} \langle\!\langle u^*, v\rangle\!\rangle.$$

With the convexity assumption it is necessary to consider the sets C_v and C_{v^*} defined as in (40).

LEMMA 9. *Suppose the parallelogram law holds in N; hence also in N^*. Then*

$$\begin{aligned} \sup_{u \in C_v} \langle\!\langle v^*, u\rangle\!\rangle &= \sup_{u^* \in C_{v^*}} \langle\!\langle u^*, v\rangle\!\rangle \\ &= \tfrac{1}{2}\langle\!\langle v^*, v\rangle\!\rangle + \tfrac{1}{2}|||v^*||| \cdot |||v|||. \end{aligned}$$

Proof. By Lemma 7 one need only establish it for the preimages in N, N^* of the vectors and sets involved.

In N, C_v is the sphere

$$C_v = \left\{ u \,\middle|\, \left\| u - \frac{v}{2} \right\| \leqq \frac{1}{2}\|v\| \right\} \tag{54}$$

by virtue of (28). That is,

$$C_v = \left\{ \frac{1}{2}v + \frac{\|v\|}{2}n \,\middle|\, \|n\| \leqq 1 \right\}.$$

For v^* in N^* the support function of C_v is

$$\begin{aligned} \sup_{u \in C_v} \langle v^*, u \rangle &= \sup_{\|n\| \leqq 1} \left\langle v^*, \frac{1}{2}v + \frac{\|v\|}{2}n \right\rangle \\ &= \frac{1}{2}\langle v^*, v \rangle + \frac{\|v\|}{2} \sup_{\|n\| \leqq 1} \langle v^*, n \rangle \qquad (55) \\ &= \frac{1}{2}\langle v^*, v \rangle + \frac{1}{2}\|v^*\| \cdot \|v\|, \end{aligned}$$

and the symmetry of this expression establishes Lemma 9.

When the parallelogram law is not assumed, the convexity assumption becomes more difficult to exploit. For the first time it will be useful then that N^* consist of *all* bounded linear functionals on N, i.e., be the norm-dual of N.

In the case where N is finite-dimensional this follows from the simple duality. But in infinite dimension the assumption is made in the next lemma to enable the use of the Hahn–Banach theorem.

LEMMA 10. *Assume N^* contains all bounded linear functionals on N. Then for all $v \in U$, $v^* \in U^*$ one has*

$$\sup_{u \in C_v} \langle\langle v^*, u \rangle\rangle \leqq \sup_{u^* \in C_{v^*}} \langle\langle u^*, v \rangle\rangle.$$

Proof. One need only consider the preimages in N, N^*. For any $v \in N$, $v^* \in N^*$, $u \in C_v$ either $\langle v^*, u \rangle \leqq 0$ and then for $u^* = 0 \in C_{v^*}$ one has $\langle u^*, v \rangle \geqq \langle v^*, u \rangle$ or else $\langle v^*, u \rangle > 0$, implying $u \neq 0$.

Then the ball $\{x \in N | \|x\| < \|u\|\}$ is disjoint from the segment $\{\theta u + (1 - \theta)v | 0 \leqq \theta \leqq 1\}$, and by the Hahn–Banach theorem there exists a nonzero linear functional which separates them.

Such a functional can be taken to have unit norm and, by assumption, is represented in N^*. Hence there exists $n^* \in N^*$ such that

(i) $\|n^*\| = 1$,

(ii) $\|x\| < \|u\| \Rightarrow \langle n^*, x \rangle \leqq \langle n^*, u \rangle$,

which entail $\langle n^*, u \rangle = \|u\|$ and

(iii) $\langle n^*, v \rangle \geqq \langle n^*, u \rangle$.

Let

$$u^* = \frac{\langle v^*, u \rangle}{\|u\|} n^*.$$

First u^* belongs to C_{v^*} because, for all θ,

$$\begin{aligned} \|\theta u^* + (1 - \theta)v^*\| &\geqq \langle \theta u^* + (1 - \theta)v^*, u/\|u\| \rangle \\ &= \frac{1}{\|u\|}\left(\theta \left\langle \frac{\langle v^*, u \rangle}{\|u\|} n^*, u \right\rangle + (1 - \theta)\langle v^*, u \rangle \right) \end{aligned}$$

$$= \langle v^*, u\rangle/\|u\| = \|u^*\|.$$

Second,

$$\begin{aligned}\langle u^*, v\rangle &= \frac{\langle v^*, u\rangle}{\|u\|}\langle n^*, v\rangle \\ &\geqq \frac{\langle v^*, u\rangle}{\|u\|}\langle n^*, u\rangle \\ &= \langle v^*, u\rangle,\end{aligned}$$

which shows that

$$\sup_{u^*\in C_{v^*}} \langle u^*, v\rangle \geqq \sup_{u\in C_v} \langle v^*, u\rangle, \tag{56}$$

establishing Lemma 10.

COROLLARY. *If N, N^* are complete reflexive, i.e., mutually norm-dual Banach spaces, then the conclusion of Lemma* 10 *holds with equality.*

3.3. Elemental duality theorems. Given a probability space $(\Omega, \mathscr{F}, P)$, an exponent $p\in[1, \infty]$, a pair of normed spaces in simple duality N, N^* define $\mathscr{N} = U + Q$ as in § 2 and $\mathscr{N}^* = U^* + Q^*$. Then without the convexity assumption, the bounds k, k^* are defined by (37) and its dual. These bounds are equal.

THEOREM 4. *For $\mathscr{N} = U + Q$ and $\mathscr{N}^* = U^* + Q^*$ constructed above, the three following expressions are numerically equal*:

(i) $$k \equiv \sup_{q\in Q}\sup_{v\in U}\sup_{u\in S_v}\frac{|||q+u|||}{|||q+v|||},$$

(ii) $$k^* \equiv \sup_{q^*\in Q^*}\sup_{v^*\in U^*}\sup_{u^*\in S_{v^*}}\frac{|||q^*+u^*|||}{|||q^*+v^*|||},$$

(iii) $$\sup_{q\in Q}\sup_{q^*\in Q^*}\sup_{v\in U}\sup_{v^*\in U^*}\frac{\langle\langle q^*, q\rangle\rangle + |||v^*|||\cdot|||v|||}{|||q^*+v^*|||\cdot|||q+v|||}.$$

Proof. By virtue of the symmetry of expression (iii) one need only show that it equals k.

By Lemma 5, $\mathscr{N}$ and $\mathscr{N}^*$ are in simple duality; hence

$$\begin{aligned}|||q+u||| &= \sup_{x^*\in\mathscr{N}^*}\frac{\langle\langle x^*, q+u\rangle\rangle}{|||x^*|||} \\ &= \sup_{q^*\in Q^*}\sup_{v^*\in U^*}\frac{\langle\langle q^*+v^*, q+u\rangle\rangle}{|||q^*+v^*|||} \qquad (57) \\ &= \sup_{q^*\in Q^*}\sup_{v^*\in U^*}\frac{\langle\langle q^*, q\rangle\rangle + \langle\langle v^*, u\rangle\rangle}{|||q^*+v^*|||}.\end{aligned}$$

Thus,

$$k = \sup_{q\in Q}\sup_{v\in U}\sup_{u\in S_v}\sup_{q^*\in Q^*}\sup_{v^*\in U^*}\frac{\langle\langle q^*, q\rangle\rangle + \langle\langle v^*, u\rangle\rangle}{|||q^* + v^*|||\cdot|||q + v|||}$$

$$= \sup_{q\in Q}\sup_{v\in U}\sup_{q^*\in Q^*}\sup_{v^*\in U^*}\frac{\langle\langle q^*, q\rangle\rangle + \sup\{\langle\langle v^*, u\rangle\rangle | u\in S_v\}}{|||q^* + v^*|||\cdot|||q + v|||},$$

and by Lemma 8 this reduces to expression (iii), proving Theorem 4.

When the bound is over convex sets only, then (39) must be used. Two cases are distinguished according to whether the parallelogram law does or does not hold in N.

THEOREM 5. *Assume the parallelogram law holds in N, N^* and let $\mathcal{N}$, $\mathcal{N}^*$ be constructed as in Theorem 4. Then the following three expressions are equal*:

(i) $$k \equiv \sup_{q\in Q}\sup_{v\in U}\sup_{u\in C_v}\frac{|||q + u|||}{|||q + v|||},$$

(ii) $$k^* = \sup_{q^*\in Q^*}\sup_{v^*\in U^*}\sup_{u^*\in C_{v^*}}\frac{|||q^* + u^*|||}{|||q^* + v^*|||},$$

(iii) $$\sup_{q\in Q}\sup_{q^*\in Q^*}\sup_{v\in U}\sup_{v^*\in U^*}\frac{\langle\langle q^*, q^*\rangle\rangle + \langle\langle v^*, v\rangle\rangle/2 + |||v^*|||\cdot|||v|||/2}{|||q^* + v^*|||\cdot|||q + v|||}.$$

Proof. The only change as compared to the proof of Theorem 4 is that Lemma 9 is invoked instead of Lemma 8.

THEOREM 6. *Suppose N^* is the norm-dual of N and let k, k^* be the same expressions as in Theorem 5. Then $k \leqq k^*$.*

Proof. Using expressions for k, k^* derived as in Theorem 4, one has

$$\begin{aligned} k &= \sup_{q\in Q}\sup_{q^*\in Q^*}\sup_{v\in U}\sup_{v^*\in U^*}\frac{\langle\langle q^*, q\rangle\rangle + \sup\{\langle\langle v^*, u\rangle\rangle | u\in C_v\}}{|||q^* + v^*|||\cdot|||q + v|||} \\ &\leqq \sup_{q\in Q}\sup_{q^*\in Q^*}\sup_{v\in U}\sup_{v^*\in U^*}\frac{\langle\langle q^*, q\rangle\rangle + \sup\{\langle\langle u^*, v\rangle\rangle | u^*\in C_{v^*}\}}{|||q^* + v^*|||\cdot|||q + v|||} \\ &= k^*, \end{aligned} \tag{58}$$

where the inequality is due to Lemma 10.

The inequality of Theorem 6 will be enough to obtain equality for bounds holding on classes of normed spaces, as in the sequel.

3.4. Duality for bounds on classes. Since bounds on classes are the most important, the duality of such bounds is of greatest interest. It requires some form of duality between the classes of normed spaces involved.

This motivates the next two definitions.

DEFINITION 6. Two *classes* of real normed spaces are said to be *in simple duality* if for each $N\in\Sigma$ there exist an $N^*\in\Sigma^*$ and a bilinear function $N\times N^*\to R$ such that N, N^* are in simple duality, and conversely for each $N^*\in\Sigma^*$ there exist an $N\in\Sigma$ and a bilinear function such that the same holds.

DEFINITION 7. Two *classes* of real normed spaces are said to be *in full duality* if for each space N in Σ there exists in Σ^* a space isometrically isomorphic to the norm-dual of N, and conversely.

Note that the members of Σ and Σ^* need not all be complete and none need be reflexive. Indeed it is sufficient that whenever $N \in \Sigma$ each even order iterated norm-dual of N be represented (modulo isometric isomorphism) in Σ and each odd order iterated norm-dual be represented in Σ^* and conversely for each space in Σ^*.

The duality theorems for classes can now be readily obtained. The exponent conjugate to p is denoted by p^*.

THEOREM 7. *If the classes* Σ *and* Σ^* *are in simple duality then*

$$k(\mathscr{P}, \Sigma, p) = k(\mathscr{P}, \Sigma^*, p^*).$$

Proof. For each $(\Omega, \mathscr{F}, P) \in \mathscr{P}$ and $N \in \Sigma$ there is, by Definition 6, an N^* in Σ^* for which Theorem 4 applies. Hence, any value of J_0/J^* that can be approached with $\mathscr{P}$, Σ and p can be approached or exceeded with $\mathscr{P}$, Σ^* and p^*. This shows that $k(\mathscr{P}, \Sigma, p) \leqq k(\mathscr{P}, \Sigma^*, p^*)$ and a symmetric argument yields the reverse inequality, establishing the theorem.

THEOREM 8. *If the classes* Σ *and* Σ^* *are quadratic and in simple duality then*

$$k_c(\mathscr{P}, \Sigma, p) = k_c(\mathscr{P}, \Sigma^*, p^*).$$

Proof. The proof proceeds as for Theorem 7 with appeal to Theorem 5 instead of 4.

THEOREM 9. *If the classes* Σ *and* Σ^* *are in full duality then*

$$k_c(\mathscr{P}, \Sigma, p) = k_c(\mathscr{P}, \Sigma^*, p^*).$$

Proof. For each $(\Omega, \mathscr{F}, P)$ in $\mathscr{P}$ and N in Σ there is, by Definition 7, a representation N^* of the norm-dual of N in the class Σ^*. By Theorem 6 any value of J_0/J^* that can be approached with $\Omega, \mathscr{F}, P, N, p$ and convex sets in N can be approached or exceeded with $\Omega, \mathscr{F}, P, N^*, p^*$ and convex sets in N^*. This establishes $k_c(\mathscr{P}, \Sigma, p) \leqq k_c(\mathscr{P}, \Sigma^*, p^*)$. For the symmetric argument, one needs only the symmetry built into Definition 7; the unsymmetric Theorem 6 is simply applied to the representative in Σ of the norm dual of a space in Σ^*. This yields the reverse inequality and proves the theorem.

Applying these results to the classes considered in § 2.5 leads to the next theorem.

THEOREM 10. *If* $\Sigma = \Sigma^\infty$ *or* Σ_p^∞ *or* Σ^d *or* Σ_p^d, *if* $\mathscr{P}$ *is any class of probability spaces and if* $1 \leqq p, p^* \leqq \infty$ *are conjugate exponents, then*

$$k(\mathscr{P}, \Sigma, p) = k(\mathscr{P}, \Sigma, p^*),$$

$$k_c(\mathscr{P}, \Sigma, p) = k_c(\mathscr{P}, \Sigma, p^*).$$

Proof. Each of the classes Σ listed in the statement is in full duality with itself because the norm-dual of a normed space is a normed space, the norm-dual of a pre-Hilbert space is a Hilbert space, the norm-dual of a space of dimension d has the same dimension. Hence, Theorems 7 and 9 suffice to establish Theorem 10.

Applying Theorem 10 with $\Sigma = \Sigma^\infty$, $\mathscr{P} = \mathscr{P}^\infty$, with $\Sigma = \Sigma^\infty$, $\mathscr{P} = \mathscr{P}_s$, with $\Sigma = \Sigma_p^\infty$, $\mathscr{P} = \mathscr{P}^\infty$, and with $\Sigma = \Sigma_p^\infty$, $\mathscr{P} = \mathscr{P}_s$, and letting $p = 1$, $p^* = \infty$, one obtains eight equalities which provide a complete explanation of the phenomena that were reported earlier [11], [12] and were a motivation for the present study.

4. Properties and values of bounds.

4.1. Auxiliary lemmas. For a specific normed space and probability space the bounds may be very difficult to compute, but for broad classes of spaces their determination is often easier because the extreme situations tend to be quite simple.

Attempts to determine the bounds are facilitated by a number of elementary lemmas.

LEMMA 11. *The values of expressions* (37), (39) *for the bounds are unchanged if one or more of the following simplifications are made*:

(i) *Replace* S_v *by* dS_v, *where*

$$dS_v = \{u \in U | \; |||u||| = |||v|||\}.$$

(ii) *Replace* C_v *by* dC_v, *where*

$$dC_v = \{u \in U | \; |||u||| = \inf_{0 \leqq \theta \leqq 1} |||\theta u + (1 - \theta)v|||\}.$$

(iii) *Remove point* v *from* S_v, C_v, dS_v *or* dC_v.

(iv) *Strengthen* $v \in U$ *by the condition* $|||v||| = 1$.

Proof. For u_0 in $S_v(C_v)$, a straight line through u_0 and the origin intersects $S_v(C_v)$ in an interval whose endpoints u_1, u_2 belong to $dS_v(dC_v)$. Since the ratio in (37) or (39) is convex in u for fixed q and v, the maximum of the ratio for $u = u_1$, u_2 cannot be less than the ratio for $u = u_0$. This establishes (i) and (ii). Since $u = v$ gives a unit ratio this case can be disregarded as in (iii). From $v = 0$ follows $u = 0$; hence, one may assume $v \neq 0$ and by homogeneity $|||v||| = 1$, that is, (iv).

For comparing bounds on different classes an appropriate subordination relation among classes is needed.

DEFINITION 8. The subordination $\Sigma_1 \leqq \Sigma_2$ holds if and only if for any normed space in class Σ_1 there exists an isometrically isomorphic subspace in a space in class Σ_2.

DEFINITION 9. The subordination $\mathscr{P}_1 \leqq \mathscr{P}_2$ holds if and only if for all $(\Omega_1, \mathscr{F}_1, P_1)$ in $\mathscr{P}_1$ there exists $(\Omega_2, \mathscr{F}_2, P_2)$ in $\mathscr{P}_2$ and $f: \Omega_2 \to \Omega_1$ such that for all sets $S \in \mathscr{F}_1$ the set $f^{-1}(S)$ belongs to $\mathscr{F}_2$ and $P_1(S) = P_2(f^{-1}(S))$.

LEMMA 12. *One has*

(i) $\Sigma_1 \subset \Sigma_2 \Rightarrow \Sigma_1 \leqq \Sigma_2$,

(ii) $\mathscr{P}_1 \subset \mathscr{P}_2 \Rightarrow \mathscr{P}_1 \leqq \mathscr{P}_2$,

(iii) $\Sigma_1 \leqq \Sigma_2, \mathscr{P}_1 \leqq \mathscr{P}_2$ *imply, for* $1 \leqq p \leqq \infty$,

$$k(\Sigma_1, \mathscr{P}_1, p) \leqq k(\Sigma_2, \mathscr{P}_2, p)$$

and

$$k_c(\Sigma_1, \mathscr{P}_1, p) \leqq k_c(\Sigma_2, \mathscr{P}_2, p).$$

Proof. (i), (ii) and (iii) for $\mathscr{P}_1 = \mathscr{P}_2$ are obvious. However, (iii) with $\Sigma_1 = \Sigma_2$ holds because for any N in $\Sigma_1 = \Sigma_2$ and any $(\Omega_1, \mathscr{F}_1, P_1)$ in $\mathscr{P}_1$ one can choose

$(\Omega_2, \mathscr{F}_2, P_2)$ in $\mathscr{P}_2$ and f as in Definition 9. Then for any q_1 in $\mathscr{N}(\Omega_1, \mathscr{F}_1, P_1, N, p)$ the function $q_2 = q_1 \circ f$ belongs to $\mathscr{N}(\Omega_2, \mathscr{F}_2, P_2, N, p)$ and generates the same function J on N. Thus any ratio feasible with $\mathscr{P}_1$ is feasible with $\mathscr{P}_2$.

LEMMA 13. *Assume* $\Sigma \leqq \Sigma^n$ *and* $\mathscr{P}^{n+1} \leqq \mathscr{P}$; *then, for* $1 \leqq p \leqq \infty$,

$$k(\mathscr{P}, \Sigma, p) = k(\mathscr{P}^{n+1}, \Sigma, p)$$

and

$$k_c(\mathscr{P}, \Sigma, p) = k_c(\mathscr{P}^{n+1}, \Sigma, p).$$

Proof. Any space $N \in \Sigma$ is of finite dimension $d \leqq n$. For $p = \infty$, $(\Omega, \mathscr{F}, P) \in \mathscr{P}$, $q \in (\Omega, \mathscr{F}, P, N, \infty)$ choose a finite ε-net on a ball in N of radius greater than $|||q|||$. Then quantize q to the nearest net-point to obtain a simple function s which satisfies $|||s - q||| \leqq \varepsilon$. For $1 \leqq p < \infty$ such an s can be found even in the infinite-dimensional case. Thus one need only consider simple functions, and, by translation, only those of zero mean. The set of all probability measures on a fixed finite set in N is convex and compact. Hence the main theorem of Dubins [3] applies; that is, any such measure with zero mean is a convex combination of measures supported on subsets of cardinality at most $d + 1$ and also of zero mean. If a bound holds for class $\mathscr{P}^{d+1}$, it holds for all these measures and as in Lemma 4 for their convex combinations. This establishes $k_{(c)}(\mathscr{P}, \Sigma, p) \leqq k_{(c)}(\mathscr{P}^{n+1}, \Sigma, p)$. Lemma 12 implies the reverse inequality, completing the proof.

For a class Σ of normed spaces, $(\Sigma)^d$ will denote the class of all normed spaces of dimension not exceeding d that are subspaces of spaces in Σ, provided with the induced norms. With this notation one has, within isomorphism, $\Sigma^d = (\Sigma^\infty)^d$, $\Sigma_p^d = (\Sigma_p^\infty)^d$ and always $(\Sigma)^d \leqq \Sigma$.

LEMMA 14. *If* $\mathscr{P} \leqq \mathscr{P}^n$, *then*

$$k_{(c)}(\mathscr{P}, \Sigma, p) = k_{(c)}(\mathscr{P}, (\Sigma)^{n+1}, p). \tag{59}$$

and if in addition Σ *is quadratic, then*

$$k_{(c)}(\mathscr{P}, \Sigma, p) = k_{(c)}(\mathscr{P}, (\Sigma)^n, p). \tag{60}$$

Proof. Any random vector q of zero mean has a range of at most n points and this set of vectors spans a space of dimension not exceeding $n - 1$ because of the zero mean condition. Thus for any choice of u, v all points involved are contained in an $(n + 1)$-dimensional subspace of a space in Σ. This establishes the first equality. If Σ is quadratic then the $(n + 1)$-dimensional space is Euclidean. The range of q and point u are contained in some n-dimensional subspace H. With expression (37), replace v by the point $\bar{v}$ in H obtained by a rotation leaving the span of range q fixed. Then $|||q + \bar{v}||| = |||q + v|||$ and $|||\bar{v}||| = |||v|||$ so that $u \in S_v$ and the ratio is unchanged. With expression (39), replace v by its orthogonal projection $\bar{v}$ upon H. Then $u \in C_v$ implies $u \in C_{\bar{v}}$ and one has $|||q + \bar{v}||| \leqq |||q + v|||$ so that the ratio is not decreased. This shows that in either case one need only consider n-dimensional subspaces, as claimed.

The bounds for infinite-dimensional spaces are limits of bounds on finite-dimensional spaces as can be seen from the previous lemma in conjunction with the following one.

LEMMA 15. *For $1 \leqq p < \infty$ and any class Σ one has*

$$k_{(c)}(\mathscr{P}^{\infty}, \Sigma, p) = \sup_{n>0} k_{(c)}(\mathscr{P}^{n}, \Sigma, p),$$

and if there exists a class Σ^* with which Σ is in full duality then this holds for $p = \infty$ as well. (Simple duality suffices without convexity or with quadratic Σ.)

Proof. The first part of the lemma follows from the fact that the simple functions are norm-dense in any space $\mathscr{N}$ formed with $1 \leqq p < \infty$. The second part follows by duality from the first part with $p = 1$.

Finally Lemma 3 can be restated as follows.

LEMMA 16. *For any class $\mathscr{P}$ and any quadratic class Σ one has*

$$k(\mathscr{P}, \Sigma, 2) = k_c(\mathscr{P}, \Sigma, 2) = 1.$$

4.2. Smooth norms and duality. When the norm in $\mathscr{N}$ is differentiable except at the origin, denote by g_x the linear functional which is the gradient at point x of the norm. One has for all $x \neq 0$ in $\mathscr{N}$

$$|||g_x||| = 1, \tag{61}$$

and, by Euler's theorem on homogeneous functions,

$$\langle\langle g_x, x \rangle\rangle = |||x|||. \tag{62}$$

Now assume that for exponent p and $\mathscr{N} = U + Q$ one has found u, v in U and q in Q to yield a maximum in the expression for the bound. Then, without convexity, one may assume $|||u||| = |||v|||$ by Lemma 11. Thus one has found a maximum of

$$\log |||q + u||| - \log |||q + v|||$$

subject to

$$|||u||| = |||v|||$$

and

$$Mq = Tu = Tv = 0.$$

Assume $\mathscr{N}^*$ is the norm-dual of $\mathscr{N}$. Then the Lagrange multipliers are a real number λ and three vectors α, β, γ in $\mathscr{N}^*$ such that

$$\begin{aligned} &\log |||q + u||| - \log |||q + v||| + \lambda(|||u||| - |||v|||) \\ &+\langle\langle \alpha, Mq \rangle\rangle + \langle\langle \beta, Tu \rangle\rangle + \langle\langle \gamma, Tv \rangle\rangle \end{aligned} \tag{63}$$

is at an unconstrained extremum.

Setting to zero the differentials with respect to q, u, v yields these equations in $\mathscr{N}^*$:

$$\frac{g_{q+u}}{|||q + u|||} - \frac{g_{q+v}}{|||q + v|||} + M^*\alpha = 0, \tag{64}$$

$$\frac{g_{q+u}}{|||q + u|||} + \lambda g_u + T^*\beta = 0, \tag{65}$$

$$\frac{g_{q+v}}{|||q + v|||} + \lambda g_v - T^*\gamma = 0. \tag{66}$$

Note that because u, v belong to U and $|||M||| = 1$, the gradients g_u, g_v belong to U^*. Define

$$u^* = \frac{M^* g_{q+v}}{|||q + v|||} \in U^*, \tag{67}$$

$$v^* = \frac{M^* g_{q+u}}{|||q + u|||} \in U^*, \tag{68}$$

$$q^* = \frac{T^* g_{q+u}}{|||q + u|||} = \frac{T^* g_{q+v}}{|||q + v|||} \in Q^*. \tag{69}$$

Then

$$u^* = -\lambda g_v, \tag{70}$$

$$v^* = -\lambda g_u, \tag{71}$$

$$|||u^*||| = |||v^*||| = |\lambda| \tag{72}$$

and

$$\frac{|||q^* + u^*|||}{|||q^* + v^*|||} = \frac{|||q + u|||}{|||q + v|||} \tag{73}$$

so that q^*, u^*, v^* yield the maximum in $\mathcal{N}^*$.

With the convexity assumption, the condition

$$|||u||| \leqq |||\theta u + (1 - \theta)v||| \quad \text{for all } \theta \in [0, 1]$$

is equivalent, by convexity of the norm, to the requirement that the directional derivative of the norm at u in the direction from u to v be nonnegative. When the gradient exists this is just

$$\langle\langle g_u, v - u \rangle\rangle \geqq 0 \tag{74}$$

and at the maximum equality will hold. Thus one has an unconstrained maximum of the function

$$\begin{aligned} &\log |||q + u||| - \log |||q + v||| + \lambda \langle\langle g_u, v - u \rangle\rangle \\ &\quad + \langle\langle \alpha, Mq \rangle\rangle + \langle\langle \beta, Tu \rangle\rangle + \langle\langle \gamma, Tv \rangle\rangle. \end{aligned} \tag{75}$$

Now assume the norm twice continuously differentiable, except at the origin. Then the Hessian H_x of the norm at point x is a self-adjoint linear mapping of $\mathcal{N}$ into $\mathcal{N}^*$. Because of the homogeneity of the norm one has, for all x,

$$H_x x = 0. \tag{76}$$

Using this Hessian one obtains three equations in $\mathcal{N}^*$:

$$\frac{g_{q+u}}{|||q + u|||} - \frac{g_{q+v}}{|||q + v|||} + M^* \alpha = 0, \tag{77}$$

$$\frac{g_{q+u}}{|||q + u|||} - \lambda g_u + \lambda H_u v + T^* \beta = 0, \tag{78}$$

$$\frac{g_{q+v}}{|||q + v|||} - \lambda g_u - T^* \gamma = 0. \tag{79}$$

Now defining u^*, v^*, q^* by the formulas (67)–(69) as before one still has

$$\frac{|||q^* + u^*|||}{|||q^* + v^*|||} = \frac{|||q + u|||}{|||q + v|||}, \tag{80}$$

but now

$$u^* = -\lambda g_u \tag{81}$$

and

$$v^* = -\lambda g_u + \lambda M^* H_u v. \tag{82}$$

The gradient of the dual norm at g_u is $u/|||u|||$ because $|||u/|||u||| \, ||| = 1$ and $\langle\langle u/|||u|||, g_u\rangle\rangle = 1 = |||g_u|||$.

Hence the gradient at $u^* = -\lambda g_u$ is $\pm u/|||u|||$. Therefore

$$\langle\langle g_{u^*}, v^* - u^*\rangle\rangle = \pm\frac{\lambda}{|||u|||}\langle\langle u, M^* H_u v\rangle\rangle \tag{83}$$

and the dual constraints are satisfied since

$$\langle\langle u, M^* H_u v\rangle\rangle = \langle\langle Mu, H_u v\rangle\rangle = \langle\langle u, H_u v\rangle\rangle = \langle\langle H_u u, v\rangle\rangle = \langle\langle 0, v\rangle\rangle = 0. \tag{84}$$

Note that the proportionality of u^* to g_u can be seen also from the fact that with a smooth norm in N the only possible choice of n^* in Lemma 10 is the gradient of the norm at u.

4.3. Symmetry. With the symmetry assumption one need only consider $\mathscr{P} = \mathscr{P}_s$ by Lemma 4. Since $\mathscr{P}_s \leqq \mathscr{P}^2$ one need only, by Lemma 14, consider classes Σ subordinated to Σ^3 or, in the quadratic case, to Σ_p^2. Since Σ_p^2 is isomorphic to $\{E^1, E^2\}$ and since $k(\mathscr{P}_s, \{E^1\}, p) = 1$ for all p, the only quadratic bounds that can be considered are $k(\mathscr{P}_s, \{E^2\}, p)$ and $k_c(\mathscr{P}_s, \{E^2\}, p)$ which will be computed later.

Without the parallelogram law there are an infinity of possibilities, essentially as many as there are classes of convex bodies in R^3. When the conditions defining such a class are involved, the determination of the bounds could be very difficult. All such bounds are upper bounded by $k(\mathscr{P}_s, \Sigma^3, p)$.

THEOREM 11.

$$k(\mathscr{P}_s, \Sigma^\infty, p) = k(\mathscr{P}_s, \Sigma^3, p) = 2 \qquad \textit{for } 1 \leqq p \leqq \infty.$$

Proof. For $|||u||| \leqq |||v|||$ one has, by (33) and (35),

$$|||q + u||| \leqq |||q||| + |||u||| \leqq |||q||| + |||v||| \leqq 2|||q + v|||$$

which shows that $k(\mathscr{P}_s, \Sigma^\infty, p) \leqq 2$.

To see that $k(\mathscr{P}_s, \Sigma^3, p) \geqq 2$ consider $N = R^3$ with l_∞-norm. Let $q(\omega_1) = (1, 1, 0)$, $q(\omega_2) = (-1, -1, 0)$, $u = (1, -1, 0)$, $v = (0, 0, 1)$. Then $\|u\| = \|v\| = 1$, $\|q(\omega_1) - u\| = \|q(\omega_2) - u\| = 2$ giving $|||q - u||| = 2$, while $\|q(\omega_1) - v\| = \|q(\omega_2) - v\| = 1$ giving $|||q - v||| = 1$ which completes the proof.

4.4. Logarithmic convexity. According to expression (iii) of Theorem 4 the bound k for given spaces $\mathscr{N}$ and $\mathscr{N}^*$, without convexity, can be written, with $x = q + v$, $x^* = q^* + v^*$,

$$k = \sup \langle\langle T^*x^*, Tx\rangle\rangle + |||M^*x^*||| \cdot |||Mx||| \tag{85}$$

subject to $x \in \mathcal{N}$, $|||x||| \leqq 1$, $x^* \in \mathcal{N}^*$, $|||x^*||| \leqq 1$ and using the dual expression for the norm:

$$k = \sup \langle\langle T^*x^*, Tx\rangle\rangle + \langle\langle M^*x^*, y\rangle\rangle\langle\langle y^*, Mx\rangle\rangle \tag{86}$$

subject to

$$x, y \in \mathcal{N}, \quad |||x||| \leqq 1, \quad |||y||| \leqq 1,$$
$$x^*, y^* \in \mathcal{N}^*, \quad |||x^*||| \leqq 1, \quad |||y^*||| \leqq 1.$$

Similarly the expression (iii) of Theorem 5 can be written as

$$k = \sup \langle\langle T^*x^*, Tx\rangle\rangle + \tfrac{1}{2}\langle\langle M^*x^*, Mx\rangle\rangle + \tfrac{1}{2}\langle\langle M^*x^*, y\rangle\rangle\langle\langle y^*, Mx\rangle\rangle \tag{87}$$

subject to the same conditions.

These expressions for the bounds satisfy all but one of the requirements of Lemma VI.10.7 as extended by Exercise VI.11.39 in Dunford and Schwartz [4]. The missing feature is the field of complex rather than real scalars.

Hence, to prove that $\log k$ is convex as a function of p^{-1}, as is the case in all examples known to the writer, one would have to construct an extension of the entire set-up to the complex field and show furthermore that this extension does not increase the bound. This will not be attempted here, but it is conjectured that the Riesz-type convexity holds at least for all quadratic classes.

Note that logarithmic convexity can hold for a class Σ even if it does not hold with some (in principle, any) of the normed spaces in Σ, because the supremum of a set of nonconvex functions can be convex.

4.5. Powers of norms and pseudonorms. The criterion K may at first be expressed by a pseudonorm rather than a norm. This would be due to the presence of "don't care" subspaces (in particular, components) in the underlying linear space. Since the bound depends only on the images of all vectors involved in the normed quotient space, all properties established in this paper are applicable. In fact the results can be stated in a slightly stronger form because for a set to have a convex image in the quotient space it is sufficient but not necessary that the set be convex. Likewise for a probability measure to induce a symmetric measure in the quotient space it is sufficient but not necessary that it be symmetric in the original space.

Another obvious but useful point concerns the case where K is a power of a norm (or pseudonorm). If one denotes by $k(r, p)$ a bound for the mean of order $p \in [1, \infty]$, of the rth power, $r \geqq p^{-1}$, of a norm, then the bounds considered in this paper are those denoted $k(1, p)$ and yield the other bounds by the formula

$$k(r, p) = k^r(1, rp). \tag{88}$$

Of most interest are the cases $k(r, \infty) = k^r(1, \infty)$ and $k(r, 1) = k^r(1, r)$.

4.6. The general one-dimensional bound. The class Σ^1 contains only spaces isomorphic to the real line; the parallelogram law holds; and symmetry yields trivially a bound 1 for all p. There remain two important problems, the first being

the determination of

$$k(p) = k(\mathscr{P}^{\infty}, \Sigma^1, p).$$

By its definition $k(p)$ is the smallest number k such that for all real random variables q with finite pth moment and any real numbers u, v one has

$$|u - \bar{q}| \leqq |v - \bar{q}| \Rightarrow \overline{(|u - q|^p)}^{1/p} \leqq k\overline{(|v - q|^p)}^{1/p}$$

(with ess sup operation when $p = \infty$). By Lemma 16, $k(2) = 1$. By Theorem 10, $k(p) = k(p^*)$ for p^* conjugate to p. It is known [12] that $k(1) = k(\infty) = 3$. Therefore one need only consider $1 < p < 2$. By Lemma 11, one need only consider $v = 1, u = -1$. By Lemma 13 one need only consider distributions taking values $-a$ with probability $b/(a + b)$, and b with probability $a/(a + b)$, for $a, b > 0$. Hence

$$k^p(p) = \sup \{R_p(a, b) | a > 0, b > 0\}, \tag{89}$$

where

$$R_p(a, b) = \frac{b \cdot |1 - a|^p + a(1 + b)^p}{b(1 + a)^p + a \cdot |1 - b|^p}. \tag{90}$$

Now for $0 < a < 1$, $R_p(1/a, b) > R_p(a, b)$, and for $b > 1$, $R_p(a, 1/b) > R_p(a, b)$. Therefore one need only consider $a \geqq 1$, $0 < b \leqq 1$ which resolves the absolute value signs. For $p = 1$ the supremum of R is approached as $a \to \infty$, with $b = 1$. But for $1 < p < 2$ the limit as $a \to \infty$ is 1, for each $b \in (0, 1]$. Also $\partial R/\partial b$ is negative at $b = 1$ and $\partial R/\partial a$ is positive at $a = 1$. Hence the maximum of R is attained for some a, b with $1 < a < \infty$, $0 < b < 1$ and these values must satisfy the necessary conditions $\partial R/\partial a = \partial R/\partial b = 0$ which can be written

$$\begin{aligned} R &\equiv \frac{b(a - 1)^p + a(1 + b)^p}{b(a + 1)^p + a(1 - b)^p} \\ &= \frac{(p - 1)a + 1}{(p - 1)a - 1} \left(\frac{a - 1}{a + 1}\right)^{p-1} \\ &= \frac{1 - (p - 1)b}{1 + (p - 1)b} \left(\frac{1 + b}{1 - b}\right)^{p-1}. \end{aligned} \tag{91}$$

This shows that at the maximum $a > 1/(p - 1)$. Let

$$\alpha = \frac{(a + 1)^{p-1}}{(p - 1)a + 1}, \qquad \beta = \frac{(1 - b)^{p-1}}{1 - (p - 1)b}; \tag{92}$$

then the expression for R can be written, using (91),

$$R = \frac{b(a - 1)[(p - 1)a - 1]R\alpha + a(1 + b)[1 + (p - 1)b]R\beta}{b(a + 1)[(p - 1)a + 1]\alpha + a(1 - b)[1 - (p - 1)b]\beta} \tag{93}$$

from which $\alpha = \beta$. Thus it is necessary that

$$\frac{(1+b)^{p-1}}{1+(p-1)b} = \frac{(a-1)^{p-1}}{(p-1)a-1} \tag{94}$$

and

$$\frac{(1-b)^{p-1}}{1-(p-1)b} = \frac{(a+1)^{p-1}}{(p-1)a+1}. \tag{95}$$

By monotonicity (94), (95) determines a as a function of b on $(0, 1)$; and one must seek the intersections of the curves $f_1 = 0, f_2 = 0$, where

$$\begin{aligned} f_1(a,b) &= (p-1)\log(1+b) - \log(1+(p-1)b) - (p-1)\log(a-1) \\ &\quad + \log((p-1)a-1), \end{aligned} \tag{96}$$

$$\begin{aligned} f_2(a,b) &= (p-1)\log(1-b) - \log(1-(p-1)b) - (p-1)\log(a+1) \\ &\quad + \log((p-1)a+1). \end{aligned} \tag{97}$$

For $1/(p-1) < a < \infty$ and $0 < b < 1$ one has

$$\frac{\partial f_1}{\partial a} > \frac{\partial f_2}{\partial a} > 0, \qquad \frac{\partial f_2}{\partial b} < \frac{\partial f_1}{\partial b} < 0. \tag{98}$$

The curve $f_1 = 0$ is a solution of the differential equation

$$\frac{da}{db} = -\frac{\partial f_1/\partial b}{\partial f_1/\partial a}, \tag{99}$$

and $f_2 = 0$ is a solution of

$$\frac{da}{db} = -\frac{\partial f_2/\partial b}{\partial f_2/\partial a} > -\frac{\partial f_1/\partial b}{\partial f_1/\partial a}. \tag{100}$$

This differential inequality shows that the curves cannot cross more than once. Hence the convergence of Newton's method to a solution of $f_1 = f_2 = 0$ within the bounds guarantees that the value of $k(p)$ is being approached.

It can be verified that if a, b yield the maximum with exponent p then

$$a^* = (p-1)b, \qquad b^* = (p-1)a \tag{101}$$

yield the maximum for $p^* = p/(p-1)$.

Computed values of $k(p)$ are given in Table 1. The convexity of $\log k(p)$ as a function of $1/p$ is apparent in Fig. 1.

As p approaches 1, b approaches 1 very rapidly so that k^p may be approximated by the maximum over a of $R_p(a, 1)$. Furthermore the maximizing a is close to $2/(p-1)$ and therefore

$$k(p) \sim \frac{2}{p+1}\left[\left(1-\frac{p-1}{2}\right)^p + 2(p-1)^{p-1}\right]^{1/p} \tag{102}$$

for p close to 1.

At $p = 2$ the values of a and b are immaterial but the limits of the optimal a, b as p approaches 2 *from below* are solutions of the asymptotic form of (94),

TABLE 1
Values of $k(\mathscr{P}^\infty, \Sigma^1, p)$

$\left\lvert \frac{1}{p} - \frac{1}{2} \right\rvert$	k	(for $1 < p < 2$) a	b
0	1	(4.9252)	(.67421)
.01	1.01684854	5.0253	.68765
.05	1.08730938	5.4633	.74028
.10	1.18347748	6.1156	.80297
.15	1.29092191	6.9306	.86089
.20	1.41287885	7.9896	.91200
.25	1.55377397	9.4388	.95351
.30	1.71978473	11.571	.98227
.35	1.91979017	15.071	.99655
.40	2.16766609	21.985	.99988
.45	2.49075092	42.460	1.00000
.49	2.85895929	203.78	1.00000
.499	2.98096328	2005.9	1.00000
.50	3	∞	1

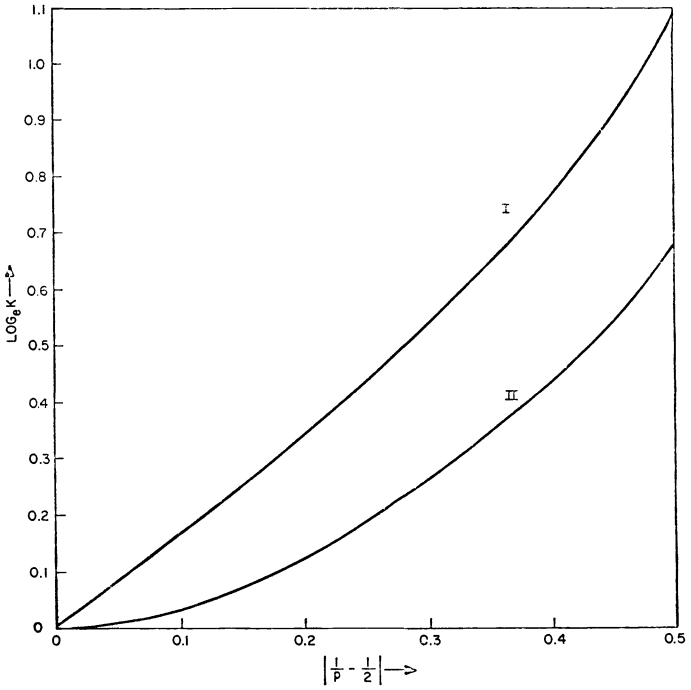


FIG. 1. *Logarithmic convexity for I:* $k(\mathscr{P}^\infty, \Sigma^1, p)$ *and II:* $k_c(\mathscr{P}^\infty, \Sigma^1, p)$

(95) which is

$$\frac{b}{1-b} + \log(1-b) = -\frac{a}{a+1} + \log(a+1), \tag{103}$$

$$\frac{b}{1+b} - \log(1+b) = \frac{a}{a-1} - \log(a-1). \tag{104}$$

The solution $\hat{a}, \hat{b}$ of these equations is given in Table 1. When $p = 2$ is approached *from above* then the optimal a, b tend, by (101), to the limits $a \to \hat{b}$, $b \to \hat{a}$. Thus a swap occurs at $p = 2$, corresponding to the corner in the graph of k.

4.7. One-dimensional bounds with convexity. For the determination of $k(p) = k_c(\mathscr{P}^\infty, \Sigma^1, p)$ one may take $v = 1$ by Lemma 11. Then $C_v = [0, 1]$ so that, again by Lemma 11, one need only consider $u = 0$.

The interpretation of this bound is thus the following: By what factor can the mean of a random variable fail to minimize the pth mean deviation? Astoundingly, it seems that these factors may not have been computed before, though their determination is very simple.

It is known that $k(p^*) = k(p)$ for conjugate exponents (Theorem 10) and that $k(2) = 1$, $k(1) = k(\infty) = 2$ (see [12]). By Lemma 13 one has $k(p) = k_c(\mathscr{P}^2, \Sigma^1, p)$ so that it suffices to consider two-point distributions of zero mean localized at $-a, +b$ $(a, b > 0)$.

Thus one has to find, for $1 < p < 2$,

$$k^p(p) = \sup \{R_p(a, b) | a, b > 0\}$$

with

$$R_p(a, b) = \frac{ba^p + ab^p}{b(a + 1)^p + a|b - 1|^p}. \tag{105}$$

For $b \in (0, 1)$ there exists $\beta > 1$ such that $(1 - b)^p/b = (\beta - 1)^p/\beta$ and $R_p(a, \beta) \geqq R_p(a, b)$; thus one may restrict to $b \geqq 1$, resolving the absolute value sign. The case $b = 1$ can be eliminated by differentiation and the case a and/or $b \to \infty$ yields $R \to 1$. Thus a maximum is reached for finite $a > 0, b > 1$. From the necessary conditions $\partial R/\partial a = \partial R/\partial b = 0$ one can derive

$$\frac{(a + 1)^{p-1}}{a} = \frac{(b - 1)^{p-1}}{b}, \tag{106}$$

$$\frac{a^{p-1}}{(p - 1)a - 1} = \frac{b^{p-1}}{(p - 1)b + 1}. \tag{107}$$

These relations first show that $a > 1/(p - 1)$. Define

$$f_1(a, b) = (p - 1)\log(a + 1) - \log a - (p - 1)\log(b - 1) + \log b, \tag{108}$$

$$f_2(a, b) = (p - 1)\log a - \log((p - 1)a - 1) - (p - 1)\log b + \log((p - 1)b + 1). \tag{109}$$

These functions have the partial derivatives

$$\frac{\partial f_1}{\partial a} = -\frac{(2 - p)a + 1}{a(a + 1)}, \qquad \frac{\partial f_1}{\partial b} = \frac{(2 - p)b - 1}{b(b - 1)}, \tag{110}$$

$$\frac{\partial f_2}{\partial a} = -\frac{(p - 1)[(2 - p)a + 1]}{a[(p - 1)a - 1]}, \qquad \frac{\partial f_2}{\partial b} = \frac{(p - 1)[(2 - p)b - 1]}{b[(p - 1)b + 1]}. \tag{111}$$

These formulas show that the left sides of (106), (107) are strictly monotone, so that $f_1 = 0$ defines $a = \varphi_1(b)$ and $f_2 = 0$ defines $a = \varphi_2(b)$ as single-valued

maps for $b > 1$. As $b \to \infty$ so does a, with

$$\varphi_1(b) \to b\left(\frac{b+1}{b-1}\right)^{(p-1)/(2-p)},$$

$$\varphi_2(b) \to b\left(\frac{1+(p-1)b}{1-(p-1)b}\right)^{1/(2-p)}$$

so that

$$\varphi_1(b) < \varphi_2(b)$$

for sufficiently large b.

However, for $b \in (1/(2-p), \infty)$, equations (110), (111) yield an inequality for the differential equations that φ_1, φ_2 satisfy:

$$-\frac{\partial f_2/\partial b}{\partial f_2/\partial a} < -\frac{\partial f_1/\partial b}{\partial f_1/\partial a}. \tag{112}$$

Hence $\varphi_1(b) < \varphi_2(b)$ holds for $b \geqq 1/(2-p)$. It therefore suffices to seek solutions of $f_1 = f_2 = 0$ in the region $a > 1/(p-1)$, $1 < b < 1/(2-p)$.

In that region (112) holds with the inequality reversed showing that there can be no more than one solution. Thus when Newton's method converges one is assured to have approached the bound. Numerical results are given in Table 2. The logarithmic convexity is apparent in Fig. 1.

TABLE 2
Values of $k_c(\mathscr{P}^\infty, \Sigma^1, p)$

$\left\lvert \frac{1}{p} - \frac{1}{2} \right\rvert$	k	(for $1 < p < 2$) a	b
0	1	∞	∞
.01	1.00035136	127.80	11.602
.05	1.00876825	28.078	2.5485
.10	1.03488172	16.222	1.4696
.15	1.07780874	12.896	1.1581
.20	1.13684798	11.925	1.0440
.25	1.21156503	12.204	1.00743
.30	1.30265981	13.602	1.00043
.35	1.41315708	16.613	1.00000
.40	1.54973722	23.180	1.00000
.45	1.72653600	43.390	1.00000
.49	1.92532002	204.52	1.00000
.499	1.99005234	2006.6	1.00000
.50	2	∞	1

As in the previous case, when a, b yield $k(p)$ then $a^* = (p-1)b$, $b^* = (p-1)a$ yield $k(p^*) = k(p)$.

4.8. Bound with symmetry and quadratic norms. The value of $k(p) = k(\mathscr{P}_s, \Sigma_p^\infty, p)$ has a simple closed form expression. By the previous lemmas $k(p) = k(\mathscr{P}_s, \{E^2\}, p)$ and geometric arguments in the Euclidean plane could be used. Instead, it may be worth giving a completely independent proof of the result.

THEOREM 12. $k(\mathscr{P}_s, \Sigma_p^\infty, p) = 2^{|1/p - 1/2|}$ *for* $1 \leqq p \leqq \infty$.

Proof. For $k > 0$ and $|u| \leqq 1$ the expression

$$(1 + u)^k + (1 - u)^k \tag{113}$$

is even in u, and its derivative

$$k[(1 + u)^{k-1} - (1 - u)^{k-1}]$$

is nonnegative for $u \in [0, 1]$, $k \in [1, \infty)$, nonpositive for $u \in [0, 1]$, $k \in (0, 1]$. Hence, for $u \in [-1, +1]$ the range of expression (113) is $[2, 2^k]$ for $k \geqq 1$ and $[2^k, 2]$ for $k \leqq 1$. It follows that for $k > 0$, $|u| \leqq 1$, $|v| \leqq 1$ one has

$$(1 + u)^k + (1 - u)^k \leqq 2^{|k-1|}[(1 + v)^k + (1 - v)^k]. \tag{114}$$

For a, b, x in an inner product space and $\|a\| \leqq \|b\|$ let

$$u = 2\frac{a \cdot x}{\|a\|^2 + \|x\|^2}, \qquad v = 2\frac{b \cdot x}{\|b\|^2 + \|x\|^2}$$

and $k = p/2$ in (114). Then multiply by the inequality

$$(\|a\|^2 + \|x\|^2)^{p/2} \leqq (\|b\|^2 + \|x\|^2)^{p/2}$$

to obtain

$$\|a + x\|^p + \|a - x\|^p \leqq 2^{|p/2-1|}(\|b + x\|^p + \|b - x\|^p). \tag{115}$$

By averaging with x a symmetrically distributed random vector of zero mean one obtains

$$E\{\|a + x\|^p\} \leqq 2^{|p/2-1|}E\{\|b + x\|^p\}, \tag{116}$$

and the power $1/p$ gives

$$J(a) \leqq 2^{|1/2-1/p|}J(b), \tag{117}$$

whenever $\|a\| \leqq \|b\|$, showing that the bound holds. To see that the bound is sharp let $\|a\| = \|b\| = \|x\| = 1$ and take a parallel (orthogonal) to x and b orthogonal (parallel) to x for $p > 2$ ($p < 2$). For $p = \infty$, a sharp bound of $\sqrt{2}$ was established earlier [11].

4.9. Bounds with symmetry, convexity and quadratic norms. In this section the values of $k(p) = k_c(\mathscr{P}_s, \Sigma_p^\infty, p)$ are investigated. One has $k(2) = 1$, $k(1) = k(\infty) = 2/\sqrt{3}$ (see [11]) and, by Theorem 10, $k(p^*) = k(p)$ for conjugate exponents p^*, p. By Lemma 14, $k(p) = k_c(\mathscr{P}_s, \Sigma_p^2, p)$ and since the one-dimensional bound is trivially 1, $k(p) = k_c(\mathscr{P}_s, \{E^2\}, p)$. Now in E^2 consider v and $u \neq v$, $u \in dC_v$ as per Lemma 11. The values $\pm q$ of the random variable may be assumed to lie on the same side of the line uv by the reflection principle. Now choose the scale so that the orthogonal projections of $q, -q$ on line uv are distant by 2 (coinciding projections would give a ratio of 1). Call a, b the distances of $-q, q$ to line uv. One may assume $a > b \geqq 0$ ($a = b$ would give a ratio of 1). Let z be the oriented distance from u to v.

Then

$$k^p(p) = \sup \frac{(a^2 + 1)^{p/2} + (b^2 + 1)^{p/2}}{(a^2 + (1 + z)^2)^{p/2} + (b^2 + (1 - z)^2)^{p/2}} \tag{118}$$

over all z, and $a > b \geqq 0$. One may first choose z to minimize the denominator of (118). This being a power of a convex function of z one need only look at the sign of the derivative at $z = 0$ which is the sign of

$$(a^2 + 1)^{p/2-1} - (b^2 + 1)^{p/2-1} \tag{119}$$

to conclude that for $p > 2$ the optimal z is closer to the farthest point, i.e., $-1 \leqq z < 0$, while for $1 < p < 2$ it is closer to the nearest point, i.e., $0 < z < 1$. One may therefore restrict attention to these ranges of z.

Assume $a > b > 0$ and z minimizing the denominator with either $1 < p < 2$ or $2 < p < \infty$. Then perform the unilateral variation, $\varepsilon > 0$,

$$b \to b - \varepsilon,$$

$$a \to a + \frac{b(b^2 + 1)^{p/2-1}}{a(a^2 + 1)^{p/2-1}}\varepsilon.$$

Then with z held fixed the first order differential of the numerator of (118) is zero while the denominator has a differential with the sign of

$$[(a^2 + (1 + z)^2)(b^2 + 1)]^{p/2-1} - [(b^2 + (1 - z)^2)(a^2 + 1)]^{p/2-1}, \tag{120}$$

that is, the sign of $z \operatorname{sgn}(p/2 - 1)$ which by the choice of z is negative. Thus b may be assumed zero. Then

$$k^p(p) = \sup_{a,z} \frac{1 + (1 + a^2)^{p/2}}{|1 + z|^p + [(1 - z)^2 + a^2]^{p/2}} = \sup R_p(a, z). \tag{121}$$

For $1 < p < 2$ or $2 < p < \infty$ the extreme cases $x \to \infty$, $z = 0, 1, -1$ can be eliminated. Then a maximum must exist satisfying the necessary conditions obtained by differentiation:

$$a^2 + (1 - z)^2 = [(1 + z)^{p-1}(1 - z)^{-1}]^{2/(p-2)}, \tag{122}$$

$$R_p(a, z) = \left(\frac{1 + a^2}{(1 - z)^2 + a^2}\right)^{p/2-1}. \tag{123}$$

Eliminating a^2 between these equations yields

$$f(z) \equiv z(2 - z) + (1 + z)^{2(p-1)/(p-2)}(1 - z)^{2/(2-p)} - (1 - 2z)^{2/(2-p)} = 0, \tag{124}$$

which shows that $-1 < z < \frac{1}{2}$, and with z determined from (124) a follows by (122).

Now for $1 < p < 2$ equation (124) must have a solution in $(-1, 0)$. The solution $z = 0$ is spurious since $\partial R/\partial z$ is not zero there for any positive a. To see that (124) has no more than one solution it suffices that $f(z)/z$ be monotone on

$(-1, 0)$ or that $zf'(z) - f(z)$ have constant sign. But indeed

$$zf'(z) - f(z)$$
$$= \left(1 - 2\frac{3-p}{2-p}z\right)(1-2z)^{p/(2-p)} + \left((1+z)^2 + 4\frac{p-1}{2-p}\right)\left(\frac{1+z}{1-z}\right)^{p/(2-p)} - z^2$$
$$\geqq (1-2z)^{p/(2-p)} - z^2$$

because $z \in (-1, 0)$. However, in that interval

$$1 - 2z > 1 > z^2$$

and $p \in (1, 2)$ implies $p/(2-p) > 1$; hence $(1-2z)^{p/(2-p)} > 1$ so that $zf'(z) - f(z) > 0$. Thus if Newton's method converges to a solution of $f(z) = 0$ in $(0, 1)$ the bound is being approached.

Computed values are given in Table 3 while logarithmic convexity can be observed in Fig. 2.

TABLE 3
Values of $k_c(\mathscr{P}_s, \Sigma_p^\infty, p)$

$\left\|\frac{1}{p} - \frac{1}{2}\right\|$	k	(for $1 < p < 2$) a^2	$-z$
0	1	(8.1630)	0
.01	1.00004829	8.3687	.02242
.05	1.00121125	9.1114	.11704
.10	1.00489560	10.0506	.24622
.15	1.01120789	10.9550	.38641
.20	1.02041526	11.7417	.53455
.25	1.03290380	12.2798	.68404
.30	1.04916350	12.3862	.82271
.35	1.06969073	11.8629	.93139
.40	1.09466559	10.6554	.98930
.45	1.12337816	9.1678	.99996
.49	1.14825189	8.2041	1.00000
.499	1.15405165	8.0198	1.00000
.50	1.15470054	8	1

If z, a yield the maximum in (121) for exponent p then for the conjugate exponent p^* the maximum is given by

$$z^* = \frac{z}{z-1}, \tag{125}$$

$$\frac{a^*}{z^*} = -\frac{a}{z}. \tag{126}$$

The values z, z^* are in the same relation as the exponents but on the other branch of the hyperbola.

When $x = |1/p - 1/2|$ is close to $\frac{1}{2}$ one has the approximation $k(p) \backsimeq (1 + 3^{-1/(2x)})^x$.

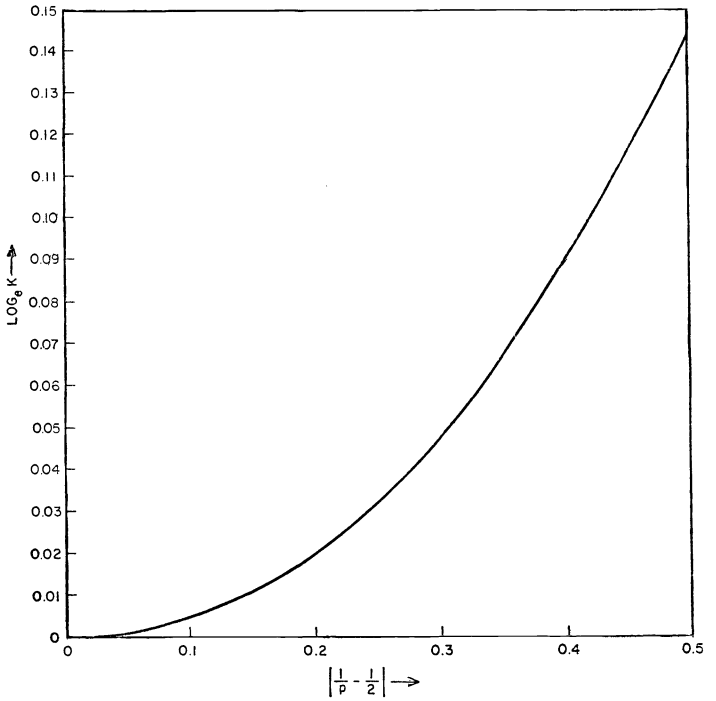


FIG. 2. *Logarithmic convexity for* $k_c(\mathscr{P}_s, \sum_p^\infty, p)$

5. Conclusions.

5.1. Other bounds and their uses. Only a few of the most interesting bounds have been determined above. The others remain to be computed. Furthermore the types of bounds considered can be widened in two directions. First, other classes of sets A, besides all sets and all convex sets, could be considered. Second, iterated evaluators such as the mean of order p_1 of means of order p_2 could be considered. This would require the set-up of a space one level above space $\mathscr{N}$.

One way in which the bounds can be used is the following. Suppose that the determination of the optimum J^* is too difficult, but that, on any grounds whatsoever, a design a_1 appears promising and yields $J_1 = J(a_1)$. Then, if a_0 and $J_0 = J(a_0)$ are easy to determine and a bound k holds, one can assert that $k^{-1}J_0 \leqq J^* \leqq J_1$. Since $J_1 < J_0$ (otherwise a_0 would be preferred to a_1) this may be a sufficiently narrow range to validate a_1 from a practical standpoint.

It is worth stating explicitly that some of the bounds considered in this paper apply to random variables, random vectors, random processes, random fields and random measures. By Theorem 3, the bounds for $p = \infty$ can also be applied to the nonrandom minimax problems.

5.2. Feedback. The results given here are of interest, by Theorem 1, as soon as the zig-zag inequality (7) holds. It would be easy to overestimate the range of applicability on this basis. In fact, as soon as the extensive form of the decision problem has sequential stages with feedback possibilities the zig-zag inequality is unlikely to hold, because a wide range of possible feedback laws is allowed.

In the sequential situation a more detailed analysis must be made and different types of bounds become of interest. Nevertheless the first results in that direction show that there are relations (the "conversion theorem" [12]) between the bounds of the present paper and some of the inequalities for two-stage problems.

5.3. Acknowledgments. The author is very much indebted to C. L. Mallows who took an active part in suggesting and programming early computer tests of what was then a duality conjecture. Many useful comments of his and of V. E. Benes, S. P. Lloyd, L. A. Shepp and A. Tromba were helpful in the development of a systematic approach. The author owes the numerical results given here to the programming work of J. B. Seery, which is gratefully acknowledged.

REFERENCES

[1] S. E. DREYFUS, *Some types of optimal control of stochastic systems*, this Journal, 2 (1964), pp. 120–134.

[2] S. DOSS, *Sur la moyenne d'un élément aléatoire abstrait*, C. R. Acad. Sci. Paris, 226 (1948), pp. 1418–1419.

[3] L. E. DUBINS, *On extreme points of convex sets*, J. Math. Anal. Appl., 5 (1962), pp. 237–244.

[4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. Part I: General Theory*, Interscience, New York, 1958.

[5] M. FRÉCHET, *Sur une nouvelle définition des positions typiques d'un élément aléatoire abstrait*, C. R. Acad. Sci. Paris, 226 (1948), pp. 1419–1420.

[6] O. HÁJEK, *Metric completion simplified*, Amer. Math. Monthly, 75 (1968), pp. 61–63.

[7] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-groups*, Colloquium Publications, American Mathematical Society, Providence, Rhode Island, 1957.

[8] R. D. LUCE AND H. RAIFFA, *Games and Decisions*, John Wiley, New York, 1957.

[9] L. J. SAVAGE, *The theory of statistical decision*, J. Amer. Statist. Assoc., 45 (1950), pp. 238–248.

[10] A. WALD, *Statistical Decision Functions*, John Wiley, New York, 1950.

[11] H. S. WITSENHAUSEN, *Minimax control of uncertain systems*, Rep. ESL-R-269, Massachusetts Institute of Technology, Cambridge, 1966. Also Rep. N66-33441, National Aeronautics and Space Administration, 1966.

[12] ———, *Inequalities for the performance of suboptimal uncertain systems*, Proc. Fourth I.F.A.C. Congress, Warsaw, 1969. Also Automatica, 5 (1969), pp. 507–512.

SIAM J. CONTROL
Vol. 8, No. 1, February 1970

REMARKS ON CONTROLLABILITY OF SECOND ORDER EVOLUTION EQUATIONS IN HILBERT SPACES*

KUNIO TSUJIOKA†

1. Introduction. We consider controllability of a second order evolution equation in a Hilbert space E;

$$\frac{d^2u}{dt^2} = Au(t) + Bf(t), \qquad 0 < t \leqq T, \tag{1}$$

with the initial condition

$$u(0) = \frac{du}{dt}(0) = 0, \tag{2}$$

where A is a self-adjoint operator in E and B is a bounded linear operator on a Hilbert space F to E. A function $f(t)$ belonging to $C^1([0, T]; F)$ is called a control. The function $u(t)$ is defined on $[0, T]$ and takes values in E.

H. O. Fattorini studied the relation between controllability of a first order evolution equation in E;

$$\frac{du}{dt} = Au(t) + Bf(t), \qquad 0 < t \leqq T, \tag{3}$$

with the initial condition

$$u(0) = 0 \tag{4}$$

and that of (1)–(2) for an operator A which is not always self-adjoint (cf. [1]). When (3)–(4) is controllable for some A and for some B, we shall ask for another operator B which makes (1)–(2) controllable at any finite time.

2. Preliminaries. Let E and F be two complex Hilbert spaces and let A be a self-adjoint operator semibounded from above with domain $D(A)$ in E. We denote by $L(X, Y)$ the set of all bounded linear operators on a Hilbert space X into a Hilbert space Y. Let B be an operator in $L(F, E)$. The norm and the scalar product in E are denoted by $\| \cdot \|$ and $(\cdot , \cdot)$ respectively. A control $f(t)$ is a function belonging to $C^1([0, T]; F)$ for some positive T. Since A is semibounded from above, we can find real numbers α and δ, $\delta > 0$, such that $((-A + \alpha)u, u) \geqq \delta \|u\|^2$ for $u \in D(A)$. We denote by $A_\alpha^{1/2}$ the positive square root of the positive operator $A_\alpha = -A + \alpha$. $D(A_\alpha^{1/2})$ becomes a Hilbert space denoted by $H_{1/2}$ with its scalar product defined by $(u, v)_{H_{1/2}} = (A_\alpha^{1/2}u, A_\alpha^{1/2}v)$ for $u, v \in D(A_\alpha^{1/2})$. Putting $u_1 = u$, $u_2 = du/dt$, the second order evolution equation (1) with the initial condition (2) is reduced to the first order equation

$$\frac{d}{dt}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}(t) = \mathfrak{A}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}(t) + \tilde{B}f(t), \tag{5}$$

* Received by the editors May 28, 1968, and in final revised form June 16, 1969.

† Institute of Mathematics and Department of Pure and Applied Sciences, College of General Education, University of Tokyo, Tokyo, Japan.

where

$$\mathfrak{A} = \begin{pmatrix} 0 & I \\ A & 0 \end{pmatrix}, \qquad \tilde{B}f(t) = \begin{pmatrix} 0 \\ Bf(t) \end{pmatrix} \tag{6}$$

with the initial condition

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}(0) = 0. \tag{7}$$

We consider (5) in the Hilbert space $\mathfrak{X} = H_{1/2} \times E$. Let $\mathfrak{A}$ be the operator in $\mathfrak{X}$ with domain $D(\mathfrak{A}) = D(A) \times D(A_\alpha^{1/2})$ such that $\mathfrak{A}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} u_2 \\ Au_1 \end{pmatrix}$ for $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \in D(\mathfrak{A})$. $\tilde{B}$ is the operator in $L(F, \mathfrak{X})$ defined in (6). The operator $\mathfrak{A}$ is the infinitesimal generator of a continuous group in $\mathfrak{X}$ (cf., e.g., [3], [4], [5]). We say that an $\mathfrak{X}$-valued function $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}(t)$ on $[0, T]$ is a *solution of* (5) with a given initial value $\begin{pmatrix} u_{10} \\ u_{20} \end{pmatrix}$ in $D(\mathfrak{A})$ if

(i) $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}(0) = \begin{pmatrix} u_{10} \\ u_{20} \end{pmatrix}$,

(ii) $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}(t) \in D(\mathfrak{A})$ for $0 < t \leqq T$,

(iii) $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}(t)$ belongs to $C^1([0, T]; \mathfrak{X})$ and satisfies (5) for every $t \in (0, T]$.

Since $\mathfrak{A}$ is the infinitesimal generator of the continuous group $e^{t\mathfrak{A}}(-\infty < t < \infty)$, the evolution equation (5) with the initial condition (7) has a unique solution

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}(t) = \int_0^t e^{(t-s)\mathfrak{A}} \tilde{B} f(s)\, ds$$

for any $f(t) \in C^1([0, T]; F)$. Let us return to the second order evolution equation (1) with the initial condition (2). We have a unique solution $u(t)$ of (1)–(2) such that

(i) $u(0) = du(0)/dt = 0$,

(ii) $u(t) \in D(A)$, $du/dt \in D(A_\alpha^{1/2})$, $0 < t \leqq T$,

(iii) $u(t)$ is twice continuously differentiable in E and satisfies (1) for every $t \in (0, T]$.

For any $T > 0$, we define the attainable set $\mathfrak{R}_T$ in $\mathfrak{X}$ by

$$\mathfrak{R}_T = \left\{ \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \int_0^T e^{(T-s)\mathfrak{A}} B f(s)\, ds, f(t) \in C^1([0, T]; F) \right\}.$$

For given A and B, we say that the evolution equation (1) with the initial condition (2) is *completely controllable* (*completely controllable at time* T) if $\overline{\bigcup_{t>0} \mathfrak{R}t} = \mathfrak{X}$ ($\overline{\mathfrak{R}_T} = \mathfrak{X}$). For a given A in E, the evolution equation (1) with the initial condition (2) is called *finitely controllable* (*finitely controllable at time* T) if it is completely controllable (completely controllable at time T) for some finite-dimensional linear space F and for some B in $L(F, E)$ (cf. [2]). For the first order equation (3)

with the initial condition (4) we define the *attainable set* R_T in E by

$$R_T = \left\{ u = \int_0^T e^{(T-s)A} Bf(s)\, ds, f(t) \in C^1([0, T]; F) \right\}.$$

Definitions of *complete controllability* (*complete controllability at time* T) and *finite controllability* (*finite controllability at time* T) for (3)–(4) are given similarly (cf. [2]). We have $\overline{R_T} = \overline{\bigcup_{t>0} R_t}$ for any finite $T > 0$. In fact, $h \in (R_T)^\perp$ (the orthogonal complement of R_T) is equivalent to

$$\left(\int_0^T e^{(T-s)A} Bf(s)\, ds, h \right) = \int_0^T (f(s), B^* e^{(T-s)A} h)\, ds = 0$$

for any $f(t) \in C^1([0, T]; F)$; that is, $B^* e^{tA} h = 0$ for $0 \leqq t \leqq T$, which can be continued analytically to $0 < t < \infty$ since e^{tA} is a holomorphic semigroup. Thus $(R_T)^\perp = (\bigcup_{t>0} R_t)^\perp$ and $\overline{R_T} = \overline{\bigcup_{t>0} R_t}$. Consequently complete controllability of (3)–(4) at some finite time is equivalent to complete controllability. As for $\mathfrak{R}_t$, we have $\overline{\mathfrak{R}_T} \subset \overline{\bigcup_{t>0} \mathfrak{R}_t}$, but the converse inclusion does not hold in general.

If E is a separable Hilbert space, E has an *ordered representation relative to the self-adjoint operator* A (cf. [6]), that is, there exist a positive measure μ defined and finite for a bounded Borel set in $(-\infty, \infty)$ vanishing outside $\sigma(A)$, a decreasing sequence of Borel sets e_n, $n = 1, 2, \cdots$, in $(-\infty, \infty)$ with $\sigma(A) = e_1$ and a unitary operator U on E into $X = \sum_{n=1}^{\infty} L^2(e_n, \mu)$ such that we have $D(UAU^{-1}) = \{f(\lambda) = (f_1(\lambda), \cdots, f_n(\lambda), \cdots) \in L^2(e_n, \mu); \lambda f(\lambda) \in \sum_{n=1}^{\infty} L^2(e_n, \mu)\}$ and that $(UAU^{-1}f)_n(\lambda) = \lambda f_n(\lambda)$ for $f(\lambda) \in D(UAU^{-1})$. If $\mu(e_m) > 0$ and $\mu(e_{m+1}) = 0$, we say that A *has multiplicity* $m(A) = m$. If $\mu(e_n) > 0$ for all n, we say that A *has infinite multiplicity.*

3. Complete controllability at any finite time of second order evolution equations. Applying the result of Fattorini [1] to a self-adjoint operator A we have the following theorem.

THEOREM 1. *Let A be a self-adjoint operator semibounded from above in a Hilbert space E. Then in order that the second order evolution equation* (1) *with the initial condition* (2) *be completely controllable it is necessary that the first order evolution equation* (3) *with the initial condition* (4) *be completely controllable. This condition is also sufficient if the resolvent set $\rho(A)$ of A intersects the negative real axis.*

Remark 1. As we remarked in § 2 complete controllability of the first order case is equivalent to complete controllability at any finite time. But in the second order system complete controllability does not always imply complete controllability at some finite time. When (3)–(4) is completely controllable for some B, we construct in Theorem 2 another operator B which makes (1)–(2) completely controllable at any finite time.

THEOREM 2. *Let A be a self-adjoint operator semibounded from above in E and let C be a bounded linear operator in E such that*:

(a) $C(E) \subset D(A^\infty) = \bigcap_{n=1}^{\infty} D(A^n)$;

(b) *if $g \in E$, $e^{tA} Cg$ can be extended to a function holomorphic in a neighborhood of the origin*;

(c) *C^* is one-to-one*;

(d) *C commutes with A.*

(*All these conditions are satisfied, for example, by* $C = e^{\varepsilon A}$, $\varepsilon > 0$.) *Let the first order evolution equation* (3) *with the initial condition* (4) *be completely controllable. Then the second order system*

$$\frac{d^2u}{dt^2} = Au + CBf \tag{8}$$

with the initial condition (2) *is completely controllable at any finite time* $T > 0$.

Remark 2. Note that $\rho(A)$ need not intersect the negative real axis in Theorem 2.

LEMMA 1. *If* $h \in E$, *then*

$$e^{tA}Ch = \sum_{n=0}^{\infty} \frac{t^n}{n!} A^n Ch. \tag{9}$$

Proof. Assumption (b) and the identities

$$\frac{d^n}{dt^n}(e^{tA}Ch)\bigg|_{t=0} = \frac{1}{n!}A^nCh$$

for $n = 0, 1, 2, \cdots$ imply (9).

LEMMA 2. *Let* $\mathfrak{C}$ *be the operator in* $\mathfrak{X}$ *such that* $\mathfrak{C}h = \begin{pmatrix} Ch_1 \\ Ch_1 \end{pmatrix}$ *for* $h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \in \mathfrak{X}$. *Then* $e^{t\mathfrak{U}}\mathfrak{C}h$ *and* $\mathfrak{C}^*e^{t\mathfrak{A}^*}h$ *can be extended to a function holomorphic in a neighborhood of the origin and we have*

$$e^{t\mathfrak{U}}\mathfrak{C}h = \begin{pmatrix} F(t)Ch_1 + G(t)Ch_2 \\ AG(t)Ch_1 + F(t)Ch_2 \end{pmatrix}, \tag{10}$$

$$\mathfrak{C}^*e^{t\mathfrak{A}^*}h = \begin{pmatrix} F(t)C^*h_1 + A_\alpha^{-1}AG(t)C^*h_2 \\ A_\alpha G(t)C^*h_1 + F(t)C^*h_2 \end{pmatrix}, \tag{11}$$

where

$$F(t) = \sum_{n=0}^{\infty} \frac{t^{2n}}{(2n)!} A^n, \tag{12}$$

$$G(t) = \sum_{n=0}^{\infty} \frac{t^{2n+1}}{(2n+1)!} A^n. \tag{13}$$

Proof. Assumption (a) implies that $\mathfrak{C}(\mathfrak{X}) \subset D(\mathfrak{A}^\infty)$ and that $e^{t\mathfrak{A}}\mathfrak{C}h$ is in C^∞. Since (9) converges, (12) and (13) converge at elements of the form Ch_i and the right-hand side of (11) can be extended holomorphically in a neighborhood of the origin. As the derivatives at the origin of both sides of (10) coincide, the equality in (10) holds. Equation (11) is shown easily using (10).

Proof of Theorem 2. If the system (8)–(2) is not completely controllable at some time $T > 0$, then there exists a nonnull $h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$ such that

$$(\widetilde{CB})^*e^{t\mathfrak{A}^*}h = 0, \qquad 0 \leqq t \leqq T,$$

or, equivalently,

$$\tilde{B}^*\mathfrak{C}^*e^{t\mathfrak{A}^*}h = 0, \qquad 0 \leqq t \leqq T.$$

Using (11), (12), (13), we have

$$B^*(G(t)A_\alpha C^*h_1 + F(t)C^*h_2) = 0$$

in a neighborhood of the origin which implies that

$$B^*A_\alpha A^nC^*h_1 = B^*A^nC^*h_2 = 0 \qquad \text{for all } n \geqq 0.$$

$$B^*e^{tA}A_\alpha C^*h_1 = B^*e^{tA}C^*h_2 = 0.$$

Thus $A_\alpha C^*h_1 = C^*h_2 \in (R_T)^\perp = \{0\}$ and $h_1 = h_2 = 0$.

4. Finite controllability of second order evolution equations. On finite controllability of the first order evolution equations, Fattorini proved the following.

THEOREM 3 (Fattorini [2]). *Let A be a self-adjoint operator semibounded from above in a separable Hilbert space E. Then in order that the first order evolution equation* (3) *with the initial condition* (4) *be finitely controllable it is necessary and sufficient that A have finite multiplicity. Moreover if A has finite multiplicity m, we can choose an m-dimensional linear space F and an operator $B \in L(F, E)$ which makes* (3)–(4) *completely controllable and such that* (3)–(4) *is not completely controllable for any F with dimension less than m.*

Remark 3. In [2], Fattorini remarked that the result of Theorem 3 can be extended further to certain normal operators with connected resolvent. We have considered finite controllability of the second order evolution equation (1) in its first order form (5). The operator $\mathfrak{A}$ is normal but it does not always have a connected resolvent and the operator $\tilde{B}$ has a special form given in (6). Therefore we cannot apply Theorem 3 directly. In Theorem 4, we obtain a result analogous to Theorem 3 for second order evolution equations.

THEOREM 4. *Let A be a self-adjoint operator semibounded from above in a separable Hilbert space E. Then in order that the second order evolution equation* (1) *with the initial condition* (2) *be finitely controllable it is necessary and sufficient that A have finite multiplicity. Moreover if A has finite multiplicity m we can choose an m-dimensional linear space and an operator B in $L(F, E)$ which makes* (1)–(2) *completely controllable at any finite time and such that* (1)–(2) *is not completely controllable for any F with dimension less than m.*

Proof of Theorem 4. Let (1)–(2) be completely controllable for F with $\dim F < \infty$ and for $B \in L(F, E)$; then (3)–(4) is completely controllable for F and B by Theorem 1. It follows immediately from Theorem 3 that $m(A)$ is finite. Conversely let $m(A)$ be finite; then we can find finite-dimensional F and $B \in L(F, E)$ which make (3)–(4) completely controllable. If we replace B by $e^{\varepsilon A}B$ in (1), then (1)–(2) is completely controllable at any finite time by Theorem 2. The second statement of Theorem 4 follows from Theorem 3.

5. Applications.

Example 1. D. L. Russell (cf. [7]) considered controllability of the boundary value problem for the one-dimensional wave equation

$$\frac{\partial^2 u(x,t)}{\partial t^2} - \frac{\partial^2 u(x,t)}{\partial x^2} + q(x)u(x,t) = g(x)f(t), \tag{14}$$

$$0 < t < T, \quad 0 < x < l,$$

with the boundary condition

$$a_0 u(0,t) + a_1 u_x(0,t) = b_0 u(l,t) + b_1 u_x(l,t) = 0, \qquad 0 < t \leqq T, \tag{15}$$

where $q(x) \in C[0, T]$, $g(x) \in L^2(0, T)$ and a_i, b_i, $i = 0, 1$, are real constants such that $a_0^2 + a_1^2 \neq 0$, $b_0^2 + b_1^2 \neq 0$. Let A be the differential operator $\partial^2/\partial x^2 - q(x)$ with domain $D(A) = \{u(x) \in \mathscr{E}^2_{L^2(0,l)}$ (cf., e.g., [8]); $u(x)$ satisfies the boundary condition (15) in $E = L^2(0, l)\}$. Then A is a self-adjoint operator semibounded from above in E, and A has a sequence of simple eigenvalues λ_n, $n = 0, 1, 2, \cdots$, strictly decreasing and diverging at $-\infty$. The multiplicity of A is 1. Let φ_n, $n = 0, 1, 2, \cdots$, be eigenfunctions corresponding to eigenvalues λ_n, $n = 1, 2, \cdots$, which form a complete orthonormal basis for $L^2(0, l)$. For $\omega_n = \sqrt{-\lambda_n}$, $n = 1, 2, \cdots$, the following properties hold;

$$\liminf_{n\to\infty} (\omega_{n+1} - \omega_n) = \frac{1}{D}, \qquad \lim_{n\to\infty} \frac{\omega_n}{n} = D, \tag{16}$$

where D is a positive constant (cf., e.g., [9]). Let us take $L^2(0, l)$ as the set of admissible controls $f(t)$. As in § 2 we treat (14)–(15) in its matrical form and we put

$$\mathfrak{R}_T(L^2) = \left\{ \int_0^T e^{(T-s)\mathfrak{A}} \begin{pmatrix} 0 \\ g \end{pmatrix} f(s)\, ds ; f \in L^2(0, T) \right\}.$$

Russell considered the following problem.

Problem 1. For any $\begin{pmatrix} u \\ v \end{pmatrix} \in D(\mathfrak{A}) = D(A) \times D(A_\alpha^{1/2})$, does there exist a control $f(t) \in L^2(0, T)$ such that the corresponding solution of (14)–(15) satisfying the initial condition $u(x, 0) = u$, $\partial u(x, 0)/\partial t = v$ satisfies the final condition $u(x, T) = \partial u(x, T)/\partial t = 0$?

For any $f(t) \in L^2(0, T)$ and any $\begin{pmatrix} u_0 \\ v_0 \end{pmatrix} \in D(\mathfrak{A})$, the solution of (14)–(15) with the initial condition $u(x, 0) = u_0$, $\partial u(x, 0)/\partial t = v_0$ in its matrical first order form is given by

$$\begin{pmatrix} u \\ v \end{pmatrix}(t) = e^{t\mathfrak{A}} \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} + \int_0^t e^{(t-s)\mathfrak{A}} \begin{pmatrix} 0 \\ g \end{pmatrix} f(s)\, ds, \qquad 0 \leqq t \leqq T. \tag{17}$$

Problem 1 is equivalent to Problem 2 given below if (17) represents the solution of (14)–(15) for any $f(t) \in L^2(0, T)$.

Problem 2. For any $\begin{pmatrix} u_0 \\ v_0 \end{pmatrix} \in D(\mathfrak{A})$, does there exist a control $f(t)$ in $L^2(0, T)$

such that

$$e^{T\mathfrak{A}}\begin{pmatrix} u_0 \\ v_0 \end{pmatrix} + \int_0^T e^{(T-s)\mathfrak{A}}\begin{pmatrix} 0 \\ g \end{pmatrix} f(s)\,ds = 0?$$

If we replace $f(s)$ by $-f(-s)$ in Problem 2 and if we take account of the fact that $e^{t\mathfrak{A}}$ is a continuous group, then the equation in Problem 2 becomes $\begin{pmatrix} u_0 \\ v_0 \end{pmatrix} = \int_0^T e^{s\mathfrak{A}}\begin{pmatrix} 0 \\ g \end{pmatrix} f(s)\,ds$. Thus Problem 2 is equivalent to the following problem.

Problem 3. Does the inclusion $D(\mathfrak{A}) \subset \mathfrak{R}_T(L^2)$ hold?

Under the assumption

$$(g, \varphi_n) \neq 0, \qquad \liminf_{n\to\infty} n|(g, \varphi_n)| > 0, \tag{18}$$

he solved Problem 2 or equivalently Problem 3 by reducing the problem to a moment problem in $L^2(0, l)$. The result is as follows (cf. [7]):

(i) If $T < 2\pi D$, then $D(\mathfrak{A}) \subset \mathfrak{R}_T(L^2)$ does not hold.

(ii) If $T > 2\pi D$, then $D(\mathfrak{A}) \subset \mathfrak{R}_T(L^2)$ holds.

(iii) If $T = 2\pi D$, then there are many cases according to the coefficients in (15). If we define the solution of (14)–(15) with an initial data $\begin{pmatrix} u_0 \\ v_0 \end{pmatrix}$ by (17), Problem 1 is equivalent to Problem 2-3. However the function (17) is not always a "strict solution" as was defined in § 2 since it does not necessarily belong to $D(\mathfrak{A})$ under the assumption (18) and the fact that $f(t) \in L^2(0, T)$. It is in general difficult to see whether (17) belongs to $D(\mathfrak{A})$ for $0 \leqq t \leqq T$ unless it is known that $f(t)$ is continuously differentiable or $\begin{pmatrix} 0 \\ g \end{pmatrix} f(t) \in D(\mathfrak{A})$. But the assumption (18) implies that $\begin{pmatrix} 0 \\ g \end{pmatrix}$ never belongs to $D(\mathfrak{A})$ and $f(t)$ may not be continuously differentiable. Taking $C^1[0, T]$ as a set of all controls instead of $L^2(0, T)$, we consider Problem 4 given below in which the solution (17) is a *strict solution* with the approximating final condition.

Problem 4. For any $\begin{pmatrix} u_0 \\ v_0 \end{pmatrix} \in D(\mathfrak{A})$ and $\varepsilon > 0$, does there exist a control $f(t) \in C^1[0, T]$ such that the solution (17) of (14)–(15) satisfies the final condition

$$\left\| e^{T\mathfrak{A}}\begin{pmatrix} u_0 \\ v_0 \end{pmatrix} - \int_0^T e^{(T-s)\mathfrak{A}}\begin{pmatrix} 0 \\ g \end{pmatrix} f(s)\,ds \right\| < \varepsilon?$$

As we reduced Problem 2 to Problem 3, we can reduce Problem 4 to the following problem.

Problem 5. Does the inclusion $D(\mathfrak{A}) \subset \overline{\mathfrak{R}_T}$ hold?

Since $\overline{D(\mathfrak{A})} = \mathfrak{X}$, the above inclusion is equivalent to $\overline{\mathfrak{R}_T} = \mathfrak{X}$ which means complete controllability of the boundary value problem for the one-dimensional wave equation (14)–(15) with the initial condition

$$u(x, 0) = \frac{\partial u}{\partial t}(x, 0) = 0. \tag{19}$$

To apply Theorem 2 to our problem, we prove the following lemma.

LEMMA 3 (see [2]). *The evolution equation*

$$\frac{\partial u(x,t)}{\partial t} = Au(x,t) + g(x)f(t), \qquad u(x,0) = 0, \tag{20}$$

$$0 < t < T, \qquad 0 < x < l,$$

in $L^2(0,l)$ *is completely controllable if and only if* $(g,\varphi_n) \neq 0$ *for* $n = 0,1,2,\cdots$.

Proof. Let $h \in (R_T)^{\perp}$; then we have

$$\left(\sum_{n=0}^{\infty} \int_0^T e^{\lambda_n(T-s)} f(s)(g,\varphi_n)\varphi_n \, ds, h\right) = 0$$

for any $f(t) \in C^1[0,T]$; that is,

$$\sum_{n=0}^{\infty} e^{\lambda_n t}(g,\varphi_n)(\varphi_n,h) = 0 \tag{21}$$

for $t \in [0,T]$. By analytic continuation, (21) holds for $t \in [0,\infty)$. For any $\lambda \neq \lambda_n$, $n = 0,1,2,3,\cdots$, with $\operatorname{Re}\lambda > \mu$,

$$0 = \sum_{n=0}^{\infty} \int_0^{\infty} e^{(\lambda_n - \lambda)t} g_n \overline{h_n} \, dt = \sum_{n=0}^{\infty} \frac{g_n \overline{h_n}}{\lambda_n - \lambda},$$

where $g_n = (g,\varphi_n)$, $h_n = (h,\varphi_n)$. By analyticity we have

$$\sum_{n=0}^{\infty} \frac{g_n \overline{h_n}}{\lambda_n - \lambda} = 0 \quad \text{for } \lambda \neq \lambda_n, \qquad n = 0,1,2,\cdots.$$

Let $\Gamma_n = \{z \in C; |z - \lambda_n| = \varepsilon_n\}$, where ε_n is a positive number such that $\varepsilon_n < \min(\lambda_{n-1} - \lambda_n, \lambda_n - \lambda_{n+1})$. Then we have

$$g_n \overline{h_n} = \frac{1}{2\pi i} \int_{\Gamma_n} \sum_{m=0}^{\infty} \frac{g_m \overline{h_m}}{z - \lambda_m} \, dz = 0.$$

Thus $(R_T)^{\perp} = \{0\}$ is equivalent to $g_n \neq 0$ for $n = 0,1,2,\cdots$.

PROPOSITION 1. *Let* $g(x) = \sum_{n=0}^{\infty} g_n \varphi_n$ *where*

$$g_n \neq 0 \quad \text{and} \quad |g_n| \leqq M e^{\varepsilon \lambda_n} \tag{22}$$

for some $M > 0$ *and* $\varepsilon > 0$, $n = 0,1,2,\cdots$. *Then the initial boundary value problem for* (14), (15), (19) *is completely controllable at any finite time* $T > 0$.

Proof. Consider controllability of the second order evolution equation

$$\frac{\partial^2 u(t)}{\partial t^2} = Au(t) + gf(t), \quad 0 < t < T, \quad u(0) = \frac{\partial u}{\partial t}(0) = 0 \tag{23}$$

in $L^2(0,l)$. If we put $g_{n,\varepsilon} = e^{\varepsilon\lambda_n/2} g_n$, we see that $\sum_{n=0}^{\infty} |g_{n,\varepsilon}|^2$ and $g_{n,\varepsilon} \neq 0$ by (22). It follows from Lemma 3 that the first order evolution equation (20) is completely controllable at time T if $g(x)$ in (20) is given by $g_\varepsilon(x) = \sum_{n=0}^{\infty} g_{n,\varepsilon}\varphi_n$. Thus $g(x) = e^{\varepsilon A/2} g_\varepsilon(x)$ makes (23) completely controllable at any time by Theorem 2.

Remark 4. If $q(x)$ is nonnegative, then $\omega_n = \sqrt{-\lambda_n} \geqq 0$ for $n = 0,1,2,\cdots$

and assumption (22) can be weakened to

$$g_n \neq 0, \qquad |g_n| \leqq Me^{-\varepsilon\omega_n}, \qquad n = 0, 1, 2, \cdots.$$

Proof. First we show that $e^{t\mathfrak{A}}\begin{pmatrix}0\\ g\end{pmatrix}$ is holomorphic in $(-\infty, \infty)$. In fact, we have

$$e^{t\mathfrak{A}}\begin{pmatrix}0\\ g\end{pmatrix} = \begin{pmatrix}I\\ 0\end{pmatrix}\sum_{n=0}^{\infty}\frac{\sin\omega_n t}{\omega_n}(g, \varphi_n)\varphi_n + \begin{pmatrix}0\\ I\end{pmatrix}\sum_{n=0}^{\infty}\cos\omega_n t(g, \varphi_n)\varphi_n$$

which is holomorphic by the assumption of Remark 4. As in the proof of Theorem 2, $\overline{\mathfrak{R}_T} = \overline{\mathfrak{R}_T(L^2)}$ is easily verified. Russell's result (ii) shows that $\overline{\mathfrak{R}_t} = \mathfrak{X}$ for any $t > 0$.

Example 2. Let A be the differential operator $\partial^2/\partial x^2$ in $L^2(-\infty, \infty)$ with domain $D(A) = \mathscr{E}^2_{L^2}(-\infty, \infty)$. As for finite controllability of (1)–(2) in this case we have the following proposition.

PROPOSITION 2. *The initial value problem for the one-dimensional wave equation*

$$\frac{\partial^2 u}{\partial t^2} = Au + \sum_{i=1}^{2} g_{i\varepsilon}(x)f_i(t), \qquad 0 < t < T, \quad -\infty < x < \infty, \tag{24}$$

$$u(x, 0) = u_t(x, 0) = 0$$

is completely controllable at any time T if

$$g_{i\varepsilon}(x) = \mathscr{F}^{-1}e^{-\varepsilon\lambda^2}\mathscr{F}g_i(x) = \frac{1}{2\sqrt{\varepsilon\pi}}\int_{-\infty}^{\infty}\exp\left(-\left(\frac{x-y}{2\varepsilon}\right)^2\right)g_i(y)\,dy, \quad i = 1, 2,$$

where

$$\mathscr{F}g(s) = \hat{g}(s) = (2\pi)^{-1/2}\,\text{l.i.m.}_{N\to\infty}\int_{-N}^{N}e^{isx}g(x)\,dx \quad \text{for } g \in L^2(-\infty, \infty)$$

and $g_1(x)$ is a nonnull function in $L^2(-\infty, \infty)$ with compact support, $g_2(x) = g_1(x - h)$ with $h \neq 0$.

Fattorini [2] proved Lemma 4 and Lemma 5 given below.

LEMMA 4. *The operator A has multiplicity* 2.

LEMMA 5. *The first order evolution equation in $L^2(-\infty, \infty)$*

$$\frac{\partial u}{\partial t} = Au + \sum_{i=1}^{2} g_i(x)f_i(t)$$

with the initial condition

$$u(0) = 0$$

is completely controllable where the g_i, $i = 1, 2$, are given in Proposition 1.

Proof of Proposition 2. The assertion is proved by Theorem 2 and Lemma 5 because $g_{i\varepsilon} = e^{\varepsilon A}g_i$, $i = 1, 2$.

Acknowledgment. The author wishes to express his hearty thanks to the editor for his encouragement and to the referee for his kind advice. Theorem 2 in the present form owes much to the latter.

REFERENCES

[1] H. O. FATTORINI, *Controllability of higher order systems*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 301–311.

[2] ———, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–462.

[3] T. USHIJIMA, *Note on the integration of the wave equation*, Sci. Papers College Gen. Ed. Univ. Tokyo, 17 (1967), pp. 155–159.

[4] K. YOSIDA, *An operator-theoretical integration of the wave equation*, J. Math. Soc. Japan, 8 (1956), pp. 79–92.

[5] ———, *Functional Analysis*, Springer, Berlin, 1965.

[6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. Part II*, Interscience, New York, 1963.

[7] D. L. RUSSELL, *Nonharmonic Fourier series in the theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–560.

[8] L. SCHWARTZ, *Théorie des distributions. I*, Actualités Sci. Indust., no. 1091, 1950, 148 pp.; *II*, no. 1122, 1951, 169 pp.

[9] F. G. TRICOMI, *Differential Equations*, Hafner, New York, 1961.

SIAM J. CONTROL
Vol. 8, No. 1, February 1970

SOLUTION OF LINEAR PURSUIT-EVASION GAMES*

YOSHIYUKI SAKAWA†

1. Introduction. This paper treats pursuit-evasion games which are played by two players and governed by linear differential equations. The pursuit-evasion games are closely related to differential games of which general theory was developed by Isaacs [1], Berkovitz [2], Pontryagin [3], Varaiya [4], and others. If the differential equations describing the evolution of the games are linear, the problems can be treated more simply in a direct way as shown in [5]–[11].

In this paper, particular attention is paid to the problems of the conditions under which the game can be completed and of finding a max-min completion time of the game. Necessary and sufficient conditions for completion of the game are presented. The optimal controls for both players are derived, respectively. Furthermore, an iterative procedure for computing the max-min completion time and the optimal controls of both players are given.

2. Formulation of the problem. Let us consider a pursuit-evasion game described by the linear differential equation:

$$dx/dt = A_1x + B_1u - C_1v, \tag{1}$$

where x is a state vector in m-dimensional Euclidean space R^m, u is an r-dimensional control vector of the first player I, v is an s-dimensional control vector of the second player II, and A_1, B_1 and C_1 are $m \times m$, $m \times r$ and $m \times s$ constant matrices, respectively. Let U and V be bounded and closed subsets of R^r and R^s, respectively. Further let U be convex. It is assumed that at each time $t \geqq 0$, $u(t)$ and $v(t)$ must satisfy the condition

$$u(t) \in U, \qquad v(t) \in V, \qquad t \geqq 0. \tag{2}$$

Let π be an $n \times m$ ($n \leqq m$) matrix corresponding to the orthogonal projection from R^m onto an n-dimensional linear subspace L, i.e.,

$$L = \{\pi x : x \in R^m\}.$$

Further let us define a subset M_ε of R^m by

$$M_\varepsilon = \{x \in R^m : \|\pi x\| \leqq \varepsilon\}.$$

The pursuit-evasion game is said to be completed from an initial point $x(0) = x_0$ if, no matter what measurable control $v(t)$ may be chosen by the second player II such that $v(t) \in V$ for all $t \geqq 0$, the first player I can choose a measurable control $u(t)$ such that $u(t) \in U$ for all $t \geqq 0$ and such that $x(T) \in M_\varepsilon$ for some finite time T, $0 \leqq T < \infty$.

* Received by the editors January 9, 1969, and in revised form June 23, 1969.

† Faculty of Engineering Science, Osaka University, Toyonaka, Osaka, Japan. This research was done while the author was visiting the school of Engineering and Applied Science, University of California, Los Angeles, California.

The rank of the matrix π is assumed to be n. Multiplying (1) by the matrix π from the left yields

$$\pi\, dx/dt = \pi A_1 x + \pi B_1 u - \pi C_1 v. \tag{3}$$

By defining a new n-dimensional vector z in L by

$$\pi x = z, \tag{4}$$

a differential equation describing a motion in the n-dimensional linear subspace L is obtained as follows:

$$dz/dt = Az + Bu - Cv, \tag{5}$$

where A, B and C are respectively $n \times n$, $n \times r$ and $n \times s$ matrices defined by

$$\begin{aligned} A &= \pi A_1 \pi'(\pi\pi')^{-1}, \\ B &= \pi B_1, \qquad C = \pi C_1. \end{aligned} \tag{6}$$

In (6), $'$ denotes the transpose of a matrix. Since rank π = rank $\pi\pi'$ = n, it is clear that the inverse of the $n \times n$ matrix $\pi\pi'$ exists.

The solution of (5) at $t = T$ with initial condition $z_0 = \pi x_0$ is given by

$$\begin{aligned} z(T) &= \Phi(T)z_0 + \int_0^T \Phi(T-t)[Bu(t) - Cv(t)]\, dt \\ &= \Phi(T)z_0 + \int_0^T \Phi(t)[Bu(T-t) - Cv(T-t)]\, dt, \end{aligned} \tag{7}$$

where $\Phi(t)$ is given by

$$\Phi(t) = e^{tA}. \tag{8}$$

Now, since the terminal condition $x(T) \in M_\varepsilon$ of the game is equivalent to the condition

$$z(T) \in S_\varepsilon = \{z \in R^n : \|z\| \leqq \varepsilon\},$$

the problem is to choose a control $u_v(t) \in U$, $t \geqq 0$, according to the opponent's control $v(t) \in V$, $t \geqq 0$, such that

$$\Phi(T_{u,v})z_0 + \int_0^{T_{u,v}} \Phi(T_{u,v} - t)[Bu_v(t) - Cv(t)]\, dt \in S_\varepsilon \tag{9}$$

for some finite time $T_{u,v}$. The subscripts u, v of $T_{u,v}$ denote the dependency of $T_{u,v}$ on the controls $u(t) \in U$ and $v(t) \in V$, $t \geqq 0$. If, no matter what measurable control $v(t) \in V$, $t \geqq 0$, may be chosen by the second player II, the first player can choose a measurable control $u_v(t) \in U$, $t \geqq 0$, such that (9) holds for some finite time $T_{u,v} < \infty$, then the game starting from the initial condition z_0 is said to be completed.

3. Main theorems. Before stating theorems on the completion of the game, it is necessary to introduce an important operation on compact sets. Let U be a

bounded and closed subset of R^r. Then a function $H_U(\eta)$ is defined by

$$H_U(\eta) = \sup_{u \in U} \eta u, \tag{10}$$

where η is an arbitrary r-dimensional row vector. Since the set U is bounded and closed, there is a vector $u(\eta) \in U$ such that

$$H_U(\eta) = \sup_{u \in U} \eta u = \eta u(\eta). \tag{11}$$

PROPOSITION 1. *The function $H_U(\eta)$ defined by* (10) *is continuous with respect to η. Furthermore, if $u(\eta)$ is uniquely determined in some neighborhood of η, then $u(\eta)$ is continuous in the neighborhood.*

Proof. Let Δ be an arbitrary r-dimensional row vector. From the definition of $u(\eta)$, it follows that

$$\begin{aligned} H_U(\eta + \Delta) - H_U(\eta) &= (\eta + \Delta)u(\eta + \Delta) - \eta u(\eta) \\ &\geqq (\eta + \Delta)u(\eta) - \eta u(\eta) = \Delta u(\eta), \end{aligned}$$

and

$$H_U(\eta + \Delta) - H_U(\eta) \leqq (\eta + \Delta)u(\eta + \Delta) - \eta u(\eta + \Delta) = \Delta u(\eta + \Delta).$$

Hence,

$$\Delta u(\eta) \leqq H_U(\eta + \Delta) - H_U(\eta) \leqq \Delta u(\eta + \Delta). \tag{12}$$

Since the set U is bounded, by letting $\Delta \to 0$,

$$\lim_{\Delta \to 0} H_U(\eta + \Delta) = H_U(\eta).$$

This proves the continuity of $H_U(\eta)$.

To prove the continuity of $u(\eta)$, let us assume $\eta_i \to \eta_0$. Since the set U is compact, we may assume that $u(\eta_i)$ converges to $\hat{u} \in U$, say. Then

$$\eta_i u(\eta_i) \geqq \eta_i u(\eta_0).$$

Passing to the limit as $i \to \infty$, we have

$$\eta_0 \hat{u} \geqq \eta_0 u(\eta_0).$$

This shows that

$$u(\eta_i) \to \hat{u} = u(\eta_0),$$

which completes the proof.

Analogously, $H_V(\xi)$ and $v(\xi)$ are defined by

$$H_V(\xi) = \sup_{v \in V} \xi v = \xi v(\xi),$$

where ξ is an arbitrary s-dimensional row vector. For convenience, let us define the $n \times r$ matrix $K(t)$ and the $n \times s$ matrix $L(t)$ by

$$K(t) = \Phi(t)B, \qquad L(t) = \Phi(t)C. \tag{13}$$

It is clear that $K(t)$ and $L(t)$ are analytic. Then (9) is rewritten as

$$\Phi(T_{u,v})z_0 + \int_0^{T_{u,v}} K(t)u_v(T_{u,v} - t)\,dt - \int_0^{T_{u,v}} L(t)v(T_{u,v} - t)\,dt \in S_\varepsilon. \tag{14}$$

Now, by using an analogous technique to [13], the following theorem is obtained.

THEOREM 1. *In order that, for any measurable control* $v(t) \in V$, $t \geqq 0$, *there exist a measurable control* $u_v(t) \in U$, $t \geqq 0$, *such that* (14) *holds for some finite time* $T_{u,v} \geqq 0$, *it is necessary and sufficient that there be a finite time* $T \geqq 0$ *such that*

$$-\varepsilon \leqq \lambda\Phi(T)z_0 + \int_0^T H_U(\lambda K(t))\,dt - \int_0^T H_V(\lambda L(t))\,dt \tag{15}$$

for all n-dimensional unit row vectors λ.

Proof. To prove the necessity, let λ be an arbitrary n-dimensional unit row vector. Then, multiplying the left-hand side of (14) by $-\lambda$ from the left and using the Schwarz inequality yields

$$-\lambda\Phi(T_{u,v})z_0 - \int_0^{T_{u,v}} \lambda K(t)u_v(T_{u,v} - t)\,dt + \int_0^{T_{u,v}} \lambda L(t)v(T_{u,v} - t)\,dt \leqq \varepsilon. \tag{16}$$

Since the inequality (16) must hold for a $v(t) \in V$ such that

$$\lambda L(t)v(T_{u,v} - t) = H_V(\lambda L(t)) = \sup_{v \in V} \lambda L(t)v,$$

$$-\varepsilon \leqq \lambda\Phi(T_{u,v})z_0 + \int_0^{T_{u,v}} \lambda K(t)u_v(T_{u,v} - t)\,dt - \int_0^{T_{u,v}} H_V(\lambda L(t))\,dt. \tag{17}$$

Since

$$\lambda K(t)u_v(T_{u,v} - t) \leqq H_U(\lambda K(t)), \tag{18}$$

by putting $T = T_{u,v}$, (15) is obtained.

To prove the sufficiency, suppose that there is a control $v(t) \in V$ for which there exists no control $u_v(t) \in U$, such that (14) holds for some finite time T. This means that the compact convex set defined by

$$\left\{ \int_0^T K(t)u(T - t)\,dt : u(T - t) \in U \right\}$$

does not intersect the compact sphere

$$-\Phi(T)z_0 + \int_0^T L(t)v(T - t)\,dt + S_\varepsilon.$$

Therefore, there exists an n-dimensional unit row vector λ such that the inequality

$$-\lambda\Phi(T)z_0 + \int_0^T \lambda L(t)v(T - t)\,dt + \lambda a > \int_0^T \lambda K(t)u(T - t)\,dt \tag{19}$$

holds for all $u(t) \in U$, $t \in [0, T]$, and for all $a \in S_\varepsilon$.

Since the inequality (19) must hold for a $u(t) \in U$ such that

$$\lambda K(t)u(T - t) = H_U(\lambda K(t)) = \sup_{u \in U} \lambda K(t)u$$

and for a vector $a = -\varepsilon\lambda' \in S_\varepsilon$, by using the inequality

$$\int_0^T H_V(\lambda L(t))\,dt \geqq \int_0^T \lambda L(t)v(T - t)\,dt,$$

it follows that

$$-\varepsilon > \lambda\Phi(T)z_0 + \int_0^T H_U(\lambda K(t))\,dt - \int_0^T H_V(\lambda L(t))\,dt. \tag{20}$$

This contradicts (15), and the proof is completed.

From Theorem 1, the following theorem is directly obtained.

THEOREM 2. *In order that, for any measurable control* $v(t) \in V$, $t \geqq 0$, *there exist a measurable control* $u_v(t) \in U$, $t \geqq 0$, *such that* (14) *holds for some finite time* $T_{u,v} \geqq 0$, *it is necessary and sufficient that there be a finite time* $T \geqq 0$ *such that*

$$\inf_{\lambda \in Q} \left[\lambda\Phi(T)z_0 + \int_0^T H_U(\lambda K(t))\,dt - \int_0^T H_V(\lambda L(t))\,dt \right] \geqq -\varepsilon, \tag{21}$$

where Q *is a set of n-dimensional unit row vectors.*

Theorem 2 here is a similar result to Theorem 1 in Pshenichniy's paper [6]. However, in this paper further results are obtained on the basis of Theorem 2.

4. Several propositions. The vectors $u \in U$ and $v \in V$ which attain the maximum of $\lambda K(t)u$ and $\lambda L(t)v$, respectively, will be henceforth denoted by $u(t, \lambda)$ and $v(t, \lambda)$; i.e.,

$$H_U(\lambda K(t)) = \sup_{u \in U} \lambda K(t)u = \lambda K(t)u(t, \lambda), \tag{22}$$

$$H_V(\lambda L(t)) = \sup_{v \in V} \lambda L(t)v = \lambda L(t)v(t, \lambda). \tag{23}$$

In what follows, we assume the following condition.

ASSUMPTION. For each fixed $\lambda \in Q$, the controls $u(t, \lambda)$ and $v(t, \lambda)$ are uniquely determined, respectively, for all $t \in [0, T]$ except a finite number of points on the interval $[0, T]$.

By this assumption and Proposition 1, it is clear that the controls $u(t, \lambda)$ and $v(t, \lambda)$ are piecewise continuous on $[0, T]$. Note that if the sets U and V are compact and strictly convex, then $u(t, \lambda)$ and $v(t, \lambda)$ are uniquely determined for all $t \in [0, T]$ and continuous on $[0, T]$ (see [12]), and that if the sets U and conv V, conv V denoting the closed convex hull of V, are compact convex polyhedrons and if $u(t, \lambda)$ and $v(t, \lambda)$ are uniquely defined by the maximum conditions, respectively, for almost all $t \in [0, T]$, then $u(t, \lambda)$ and $v(t, \lambda)$ are piecewise constant on $[0, T]$ (see [14]).

Let us define a scalar function by

$$\begin{aligned} F(T, \lambda; z_0) &= \lambda\Phi(T)z_0 + \int_0^T H_U(\lambda K(t))\,dt - \int_0^T H_V(\lambda L(t))\,dt \\ &= \lambda\Phi(T)z_0 + \lambda \int_0^T K(t)u(t, \lambda)\,dt - \lambda \int_0^T L(t)v(t, \lambda)\,dt. \end{aligned} \tag{24}$$

PROPOSITION 2. *The gradient vector of the function* $F(T, \lambda; z_0)$ *with respect to* λ *is given by*

$$\operatorname{grad}_\lambda F(T, \lambda; z_0) = z(T, \lambda; z_0), \tag{25}$$

where $z(T, \lambda; z_0)$ *is defined by*

$$z(T, \lambda; z_0) = \Phi(T)z_0 + \int_0^T K(t)u(t, \lambda)\, dt - \int_0^T L(t)v(t, \lambda)\, dt. \tag{26}$$

Moreover, $\operatorname{grad}_\lambda F(T, \lambda; z_0)$ *is continuous in* T, λ *and* z_0.

Proof. Let Δ be an arbitrary n-dimensional row vector. From the definition of $u(t, \lambda)$ it follows that

$$\begin{aligned} H_U((\lambda + \Delta)K(t)) - H_U(\lambda K(t)) &\geqq (\lambda + \Delta)K(t)u(t, \lambda) - \lambda K(t)u(t, \lambda) \\ &= \Delta K(t)u(t, \lambda), \\ H_U((\lambda + \Delta)K(t)) - H_U(\lambda K(t)) &\leqq (\lambda + \Delta)K(t)u(t, \lambda + \Delta) - \lambda K(t)u(t, \lambda + \Delta) \\ &= \Delta K(t)u(t, \lambda + \Delta). \end{aligned}$$

Thus, the following inequality is obtained:

$$\Delta K(t)u(t, \lambda) \leqq H_U((\lambda + \Delta)K(t)) - H_U(\lambda K(t)) \leqq \Delta K(t)u(t, \lambda + \Delta). \tag{27}$$

Integration of (27) with respect to t yields

$$\begin{aligned} \Delta \int_0^T K(t)u(t, \lambda)\, dt &\leqq \int_0^T H_U((\lambda + \Delta)K(t))\, dt - \int_0^T H_U(\lambda K(t))\, dt \\ &\leqq \Delta \int_0^T K(t)u(t, \lambda + \Delta)\, dt. \end{aligned} \tag{28}$$

Let $t_1, t_2, \cdots, t_N$ $(0 < t_1 < t_2 < \cdots < t_N < T)$ be the points on $[0, T]$ except where the control $u(t, \lambda)$ is continuous. Let us define subintervals of $[0, T]$ by

$$\begin{aligned} &I_0(\varepsilon) = [0, \varepsilon), \qquad I_{N+1}(\varepsilon) = (T - \varepsilon, T], \\ &I_i(\varepsilon) = (t_i - \varepsilon, t_i + \varepsilon), \qquad i = 1, \cdots, N, \\ &I(\varepsilon) = [0, T] - \bigcup_{i=0}^{N+1} I_i(\varepsilon). \end{aligned} \tag{29}$$

By the continuity argument (Proposition 1), it is clear that for sufficiently small $\varepsilon > 0$ there is a $\delta(\varepsilon) > 0$ such that, if $\|\Delta\| < \delta(\varepsilon)$, then for $t \in I(\varepsilon)$,

$$\|u(t, \lambda + \Delta) - u(t, \lambda)\| < \varepsilon; \tag{30}$$

while, since U is bounded there is a constant $k > 0$ such that

$$\|u(t, \lambda + \Delta) - u(t, \lambda)\| < k \quad \text{if } t \in \bigcup_{i=0}^{N+1} I_i(\varepsilon). \tag{31}$$

Therefore, we obtain

$$\int_0^T \|u(t, \lambda + \Delta) - u(t, \lambda)\|\, dt < \varepsilon T + 2\varepsilon(N + 1)k. \tag{32}$$

Relations (28) and (32) imply that

$$\operatorname{grad}_\lambda \int_0^T \lambda K(t)u(t, \lambda)\, dt = \int_0^T K(t)u(t, \lambda)\, dt. \tag{33}$$

Similarly it can be shown that

$$\text{grad}_\lambda \int_0^T \lambda L(t)v(t,\lambda)\,dt = \int_0^T L(t)v(t,\lambda)\,dt. \tag{34}$$

Hence (25) has been proved. The continuity of $\text{grad}_\lambda F(T,\lambda;z_0)$ is evident from the course of the proof. This completes the proof.

For simplicity, when it is clear that the initial condition is fixed to z_0, one writes $F(T,\lambda;z_0)$ and $z(T,\lambda;z_0)$ simply as $F(T,\lambda)$ and $z(T,\lambda)$, respectively.

Now, since the function $F(T,\lambda)$ given by (24) is continuous in λ and the set $Q = \{\lambda \in R^n : \|\lambda\| = 1\}$ is compact, there is a $\lambda \in Q$ which attains the infimum of $F(T,\lambda)$. Let us denote the λ which attains the infimum of $F(T,\lambda)$ by $\lambda(T)$; i.e.,

$$\inf_{\lambda\in Q} F(T,\lambda) = F(T,\lambda(T)).$$

PROPOSITION 3.

$$\inf_{\lambda\in Q} F(T,\lambda) = F(T,\lambda(T)) = -\|z(T,\lambda(T))\|, \tag{35}$$

where $z(T,\lambda)$ is given by (26).

Proof. Since the minimum of $F(T,\lambda)$, T being fixed, is sought under the condition $\|\lambda\|^2 - 1 = 0$, let us define

$$\bar{F}(T,\lambda,\mu) = F(T,\lambda) + \mu(\|\lambda\|^2 - 1), \tag{36}$$

where μ is a Lagrange multiplier. Put

$$\partial\bar{F}/\partial\lambda_i = z_i(T,\lambda) + 2\mu\lambda_i = 0, \qquad i = 1,\cdots,n, \tag{37}$$

where z_i and λ_i are ith components of the vectors z and λ, respectively. Eliminating the Lagrange multiplier μ from (37) yields

$$\lambda(T) = -z'(T,\lambda(T))/\|z(T,\lambda(T))\|, \tag{38}$$

where $'$ denotes the transpose of a vector. Substituting (38) into (24) yields (35).

PROPOSITION 4. *Let us assume that for any time $T > 0$ and for any $\lambda_1, \lambda_2 \in Q$,*

$$\|z(T,\lambda_1)\| = \|z(T,\lambda_2)\| \quad \textit{implies } \lambda_1 = \lambda_2. \tag{38'}$$

Then it follows that

$$\frac{dF(T,\lambda(T))}{dT} = \lambda(T)A\Phi(T)z_0 + H_U(\lambda(T)K(T)) - H_V(\lambda(T)L(T)). \tag{39}$$

Proof. Let δ be an arbitrary real number. In view of the definition (24) of $F(T,\lambda)$ and the relation

$$\Phi(T+\delta) = \Phi(T) + \int_T^{T+\delta} A\Phi(t)\,dt,$$

it follows that

$$F(T+\delta,\lambda) = F(T,\lambda) + \int_T^{T+\delta} [\lambda A\Phi(t)z_0 + H_U(\lambda K(t)) - H_V(\lambda L(t))]\,dt. \tag{40}$$

When $\lambda = \lambda(T + \delta)$, by using the relation

$$F(T, \lambda(T + \delta)) \geqq F(T, \lambda(T)) = \inf_{\lambda \in Q} F(T, \lambda),$$

it follows from (40) that

$$\begin{aligned} &F(T + \delta, \lambda(T + \delta)) - F(T, \lambda(T)) \\ &\geqq \int_T^{T+\delta} [\lambda(T + \delta)A\Phi(t)z_0 + H_U(\lambda(T + \delta)K(t)) - H_V(\lambda(T + \delta)L(t))]\, dt. \end{aligned} \tag{41}$$

On the other hand, it is clear that

$$F(T + \delta, \lambda(T + \delta)) - F(T, \lambda(T)) \leqq F(T + \delta, \lambda(T)) - F(T, \lambda(T)). \tag{42}$$

Since $F(T, \lambda)$ is continuous in T, the relations (41) and (42) show the continuity of $F(T, \lambda(T))$; i.e.,

$$F(T + \delta, \lambda(T + \delta)) \to F(T, \lambda(T)) \quad \text{if } \delta \to 0. \tag{43}$$

From (35) it is clear that the assumption (38′) implies uniqueness of the $\lambda(T) \in Q$ which attains the infimum of $F(T, \lambda)$ over Q. Therefore, it follows from (43) that $\lim_{\delta \to 0} \lambda(T + \delta)$ has a unique limit $\lambda(T)$; i.e.,

$$\lambda(T + \delta) \to \lambda(T) \quad \text{if } \delta \to 0. \tag{44}$$

If $\delta > 0$, it follows from (41) and (42) that

$$\begin{aligned} &\frac{1}{\delta} \int_T^{T+\delta} [\lambda(T + \delta)A\Phi(t)z_0 + H_U(\lambda(T + \delta)K(t)) - H_V(\lambda(T + \delta)L(t))]\, dt \\ &\leqq [F(T + \delta, \lambda(T + \delta)) - F(T, \lambda(T))]/\delta \\ &\leqq [F(T + \delta, \lambda(T)) - F(T, \lambda(T))]/\delta. \end{aligned} \tag{45}$$

In view of (44) and the continuity of $H_U(\lambda K(t))$ and $H_V(\lambda L(t))$ in λ and t (by Proposition 1), it follows from (45) that

$$\frac{dF(T, \lambda(T))}{dT} = \lambda(T)A\Phi(T)z_0 + H_U(\lambda(T)K(T)) - H_V(\lambda(T)L(T)). \tag{46}$$

In the case where $\delta < 0$, the same result (46) is obtained.

5. Completion of the game. Suppose that $\|z_0\| > \varepsilon$ and there exists a time T $(0 \leqq T < \infty)$ such that

$$\inf_{\lambda \in Q} F(T, \lambda; z_0) = F(T, \lambda(T); z_0) = -\varepsilon. \tag{47}$$

Let T_0 be the smallest nonnegative time satisfying (47). Then the following theorem is obtained.

THEOREM 3. *No matter what measurable control $v(t) \in V$, $t \geqq 0$, may be chosen by the second player* II, *the game can be completed in a time not greater than T_0, where T_0 is the smallest nonnegative time satisfying* (47). *Further, no matter what measurable control $u(t) \in U$, $t \geqq 0$, may be chosen by the first player* I, *there is a control $v(t) \in V$, $t \geqq 0$, of the second player* II *such that the game cannot be completed in a time smaller than T_0.*

Proof. Corresponding to an arbitrary control $v(t) \in V$, $t \geqq 0$, let us define

$$F_v(T, \lambda; z_0) = \lambda\Phi(T)z_0 + \lambda \int_0^T K(t)u(t, \lambda)\, dt - \lambda \int_0^T L(t)v(T - t)\, dt. \tag{48}$$

In view of (23), it is clear that

$$F_v(T, \lambda; z_0) \geqq F(T, \lambda; z_0) \quad \text{for all } \lambda \in Q. \tag{49}$$

Hence,

$$\inf_{\lambda \in Q} F_v(T_0, \lambda; z_0) \geqq \inf_{\lambda \in Q} F(T_0, \lambda; z_0) = -\varepsilon. \tag{50}$$

From Proposition 3, it is obvious that

$$\inf_{\lambda \in Q} F_v(T, \lambda; z_0) = F_v(T, \lambda_T; z_0) = -\|z_v(T, \lambda_T)\|, \tag{51}$$

where

$$z_v(T, \lambda_T) = \Phi(T)z_0 + \int_0^T K(t)u(t, \lambda_T)\, dt - \int_0^T L(t)v(T - t)\, dt, \tag{52}$$

and $\lambda_T \in Q$ attains the infimum of $F_v(T, \lambda; z_0)$ when T and z_0 are fixed.

Since $z_v(T, \lambda_T)$ is continuous in time T, and it holds from (50) and (51) that

$$-\|z_v(T_0, \lambda_{T_0})\| \geqq -\varepsilon, \tag{53}$$

there exists a time T^* such that

$$0 \leqq T^* \leqq T_0, \qquad -\|z_v(T^*, \lambda_{T^*})\| = -\varepsilon. \tag{54}$$

Equation (54) shows that the game can be completed in a time T^* which is not greater than T_0.

In the same way, let us define

$$F_u(T, \lambda; z_0) = \lambda\Phi(T)z_0 + \lambda \int_0^T K(t)u(T - t)\, dt - \lambda \int_0^T L(t)v(t, \lambda)\, dt. \tag{55}$$

In view of (22), it is clear that

$$F_u(T, \lambda; z_0) \leqq F(T, \lambda; z_0) \quad \text{for all } \lambda \in Q.$$

Therefore,

$$\inf_{\lambda \in Q} F_u(T_0, \lambda; z_0) \leqq \inf_{\lambda \in Q} F(T_0, \lambda; z_0) = -\varepsilon. \tag{56}$$

From Proposition 3, it is obvious that

$$\inf_{\lambda \in Q} F_u(T, \lambda; z_0) = F_u(T, \lambda_T; z_0) = -\|z_u(T, \lambda_T)\|, \tag{57}$$

where

$$z_u(T, \lambda_T) = \Phi(T)z_0 + \int_0^T K(t)u(T - t)\, dt - \int_0^T L(t)v(t, \lambda_T)\, dt, \tag{58}$$

and $\lambda_T \in Q$ attains the infimum of $F_u(T, \lambda; z_0)$. From (56) and (57),

$$-\|z_u(T_0, \lambda_{T_0})\| \leqq -\varepsilon. \tag{59}$$

Equation (59) shows that the game cannot be completed in a time smaller than T_0. This completes the proof.

Theorem 3 shows that the controls $u(t) = u(T_0 - t, \lambda(T_0))$ and $v(t) = v(T_0 - t, \lambda(T_0))$, $t \in [0, T_0]$, are respectively optimal for both players, in the sense that the first player I wishes to complete the game as soon as possible, and the second player II wishes to prevent the completion of the game as long as possible. The time T_0 is clearly the smallest max-min completion time of the game.

Now, a natural question may occur. Under what initial condition does a finite time T exist satisfying (47)? The following theorems give sufficient conditions for the existence of the finite time T satisfying (47).

THEOREM 4. *If the homogeneous differential equation*

$$dz/dt = Az \tag{60}$$

is asymptotically stable and it holds that

$$BU \supset CV, \tag{61}$$

where BU and CV are subsets of R^n defined by

$$BU = \{Bu : u \in U\}, \qquad CV = \{Cv : v \in V\}, \tag{62}$$

then the game can be completed, no matter what the initial condition $z_0 \in R^n$ may be.

Proof. Since $CV \subset BU$, whatever control $v(t) \in V$, $t \geqq 0$, may be chosen by the second player II, it is possible for the first player I to choose a control $u(t) \in U$, $t \geqq 0$, such that

$$Bu(t) = Cv(t) \qquad \text{for all } t \geqq 0. \tag{63}$$

Since the system described by (60) is asymptotically stable, there exists a finite time T such that

$$\|z(T)\| \leqq \varepsilon. \tag{64}$$

This completes the proof of the theorem.

Another sufficient condition for the completion of the game is obtained by using (39). Since

$$\Phi(t) = e^{tA} = I + tA + \tfrac{1}{2}t^2A^2 + \cdots,$$

it is clear that

$$A\Phi(t) = \Phi(t)A. \tag{65}$$

Using (65), equation (39) can be rewritten as

$$\frac{dF(T, \lambda(T))}{dT} = \lambda(T)\Phi(T)Az_0 + \max_{\hat{u} \in BU} \lambda(T)\Phi(T)\hat{u} - \max_{\hat{v} \in CV} \lambda(T)\Phi(T)\hat{v}. \tag{66}$$

THEOREM 5. *Let us assume that for any time $T > 0$ and for any $\lambda_1, \lambda_2 \in Q$,*

$$\|z(T, \lambda_1)\| = \|z(T, \lambda_2)\| \quad \textit{implies } \lambda_1 = \lambda_2. \tag{67}$$

If there exists a $\delta > 0$ such that

$$-Az_0 + CV + S_\delta \subset BU, \tag{68}$$

$$\|\lambda(T)\Phi(T)\| \geqq \delta \quad \textit{for all } T > 0, \tag{69}$$

where S_δ is a closed sphere in R^n of radius δ about the origin, then the game starting from z_0 can be completed.

Proof. Let μ be an arbitrary n-dimensional nonzero row vector such that $\|\mu\| \geqq \delta > 0$. Then it is clear that

$$\max_{x \in S_\delta} \mu x = \mu x(\mu) \geqq \delta^2,$$

where $x(\mu)$ is a point of S_δ at which the maximum is attained. From (68), for arbitrary $\hat{v} \in CV$ and $x \in S_\delta$, there is a $\hat{u} \in BU$ such that

$$-Az_0 + \hat{v} + x = \hat{u}.$$

Hence, for all $\hat{v} \in CV$ and for all μ satisfying $\|\mu\| \geqq \delta$, there is a $\hat{u} \in BU$ such that

$$\mu(\hat{u} - \hat{v} + Az_0) \geqq \delta^2 > 0. \tag{70}$$

Inequality (70) still holds for such a $\hat{v}(\mu)$ that

$$\mu\hat{v}(\mu) = \max_{\hat{v} \in CV} \mu\hat{v},$$

and it holds that

$$\mu\hat{u} \leqq \mu\hat{u}(\mu) = \max_{\hat{u} \in BU} \mu\hat{u}.$$

Hence, for all μ satisfying $\|\mu\| \geqq \delta$,

$$\max_{\hat{u} \in BU} \mu\hat{u} - \max_{\hat{v} \in CV} \mu\hat{v} + \mu Az_0 \geqq \delta^2. \tag{71}$$

Under the assumption (67), it is clear from Proposition 4 that (66) holds. By putting $\mu = \lambda(T)\Phi(T)$, it is evident from (66) and (71) that

$$dF(T, \lambda(T))/dT \geqq \delta^2 > 0 \quad \text{for all } T > 0. \tag{72}$$

Since $F(0, \lambda(0)) = -\|z_0\| < -\varepsilon < 0$, it is clear that the game which starts from z_0 can be completed, if z_0 satisfies (68).

6. Iterative procedures for determining optimal controls. Let us assume that the game starting from z_0 can be completed. The minimum time T_0 and the vector $\lambda(T_0)$ satisfying (47) can be computed as follows:

1. Put $\lambda_1 = -z_0'/\|z_0\|$. Compute $F(T, \lambda_1)$, $T \geqq 0$, λ_1 being fixed, up to the time T_1, where $F(T_1, \lambda_1) = -\varepsilon$. Clearly, $T_1 \leqq T_0$.

2. Let $F(T_i, \lambda_i) = -\varepsilon$, $i = 1, 2, \cdots$. Minimize $F(T_i, \lambda)$ with respect to λ by using the gradient method, T_i being fixed. By Proposition 2, the gradient of $F(T_i, \lambda)$ can be computed easily. Put

$$\min_{\lambda \in Q} F(T_i, \lambda) = F(T_i, \lambda_{i+1}) \leqq -\varepsilon.$$

3. Compute $F(T, \lambda_{i+1})$, $T \geqq T_i$, λ_{i+1} being fixed, up to the time T_{i+1}, where $F(T_{i+1}, \lambda_{i+1}) = -\varepsilon$. It is clear that

$$F(T, \lambda_{i+1}) \geqq F(T, \lambda(T)) \quad \text{for all } T \in [0, T_{i+1}]. \tag{73}$$

4. Iterate procedure 2 and procedure 3.

Since $T_i \leqq T_{i+1} \leqq T_0$, $\lim T_i$ exists which will be denoted by T_0'. Clearly

$T'_0 \leqq T_0$. It holds that

$$F(T_{i+1}, \lambda_{i+1}) = F(T_{i+1}, \lambda(T_i)) = -\varepsilon, \tag{74}$$

and by (43), $F(T, \lambda(T))$ is continuous in T, hence

$$F(T'_0, \lambda(T'_0)) = -\varepsilon. \tag{75}$$

Since T_0 is the smallest nonnegative time satisfying (47), it follows that $T'_0 = T_0$, and $\lambda_{i+1} = \lambda(T_i) \to \lambda(T_0 - 0)$, if $\lambda(T_0)$ is not unique and $\lambda(T_0 - 0) \neq \lambda(T_0 + 0)$. It is clear from Theorem 3 that the controls $u(t) = u(T_0 - t, \lambda(T_0 - 0))$ and $v(t) = v(T_0 - t, \lambda(T_0 - 0))$, $t \in [0, T_0]$, are optimal for both players, respectively.

7. Concluding remarks. As we have seen in the preceding section, $\lambda(T_0 - 0)$ can be determined for each initial condition z_0. Hence, the optimal controls have been synthesized in the form of $u(t) = u(T_0 - t, z_0)$ and $v(t) = v(T_0 - t, z_0)$, respectively. In an actual game, however, it is desirable to synthesize the controls as a function of the current state of the game. In this sense, $u(T_0, z_0)$ and $v(T_0, z_0)$ are optimal feedback controls at $t = 0$. Thus, we can synthesize the optimal controls at $t = t$ in the form of $u(T_0 - t, z(t))$ and $v(T_0 - t, z(t))$, respectively.

In view of Proposition 3, the λ which minimizes $F(T, \lambda; z_0)$ under $\lambda \in Q$ depends on z_0. Thus, let us define $\lambda(T, z)$ by

$$\inf_{\lambda \in Q} F(T, \lambda; z) = F(T, \lambda(T, z); z). \tag{76}$$

Further let us define $u^0(T, z)$ by

$$u^0(T, z) = u(T, \lambda(T, z)), \tag{77}$$

where $u(T, \lambda(T, z))$ satisfies

$$\max_{u \in U} \lambda(T, z)K(T)u = \lambda(T, z)K(T)u(T, \lambda(T, z)). \tag{78}$$

Pshenichniy [15] proved, under several conditions, that if the first player employs the control $u^0(T, z)$ thus obtained, then the game governed by

$$dz(t)/dt = Az(t) + Bu^0(T_0 - t, z(t)) - Cv(t), \qquad z(0) = z_0, \tag{79}$$

can be completed in a time not greater than T_0, where T_0 is the smallest nonnegative time satisfying (47). However, Pshenichniy did not show how to compute $\lambda(T, z)$, T_0, and so on. These computations can be done by using the algorithm shown in this paper. Therefore, the results obtained in this paper will be useful for synthesizing the optimal open-loop controls, as well as the feedback controls.

REFERENCES

[1] R. Isaacs, *Differential Games*, John Wiley, New York, 1965.

[2] L. D. Berkovitz, *Necessary conditions for optimal strategies in a class of differential games and control problems*, this Journal, 5 (1967), pp. 1–24.

[3] L. S. Pontryagin, *On the theory of differential games*, Uspekhi Mat. Nauk, 21 (1966), pp. 219–274.

[4] P. P. Varaiya, *On the existence of solutions to a differential game*, this Journal, 5 (1967), pp. 153–162.

[5] L. S. Pontryagin, *Linear differential games*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 330–334.

[6] B. N. Pshenichniy, *Linear differential games*, Ibid., pp. 335–341.

[7] Y. SAKAWA, *On linear differential games*, Ibid., pp. 373–385.
[8] F. M. KIRILLOVA, *The application of functional analysis to problems of pursuit*, Ibid., pp. 386–397.
[9] L. S. PONTRYAGIN, *Linear differential games* 1, Soviet Math. Dokl., 8 (1967), no. 3, pp. 769–771.
[10] ———, *Linear differential games* 2, Ibid., 8 (1967), no. 4, pp. 910–912.
[11] E. SHIMEMURA AND T. MATSUMOTO, *On the solvability of linear differential games*, to appear.
[12] E. H. EGGLESTON, *Convexity*, Cambridge University Press, Cambridge, 1958.
[13] H. A. ANTOSIEWICZ, *Linear control systems*, Arch. Rational Mech. Anal., 12 (1963), pp. 313–324.
[14] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
[15] B. N. PSHENICHNIY, *Linear differential games*, Avtomat. i Telemekh., (1968), no. 1, pp. 65–78.

SIAM J. CONTROL
Vol. 8, No. 1, February 1970

ON CONTROLLABILITY OF NONLINEAR SYSTEMS*

V. A. CHEPRASOV†

We consider the problem of steering a continuous mechanical system to a position of equilibrium, when the system motion is described by a vector differential equation

$$\dot{z} = \varphi(z) + Q(t). \tag{1}$$

Here $z(t)$ is the vector of the phase coordinates of the system, $\varphi(z)$ is some continuous nonlinear vector function which does not depend explicitly on time, and $Q(t)$ is a vector of generalized control forces which is to be found. The problem is to find a force $Q(t)$ such that system (1) is transferred from a given point $z(0)$ of the phase space at $t = 0$ to a position of equilibrium $z(T) = 0$ at a previously assigned time T.

Equation (1) can be written

$$\dot{z} = Az + f(z) + Q(t), \tag{2}$$

if we assume that $\varphi(z) = Az + f(z)$, where A is a constant square matrix and $f(0) = 0$.

Let us examine the linear system

$$\dot{u} = Au + Q_0(t). \tag{3}$$

The problem of finding the control force for the system (3) has been solved by Ya. N. Roĭtenberg [1].

If the interval $[0, T]$ is subdivided into a number of subintervals, on each of which the components of the vector $Q_0(t)$ are assumed constant, then the number of subintervals should be such that the total number of steps of all nonzero components of the control force is equal to the order of the system. This condition ensures uniqueness in the determination of forces from a given class of functions for a given method of partitioning the interval.

The general solution of the matrix equation (3) is of the form

$$u(t) = N(t)u(0) + \int_0^t N(t - \tau)Q_0(\tau)\,d\tau, \tag{4}$$

where $N(t)$ is the fundamental matrix of the homogeneous system $\dot{u} = Au$, and $N(t - \tau) = N(t)N^{-1}(\tau)$ is the matrix weighting function for the system (3).

Let us assume that the vector $Q_0(t)$ has only one nonzero component q_{0l}. Then the solution (4) in scalar form becomes

$$u_j(t) = \sum_{k=1}^{r} N_{jk}(t)u_k(0) + \int_0^t N_{jl}(t - \tau)q_{0l}(\tau)\,d\tau, \quad j = 1, 2, \cdots, r.$$

* Originally published in *Vestnik Moskovskogo Universiteta, Matematika, Mekhanika*, 1968, no. 4, pp. 55–64. Submitted on June 15, 1967. This translation into English has been prepared by R. N. and N. B. McDonough.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† Computing Center, Moscow State University, Moscow, USSR.

By assuming that $u(T) = 0$, the interval $[0, T]$ is subdivided into subintervals $[t_{\nu-1}, t_\nu]$, $\nu = 1, 2, \cdots, r$, $t_0 = 0$, $t_r = T$, and the values $q_{0l}^{(\nu)}$ of the force q_{0l} are considered constant on these subintervals. We then have the system of algebraic equations

$$\sum_{\nu=1}^{r} q_{0l}^{(\nu)} \int_{t_{\nu-1}}^{t_\nu} N_{jl}(T - \tau)\, d\tau = -\sum_{k=1}^{r} N_{jk}(T) u_k(0), \qquad j = 1, 2, \cdots, r, \tag{5}$$

for the $q_{0l}^{(\nu)}$. If the determinant of this system is not zero, then

$$Q_0(t) = -BN(T)u(0), \tag{6}$$

where the components of the vector $Q_0(t)$ are the values $q_{0l}^{(\nu)}$ of the force q_{0l} on the subintervals $[t_{\nu-1}, t_\nu]$, and B is the inverse of the matrix of coefficients of the system (5).

To determine the force $Q(t)$ in (1), we will consider the following process of successive approximations. Let us pass from the nonlinear differential equation (2) to the equivalent integral equation

$$z(t) = N(t)z(0) + \int_0^t N(t - \tau) f[z(\tau)]\, d\tau + \int_0^t N(t - \tau) Q(\tau)\, d\tau,$$

and define the iterative process

$$\begin{aligned} &N(T)z(0) + \int_0^T N(T - \tau) f[z_n(\tau)]\, d\tau + \int_0^T N(T - \tau) Q_n(\tau)\, d\tau = 0, \\ &z_{n+1}(t) = N(t)z(0) + \int_0^t N(t - \tau) f[z_{n+1}(\tau)]\, d\tau + \int_0^t N(t - \tau) Q_n(\tau)\, d\tau, \\ &\qquad\qquad n = 0, 1, 2, \cdots. \end{aligned} \tag{7}$$

If we take $n = 0$ in the first equation of (7), and put $z_0(t) \equiv 0$, then from the condition $f(0) = 0$ there follows the equation for determining the force $Q_0(t)$ which takes the linear system (3) from the state $u(0) = z(0)$ to the state $u(T) = z(T) = 0$. If the force $Q_0(t)$ thus determined is applied to the nonlinear system (2), then we have, from the second equation of (7),

$$z_1(t) = N(t)z(0) + \int_0^t N(t - \tau) f[z_1(\tau)]\, d\tau + \int_0^t N(t - \tau) Q_0(\tau)\, d\tau.$$

Knowing the solution of this, we can calculate the integral

$$\int_0^T N(T - \tau) f[z_1(\tau)]\, d\tau,$$

needed to determine the force $Q_1(t)$ from the first equation of (7). This first approximation $Q_1(t)$ to the required force will allow one to determine the second approximation $z_2(t)$ to the solution, which will determine the second approximation $Q_2(t)$ to the force, etc.

An analogous process was examined in [2]. The approximation $z_n(t)$ will result in a certain residual at time T, this being the deviation from the equilibrium position.

The residual $z_n(T)$ of the nth approximation, i.e., the value of the solution $z_n(t)$ for $t = T$, can be found from the second equation of (7) as

$$z_n(T) = N(T)z(0) + \int_0^T N(T - \tau) f[z_n(\tau)]\, d\tau + \int_0^T N(T - \tau) Q_{n-1}(\tau)\, d\tau.$$

Substituting $\int_0^T N(T-\tau)f[z_n(\tau)]\,d\tau$ into the first equation of (7), we obtain

$$\int_0^T N(T-\tau)q_n(\tau)\,d\tau = -z_n(T), \qquad q_n(t) = Q_n(t) - Q_{n-1}(t),$$

which is an equation for determining the supplementary force $q_n(t)$ as that force which transfers the linear system (3) from the equilibrium position $u(0) = 0$ to the position $u(T) = -z_n(T)$. The solution of the linear system on which the correcting stepwise force $q_n(t)$ acts is in this case given by the formula

$$q_n(t) = -Bz_n(T), \tag{8}$$

and will have the form

$$u_n(t) = \int_0^t N(t-\tau)q_n(\tau)\,d\tau, \tag{9}$$

with $u_n(T) = -z_n(T)$.

It is seen that for the iterative process under discussion, it is only necessary to know the value of the solutions of system (1) at the time $t = T$.

For use in the convergence proof of the suggested method of successive approximations, we introduce norms for all quantities, vector and matrix, entering into the system (2). For the norm of a vector function, we take the sum of the maxima of the absolute values of the functions which are its components, and as the norm of a matrix, we take the sum of the absolute values of the elements of the matrix.

Since the process of bringing the nonlinear system (1) to equilibrium is to take place over a finite interval of time, the solutions of the homogeneous system corresponding to (3) are bounded on this interval. Thus, we can find a constant $C \geqq 1$ such that

$$\|N(t)\| \leqq C$$

for all $t \in [0, T]$, where $\|N(t)\|$ is the norm of the fundamental matrix of the homogeneous equation. If α is the norm of the matrix B entering into (6) and (8), then the zeroth approximation $Q_0(t)$ to the sought force will satisfy

$$\|Q_0(t)\| \leqq \alpha A_3^{(0)}\|z(0)\| \tag{10}$$

where $A_3^{(0)} = C$.

Further, we assume that, for some number m satisfying

$$me^{CmT} < \frac{1}{C^2\alpha T^2}, \tag{11}$$

there is a closed region $R(\|z\| \leqq A_0)$ in which the nonlinear function $f(z)$ in (2) satisfies a Lipschitz condition for the norm:

$$\|f(z_1) - f(z_2)\| \leqq m\|z_1 - z_2\|. \tag{12}$$

Introducing the notation

$$C\alpha T = \kappa, \qquad CmT = \mu, \tag{13}$$

we determine a number

$$A_1 = \left\{ C + \kappa \left[C + \frac{C\mu e^{\mu}(1 + \kappa)}{1 - \kappa\mu e^{\mu}} \right] \right\} e^{\mu}, \tag{14}$$

and assume that the norm $\|z(0)\|$ is such that

$$A_1 \|z(0)\| < A_0 \tag{15}$$

is always satisfied.

From the condition $C \geqq 1$, we have

$$A_1 > 1. \tag{16}$$

From (12) and the condition $f(0) = 0$, there follows the estimate

$$\|f(z)\| \leqq m\|z\|, \tag{17}$$

for the nonlinear function $f(z)$.

In the future, we will need the following estimate [3]. If for $t \geqq 0$ a continuous function $z(t)$ is positive, $z(t) \geqq 0$, and if we can find two nonnegative numbers C_1 and C_2 such that

$$z(t) \leqq C_1 + \int_0^t C_2 z(\tau)\, d\tau,$$

then

$$z(t) \leqq C_1 e^{C_2 t}. \tag{18}$$

Let us estimate the first approximation $z_1(t)$. From the second formula of (7) it follows that

$$z_1(t) = N(t)z(0) + \int_0^t N(t - \tau) f[z_1(\tau)]\, d\tau + \int_0^t N(t - \tau) Q_0(\tau)\, d\tau. \tag{19}$$

It is necessary to prove that the action of the force $Q_0(t)$ on the nonlinear system (2) will be such that the solution $z_1(t)$ remains within the region R on the entire interval of interest, i.e., to prove that $\|z_1(t)\| \leqq A_0$, $t \in [0, T]$.

Let us assume the contrary. Then by virtue of the continuity of the solutions, a point $t_1 \in [0, T]$ can be found such that $\|z_1(t)\| = A_1\|z_0\|$ on the interval $[0, t_1]$. From the condition $z_1(0) = z(0)$ and (16) it follows that $t_1 > 0$. If we estimate the solution $z_1(t)$ on the interval $0 \leqq t \leqq t_1$, where (17) is valid, replacing the upper limit of the last integral in (19) by T, there results

$$\|z_1(t)\| \leqq C\|z(0)\| + \int_0^t Cm\|z_1(\tau)\|\, d\tau + \int_0^T C\alpha A_3^{(0)} \|z(0)\|\, d\tau, \qquad t \in [0, t_1]. \tag{20}$$

From this, using (18), we have

$$\|z_1(t)\| \leqq A_1^{(1)} \|z(0)\|, \tag{21}$$

where $A_1^{(1)} = (C + \kappa C)e^{\mu}$. But, by hypothesis, on the interval $[0, t_1]$, $\|z_1(t)\| = A_1\|z(0)\|$. Thus $A_1 \leqq A_1^{(1)}$, in contradiction to (14); thus such a point t_1 cannot exist on the interval $[0, T]$. Thus, the force $Q_0(t)$ which brings the linear system (3)

to equilibrium does not carry the solution of the nonlinear system (2) out of the region R at any point of the interval $[0, T]$. Consequently, the estimate (21) is valid on the interval $[0, T]$.

Let us estimate the residual of the solution $z_1(t)$ at $t = T$. Since the point in phase space from which the linear and nonlinear systems are brought to equilibrium is the same for both, from (4) and (19) we have

$$z_1(t) - u(t) = \int_0^t N(t - \tau) f[z_1(\tau)]\, d\tau. \tag{22}$$

In order to make use of inequality (18) in the estimation of the difference (22), in the integrand we add and subtract the function $N(t - \tau)f[u(\tau)]$. To distinguish the solution $u(t)$ of the linear system for a given step of the iteration with the initial condition $u(0) = z(0)$ from all following solutions, we use the subscript zero. Then

$$z_1(t) - u_0(t) = \int_0^t N(t - \tau)\{f[z_1(\tau)] - f[u_0(\tau)]\}\, d\tau + \int_0^t N(t - \tau) f[u_0(\tau)]\, d\tau.$$

It must be shown that the solution $u_0(t)$ lies within the region R. In fact, according to (4), (10), and (15),

$$\|u_0(t)\| \leqq (C + \kappa C)\|z(0)\| < A_1\|z(0)\| < A_0. \tag{23}$$

Using the estimates (12), (17), and (23), in estimating the difference (22), we obtain

$$\|z_1(t) - u_0(t)\| \leqq \int_0^t Cm\|z_1(\tau) - u_0(\tau)\|\, d\tau + \int_0^T Cm(C + \kappa C)\|z(0)\|\, d\tau,$$

from which, taking into account (18) and the fact that $u_0(T) = 0$, we have

$$\|z_1(T)\| \leqq A_2^{(1)}\|z(0)\|,$$

where

$$A_2^{(1)} = C\mu(1 + \kappa)e^{\mu}. \tag{24}$$

Knowing the estimate of the residual of the first approximation, from (8) it is possible to estimate the correcting force $q_1(t)$, which in turn allows us to estimate the first approximation $Q_1(t) = Q_0(t) + q_1(t)$ of the control force.

The inequality

$$\|q_1(t)\| \leqq \alpha A_2^{(1)}\|z(0)\| \tag{25}$$

holds, where, as before, α is the norm of the matrix B. From this it follows that

$$\|Q_1(t)\| \leqq \|Q_0(t)\| + \|q_1(t)\| = \alpha A_3^{(1)}\|z(0)\|;$$

here,

$$A_3^{(1)} = A_3^{(0)} + A_2^{(1)} = C + C\mu e^{\mu}(1 + \kappa).$$

Let us estimate the difference between the first and zeroth approximations, $z_1(t) - z_0(t)$. Since, in constructing the zeroth approximation, we assumed

$z_0(t) \equiv 0$ in the first of equations (7), at this step of the iterative process the estimate of the difference $z_1(t) - z_0(t)$ will be the same as the estimate (21) of the first approximation $z_1(t)$. In the following it is convenient to use the notation $A_1^{(1)} = A_4^{(1)}$. Then $\|z_1(t) - z_0(t)\| \leqq A_4^{(1)}\|z(0)\|$.

It remains to be shown that the force $Q_1(t)$ which is determined from the residual $z_1(T)$ of the first approximation will not result in a second approximation which goes outside the region R.

In fact, from (7) we have

$$z_2(t) = N(t)z(0) + \int_0^t N(t-\tau)f[z_2(\tau)]\,d\tau + \int_0^t N(t-\tau)Q_1(\tau)\,d\tau.$$

As before, we assume there exists a point $t_1 > 0$ such that $\|z_2(t)\| = A_1\|z(0)\|$ on the interval $[0, t_1]$.

Analogous to the estimate (20), we obtain

$$\|z_2(t)\| \leqq C\|z(0)\| + \int_0^t Cm\|z_2(\tau)\|\,d\tau + \int_0^T C\alpha A_3^{(1)}\|z(0)\|\,d\tau, \quad t \in [0, t_1],$$

from which, according to (18),

$$\|z_2(t)\| \leqq A_1^{(2)}\|z(0)\|, \tag{26}$$

where

$$A_1^{(2)} = (C + \kappa A_3^{(1)})e^{\mu} = \{C + \kappa[C + C\mu e^{\mu}(1 + \kappa)]\}e^{\mu}.$$

But, by hypothesis, there exists a point t_1 such that $\|z_2(t)\| = A_1\|z(0)\|$ for $t \in [0, t_1]$, from which follows the inequality $A_1 \leqq A_1^{(2)}$. However, from (11) and (13) we have $\kappa\mu e^{\mu} < 1$; hence from (14) this last inequality is false. Thus we have a contradiction, which is to say that $z_2(t)$ does not leave the region R anywhere on the interval $[0, T]$.

In the estimation of the residual $z_2(T)$ we need to take into account that $u_1(T) + z_1(T) = 0$, where $u_1(t)$ is the solution of system (3) with initial condition $u_1(0) = 0$. Thus taking into account (7) and (9), we can obtain

$$z_2(t) - z_1(t) - u_1(t) = \int_0^t N(t-\tau)\{f[z_2(\tau)] - f[z_1(\tau)]\}\,d\tau. \tag{27}$$

For $t = T$, expression (27) is the residual $z_2(T)$. In the integrand of (27), let us add and subtract the quantity $N(t-\tau)f[z_1(\tau) + u_1(\tau)]$. We obtain

$$\begin{aligned} z_2(t) - z_1(t) - u_1(t) = &\int_0^t N(t-\tau)\{f[z_2(\tau)] - f[z_1(\tau) + u_1(\tau)]\}\,d\tau \\ &+ \int_0^t N(t-\tau)\{f[z_1(\tau) + u_1(\tau)] - f[z_1(\tau)]\}\,d\tau. \end{aligned} \tag{28}$$

To make use of the condition (12) in the estimation of (28), we need to show that the sum $z_1(t) + u_1(t)$ also lies in R. To do this, we estimate the solution $u_1(t)$.

According to (9), $u_1(t) = \int_0^t N(t-\tau)q_1(\tau)\,d\tau$, from which, using (25), we have $\|u_1(t)\| \leqq \kappa A_2^{(1)}\|z(0)\|$. Using this and (21), we obtain

$$\|z_1(t) + u_1(t)\| \leqq \|z_1(t)\| + \|u_1(t)\| = (A_1^{(1)} + \kappa A_2^{(1)})\|z(0)\|,$$

where

$$A_1^{(1)} + \kappa A_2^{(1)} = (C + \kappa C)e^{\mu} + \kappa C\mu e^{\mu}(1 + \kappa).$$

Since $\mu > 0$,

$$A_1^{(1)} + \kappa A_2^{(1)} < \{C + \kappa[C + C\mu e^{\mu}(1 + \kappa)]\}e^{\mu} = A_1^{(2)}.$$

As was shown by the inequality (26), $A_1^{(2)}\|z(0)\| < A_0$. Thus, also, $(A_1^{(1)} + \kappa A_2^{(1)}) \cdot \|z(0)\| < A_0$, i.e., the sum $z_1(t) + u_1(t)$ lies in the region R.

It is now possible to write that

$$\|z_2(t) - z_1(t) - u_1(t)\| \leqq \int_0^t Cm\|z_2(\tau) - z_1(\tau) - u_1(\tau)\|\,d\tau + \int_0^T Cm\|u_1(\tau)\|\,d\tau,$$

from which $\|z_2(t) - z_1(t) - u_1(t)\| \leqq \kappa\mu e^{Cmt}A_2^{(1)}\|z(0)\|$. For $t = T$, we thus have the estimate $\|z_2(T)\| \leqq A_2^{(2)}\|z(0)\|$, where $A_2^{(2)} = \kappa\mu e^{\mu}A_2^{(1)}$. Thus we have estimated the residual $z_2(T)$.

Let us now estimate the second correcting force $q_2(t)$. From (8), $\|q_2(t)\| \leqq \alpha A_2^{(2)}\|z(0)\|$. Since $Q_2(t) = Q_1(t) + q_2(t)$, we have for $Q_2(t)$ the estimate

$$\|Q_2(t)\| \leqq \alpha A_3^{(2)}\|z(0)\|,$$

where

$$A_3^{(2)} = A_3^{(1)} + A_2^{(2)}.$$

We will now estimate the difference between the second and first approximations:

$$z_2(t) - z_1(t) = \int_0^t N(t-\tau)\{f[z_2(\tau)] - f[z_1(\tau)]\}\,d\tau + \int_0^t N(t-\tau)q_1(\tau)\,d\tau.$$

From (12) and (25) we have

$$\|z_2(t) - z_1(t)\| \leqq \int_0^t Cm\|z_2(\tau) - z_1(\tau)\|\,d\tau + \int_0^T C\alpha A_2^{(1)}\|z(0)\|\,d\tau,$$

or, taking into account the estimate (18),

$$\|z_2(t) - z_1(t)\| \leqq A_4^{(2)}\|z(0)\|,$$

where

$$A_4^{(2)} = \kappa e^{\mu}A_2^{(1)}.$$

It remains to be shown that $z_3(t)$ does not depart from the region R. For this, we estimate the third approximation to obtain

$$\|z_3(t)\| \leqq C\|z(0)\| + \int_0^t Cm\|z_3(\tau)\|\,d\tau + \int_0^T C\alpha A_3^{(2)}\|z(0)\|\,d\tau.$$

As before, we assume that this inequality is true only on the interval $0 \leqq t \leqq t_1$, where the point t_1 is defined as before, from which follows the contradictory inequality $A_1 \leqq A_1^{(3)}$, since

$$A_1^{(3)} = \left\{ C + \kappa \left[C + \frac{C\mu e^{\mu}(1 + \kappa)(1 - (\mu\kappa e^{\mu})^2)}{1 - \mu\kappa e^{\mu}} \right] \right\} e^{\mu} < A_1 .$$

This means that $z_3(t)$ does not depart from the region R anywhere on the interval $[0, T]$.

Thus we have shown that the forces $Q_1(t)$ and $Q_2(t)$, determined from the corresponding approximations through the residuals $z_1(T)$ and $z_2(T)$, do not result in solutions which depart from the region R.

The following general assertion is valid: for any n, $n = 1, 2, \cdots$, there exist constants $A_1^{(n)}, A_2^{(n)}, A_3^{(n)}, \cdots$, depending on n, such that, for $t \in [0, T]$,

$$\|z_n(t)\| \leqq A_1^{(n)} \|z(0)\| , \tag{29}$$

$$\|z_n(T)\| \leqq A_2^{(n)} \|z(0)\| , \tag{30}$$

$$\|q_n(t)\| \leqq \alpha A_2^{(n)} \|z(0)\| , \tag{31}$$

$$\|Q_n(t)\| \leqq \alpha A_3^{(n)} \|z(0)\| , \tag{32}$$

$$\|z_n(t) - z_{n-1}(t)\| \leqq A_4^{(n)} \|z(0)\| , \tag{33}$$

and such that $z_{n+1}(t)$ does not leave region R for $t \in [0, T]$. Further,

$$A_1^{(n)} = (C + \kappa A_3^{(n-1)}) e^{\mu}, \qquad n = 1, 2, \cdots , \tag{34}$$

$$A_2^{(n)} = \kappa\mu e^{\mu} A_2^{(n-1)} , \qquad n = 2, 3, \cdots , \tag{35}$$

$$A_3^{(n)} = A_3^{(n-1)} + A_2^{(n)}, \qquad n = 1, 2, \cdots , \tag{36}$$

$$A_4^{(n)} = \kappa e^{\mu} A_2^{(n-1)}, \qquad n = 2, 3, \cdots . \tag{37}$$

In fact, for $n = 2$, this assertion has been proved above. Assuming that it is true for any n, we can show that it is true also for $n + 1$.

For the $(n + 1)$th iteration, according to the second of equations (7), we have

$$z_{n+1}(t) = N(t)z(0) + \int_0^t N(t - \tau) f[z_{n+1}(\tau)] \, d\tau + \int_0^t N(t - \tau) Q_n(\tau) \, d\tau .$$

By hypothesis, the force $Q_n(t)$ produces a $z_{n+1}(t)$ which does not move outside the region R. Hence from (17) and (32),

$$\|z_{n+1}(t)\| \leqq C\|z(0)\| + \int_0^t Cm\|z_{n+1}(\tau)\| \, d\tau + \int_0^T C\alpha A_3^{(n)} \|z(0)\| \, d\tau ,$$

from which

$$\|z_{n+1}(t)\| \leqq A_1^{n+1} \|z(0)\| ,$$

where

$$A_1^{(n+1)} = (C + \kappa A_3^{(n)}) e^{\mu} ,$$

which is the recursion relation (34).

To estimate the residual $z_{n+1}(T)$, we use the fact that $z_n(T) + u_n(T) = 0$, where $u_n(t)$ is the solution of the linear system (3) with initial condition $u_n(0) = 0$, from which, just as in the case of $n = 2$, there follows

$$z_{n+1}(t) - z_n(t) - u_n(t) = \int_0^t N(t-\tau)\{f[z_{n+1}(\tau)] - f[z_n(\tau)]\}\, d\tau.$$

For $t = T$, this difference is $z_{n+1}(T)$.

This last equation can be written

$$z_{n+1}(t) - z_n(t) - u_n(t) = \int_0^t N(t-\tau)\{f[z_{n+1}(\tau)] - f[z_n(\tau) + u_n(\tau)]\}\, d\tau + \int_0^t N(t-\tau)\{f[z_n(\tau) + u_n(\tau)] - f[z_n(\tau)]\}\, d\tau.$$

It is necessary to show that the sum of solutions $z_n(t) + u_n(t)$ lies in the region R. We will estimate the solution $u_n(t)$. Since $u_n(0) = 0$, according to (9),

$$u_n(t) = \int_0^t N(t-\tau) q_n(\tau)\, d\tau,$$

so that, using (31), we have

$$\|u_n(t)\| \leqq \kappa A_2^{(n)} \|z(0)\|. \tag{38}$$

Thus, using (29) and (38),

$$\|z_n(t) + u_n(t)\| \leqq \|z_n(t)\| + \|u_n(t)\| = (A_1^{(n)} + \kappa A_2^{(n)})\|z(0)\|.$$

Since

$$A_1^{(n)} + \kappa A_2^{(n)} = (C + \kappa A_3^{(n-1)})e^{\mu} + \kappa A_2^{(n)} < [C + \kappa(A_3^{(n-1)} + A_2^{(n)})]e^{\mu} < A_1^{(n+1)},$$

the sum $z_n(t) + u_n(t)$ is in the region R.

Thus

$$\|z_{n+1}(t) - z_n(t) - u_n(t)\| \leqq \int_0^t Cm\|z_{n+1}(\tau) - z_n(\tau) - u_n(\tau)\|\, d\tau + \int_0^T Cm\|u_n(\tau)\|\, d\tau,$$

or

$$\|z_{n+1}(T)\| \leqq A_2^{(n+1)} \|z(0)\|,$$

where

$$A_2^{(n+1)} = \kappa\mu e^{\mu} A_2^{(n)}.$$

Thus, the recursion relation (35) is proved.

For the $(n+1)$th correcting force, from (8) it follows that

$$\|q_{n+1}(t)\| \leqq \alpha A_2^{(n+1)} \|z(0)\|. \tag{39}$$

Using (32) and (39), we have

$$\|Q_{n+1}(t)\| \leqq \|Q_n(t)\| + \|q_{n+1}(t)\| = \alpha A_3^{(n+1)} \|z(0)\|,$$

where

$$A_3^{(n+1)} = A_3^{(n)} + A_2^{(n+1)},$$

and thus we have proved the recursion formula (36).

Let us now estimate the difference $z_{n+1}(t) - z_n(t)$:

$$z_{n+1}(t) - z_n(t) = \int_0^t N(t-\tau)\{f[z_{n+1}(\tau)] - f[z_n(\tau)]\}\,d\tau + \int_0^t N(t-\tau)q_n(\tau)\,d\tau.$$

Using (12) and (31), we have

$$\|z_{n+1}(t) - z_n(t)\| \leqq \int_0^t Cm\|z_{n+1}(\tau) - z_n(\tau)\|\,d\tau + \int_0^T C\alpha A_2^{(n)}\|z(0)\|\,d\tau,$$

which, by using (18), yields

$$\|z_{n+1}(t) - z_n(t)\| \leqq A_4^{(n+1)}\|z(0)\|,$$

where

$$A_4^{(n+1)} = \kappa e^{\mu} A_2^{(n)}.$$

Thus the recursion formula (37) is true.

Let us now show that $z_{n+2}(t)$ does not go outside the region R. As before, we assume that $\|z_{n+2}(t)\|$ has the value $A_1\|z(0)\|$ on a certain interval $[0, t_1]$, and we estimate the approximation on that interval. From (29) and (34),

$$\|z_{n+2}(t)\| \leqq A_1^{(n+2)}\|z(0)\|,$$

where

$$A_1^{(n+2)} = (C + \kappa A_3^{(n+1)})e^{\mu},$$

from which it follows that $A_1 \leqq A_1^{(n+2)}$. But, on the other hand, from (35) and (36) we have

$$A_2^{(n+1)} = (\mu\kappa e^{\mu})^n A_2^{(1)}, \tag{40}$$

$$A_3^{(n+1)} = A_2^{(0)} + \sum_{k=1}^{n+1} A_2^{(k)} = A_3^{(0)} + A_2^{(1)} \sum_{k=0}^{n} (\mu\kappa e^{\mu})^k. \tag{41}$$

Since the sum in (41) is a geometric progression with ratio $\mu\kappa e^{\mu}$, with (10) and (24) we have

$$A_1^{(n+2)} = \left\{ C + \kappa\left[C + \frac{C\mu e^{\mu}(1+\kappa)(1-(\mu\kappa e^{\mu})^{n+1})}{1-\mu\kappa e^{\mu}} \right] \right\} e^{\mu},$$

i.e., $A_1 > A_1^{(n+2)}$, since $\mu\kappa e^{\mu} < 1$. This contradiction shows that $z_{n+2}(t) \in R$, $t \in [0, T]$. Thus the inequalities (29)–(33) and the recursion relations (34)–(37) are proved.

From (40) it follows that the sequence $\{A_2^n\}$ converges to zero as $n \to \infty$, provided $\kappa\mu e^{\mu} < 1$. Thus the discrepancy $z_n(T)$ tends to zero as $n \to \infty$, by virtue of (30), and thus so also do the additional forces $q_n(t)$. Since the series (41) converges, the sequence of forces $\{Q_n(t)\}$ has a finite limit. According to (37), the sequence $\{A_4^{(n)}\}$ converges to zero, from which it follows that $\lim_{n\to\infty}\|z_n(t) - z_{n-1}(t)\| = 0$.

Thus we have proved the convergence of the iterative process under condition (11) and the other conditions assumed above. But as was shown, condition (11) is always satisfied if condition (15) is satisfied, from which it follows that the iterative process converges under the condition

$$\|z(0)\| < \frac{A_0}{A_1}, \tag{42}$$

where A_1 is given by (14) and A_0, as follows from (12), is given by $\max_{\|z\| \leqq A_0} \|f'(z)\| \leqq m$.

Thus condition (42) is sufficient for convergence of the iterative process which determines the control force $Q(t)$ which takes the nonlinear system (1) to equilibrium in a given time.

The iterative process described here was applied to the problem of bringing a gyrocompass to the meridian, with good results.

The author is deeply grateful to I. A. Balaeva and Ya. N. Roĭtenberg, who read the manuscript and made valuable comments.

REFERENCES

[1] Ya. N. Roĭtenberg, *Some Problems of Control of Motion*, Fizmatgiz, Moscow, 1963.

[2] I. A. Balaeva, *On the problem of determining controlling forces*, Mekhanika Tverdogo Tela, no. 1, 1966.

[3] R. Bellman, *Stability Theory of Differential Equations*, McGraw-Hill, New York, 1953.

SIAM J. CONTROL
Vol. 8, No. 1, February 1970

THE DECISION PROBLEMS OF DEFINITE STOCHASTIC AUTOMATA*

I-NGO CHEN† AND C. L. SHENG‡

Abstract. A k-definite stochastic automaton is defined as a finite stochastic automaton in which the internal state probability distribution depends only on the last k input symbols for some definite integer k. Whether or not a stochastic automaton is definite depends on its state transition matrices. A set of stochastic matrices is called definite of order k by Paz if, for a fixed integer k, any product of k or more of the matrices is a matrix with all rows equal. A study is made of conditions in which linearly independent row vectors in the component matrices will result in identical row vectors in the product matrix. It is established that a definite stochastic automaton with n internal states and with highest rank of its transition matrices $n - m$ is at least $(n - m)/m$-definite. A time saving decision procedure for the definiteness of stochastic automata is presented and illustrated.

1. Introduction. The notion of a definite event was first introduced by Kleene in 1956 [1]. The theory of definite automata was then developed by Rabin and Scott [6], and by Perles et al. [5]. By their definition, a finite sequence of symbols on a certain alphabet is called a tape; and a set of tapes is called a definite event if for some integer k, two tapes coinciding on the last k squares are either both in the set or both not in the set. Automata are used for classifying tapes. An automaton defining a definite event is called a definite automaton. The notion of definiteness has been applied by Paz [3] to stochastic matrices where a finite set of stochastic matrices of the same order is called a definite set of matrices of order k, if there exists an integer k such that for $n \geqq k$, any product of n matrices from the set is a matrix with all rows the same. Such a matrix is called a *stable matrix* by Paz and Reichaw [4]. Following this line, we define a stochastic finite automaton as a system.

$$\Omega = (\Sigma, S, \{A_{\sigma_i}\} : \sigma_i \in \Sigma, \pi_0, \Lambda, F),$$

where
- Σ: input alphabet,
- S: finite set of internal states,
- $\{A_{\sigma_i}\}$: finite set of transition matrices where any element A_{σ_i} is a stochastic matrix. The entry $a_{ij}(\sigma_i)$ of A_{σ_i} is the probability of transition from state s_i to state s_j after the input symbol σ_i is applied, where $s_i, s_j \in S$.
- π_0: initial state probability distribution,
- Λ: set of all state probability distributions,
- F: $F \subseteq \Lambda$, set of final state probability distributions.

Obviously, if the initial state distribution is π_0, then after an input symbol σ_i is applied, the state distribution will be

$$\pi_0 \cdot A_{\sigma_i}.$$

* Received by the editors January 28, 1969, and in revised form August 15, 1969. This work was supported in part by the National Research Council of Canada under Grant A-1690.

† Department of Computing Science, University of Alberta, Edmonton, Alberta.

‡ Department of Electrical Engineering, University of Ottawa, Ottawa, Ontario.

Since a tape is a sequence of elements of Σ, the transition matrix for a tape $x = \sigma_1\sigma_2 \cdots \sigma_n$ is

$$A_x = A_{\sigma_1} \cdot A_{\sigma_2} \cdots A_{\sigma_n},$$

where σ_i might equal σ_j for $1 \leqq i, j \leqq n$. If x and y are tapes, then

$$A_{xy} = A_x \cdot A_y.$$

Now we define a k-definite stochastic automaton (or definite stochastic automaton of order k) as a finite stochastic automaton such that for any tape x of length $n \geqq k$, for certain fixed integer k, A_x is stable. A k-definite stochastic automaton Ω will thus forget its past except for the last k intervals of time. In other words, for x with length $n \geqq k$, the state distribution of the stochastic automaton will depend only on the last k inputs and will be independent of the initial state distribution. The state probability distributions in the first k intervals of time might be different, depending on different initial state probability distributions when the input tape x applies. However, after input of k symbols, Ω will reach a situation where the state probability distribution can be predicted if only the last k input symbols are known. We call such a state probability distribution a *stable distribution.* All other state probability distributions of Ω which are not independent of the initial state probability distribution are called *transient distributions.* Thus, similar to the deterministic case (where for a definite automaton of order k, a destined state is reached following a sequence of transient states of length k), for a k-definite stochastic automaton Ω, a stable distribution will be reached through a sequence of transient distributions of length k. The number of stable distributions of Ω is at most equal to $(\#(\Sigma))^k$, where $\#(\Sigma)$ denotes the number of elements of Σ. Whether a stochastic automaton is definite or not thus depends on its set of transition matrices $\{A_{\sigma_i}\}$. A stochastic automaton is definite if and only if its set of transition matrices is definite. In this paper, we investigate some properties of the transition matrix A_{σ_i}, particularly conditions under which a stable matrix can be obtained through multiplication of stochastic matrices which are not stable, and thereupon obtain a time saving decision procedure for the definiteness of stochastic automata.

2. Some properties of k-definite stochastic automata. Consider any transition matrix A_{σ_i} of a stochastic automaton Ω. Before the input symbol σ_i is applied, the state probability distribution is a certain vector, say π, where $\pi \in \Lambda$. After the input symbol σ_i is applied, the state probability distribution changes from π to δ, $\delta \in \Lambda$. Thus from the point of view of state probability distributions, Ω is nothing but a set of mappings, taking a point π in Λ into a point δ in Λ. The domain of every mapping is Λ. If Ω has n internal states, then Λ is a polyhedral convex set W_n determined by n unit vectors $e_1, e_2, \cdots, e_n$, where e_i is an n-component vector with the ith component 1 and all other components 0. The ranges of the mappings are different for different input symbols. For an input symbol σ_i, the range of the mapping is R_{σ_i} which is also a polyhedral convex set determined by points which are the row vectors of the matrix A_{σ_i}. We call these points the *determining points* of R_{σ_i}. Since A_{σ_i} is stochastic, R_{σ_i} is contained in W_n. If a tape x is applied, the result is a successive mapping. The range of the resultant mapping is R_x, which is

again a polyhedral convex set determined by points which are row vectors of the matrix A_x. If A_x is stable, then R_x is a single point set. Now the interesting problem is under what conditions R_x will be a single point set. From the theory of Markov chains, we know that if A_{σ_i} is regular, and if $x = \sigma_i\sigma_i \cdots \sigma_i = \sigma_i^n$, then R_x approaches a limiting point as n approaches infinity. Since we are interested only in the condition that R_x is a single point set for definite length of x, we consider first the situation when a stochastic matrix is multiplied by another one of the same order. Suppose A, B are stochastic matrices of the same order and

$$A \cdot B = C. \tag{1}$$

Then C is stochastic also. As mentioned before, each matrix can be regarded as a mapping, and the range of each mapping is a polyhedral convex set determined by points which are row vectors of the matrix. Thus corresponding to each matrix in (1), we have ranges R_A, R_B and R_C respectively. Since any row of C is a convex combination of all rows of B,

$$R_B \supseteq R_C$$

which implies not only that the dimension of R_C is equal to or less than that of R_B, but also that every point of R_C is contained in R_B. In (1), we may think of B as a mapping M_B which maps the point set R_A onto the point set R_C, i.e.,

$$M_B(R_A) = R_C.$$

Let the ith row vector of A be a_i and that of C be c_i. As defined before, a_i is a determining point of R_A and c_i is a determining point of R_C. Thus

$$M_B(a_i) = c_i \quad \text{for } i = 1, 2, \cdots, n,$$

if the order of A and C is n. Since B is stochastic, M_B is affine. Therefore, if a linear relationship exists among the determining points of R_A, the same relationship will exist among the corresponding determining points of R_C. Hence, we have only to consider those linearly independent points of R_A in deciding whether or not R_A can be mapped by a mapping or a string of mappings into a single point set. If this is possible, the upper bound of the length of the string of mappings is determined by the order of A. The following theorem has been proved by Paz [3].

THEOREM 1 (Paz). *If a stochastic finite automaton Ω is definite, and the number of internal states of Ω is n, then Ω is at most $(n-1)$-definite.*

LEMMA 1. *If $\Omega = (\Sigma, S, \{A_{\sigma_i}\}, \pi_0, \Lambda, F)$ is a definite stochastic automaton, then all of the matrices in $\{A_{\sigma_i}\}$ are singular.*

Proof. If any matrix in $\{A_{\sigma_i}\}$, say A_{σ_j}, is nonsingular, then $A_{\sigma_j}^2$ is nonsingular, and for any definite integer k, $A_{\sigma_j}^k$ is not stable. This is contradictory to the assumption that Ω is definite.

LEMMA 2. *Let B be an $m \times n$ stochastic matrix where all m row vectors are linearly independent. Then points which are linearly independent in the domain of M_B will be mapped into linearly independent points in the range of M_B.*

Proof.[1] Let $\{a_i\}$ be a finite set of stochastic m-component row vectors. Let $a_1, a_2, \cdots, a_l$ be linearly independent vectors of $\{a_i\}$. Let

$$A = \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_l \end{bmatrix}.$$

By Sylvester's inequality,

$$\text{rank}\,(A \cdot B) \geqq \text{rank}\,(A) + \text{rank}\,(B) - \#\text{ of rows of } B.$$

Since, by assumption, all rows of A and B are respectively linearly independent,

$$l \leqq m \leqq n$$

and

$$\text{rank}\,(B) = m.$$

Hence

$$\text{rank}\,(A \cdot B) \geqq \text{rank}\,(A) = l.$$

But

$$\text{rank}\,(A \cdot B) = \min(\text{rank}\,(A), \text{rank}\,(B)).$$

Therefore

$$\text{rank}\,(A \cdot B) = \text{rank}\,(A).$$

This completes the proof.

As mentioned before, let W_n be a polyhedral convex set determined by n unit vectors $e_1, e_2, \cdots, e_n$ in n-dimensional space. We then have the following lemma.

LEMMA 3. *If B is of order n and rank $n - m$, then at most $m + 1$ points which are linearly independent in W_n will be mapped by M_B into a single point in W_n.*

Proof. Let A be a nonsingular stochastic matrix of order n. Then

$$\text{rank}\,(A \cdot B) \leqq \text{rank}\,(B) = n - m,$$

$$\text{rank}\,(A \cdot B) > \text{rank}\,(A) + \text{rank}\,(B) - \#\text{ of rows of } B$$

$$= n + n - m - n = n - m.$$

Hence,

$$\text{rank}\,(A \cdot B) = n - m,$$

and no more than $m + 1$ rows of $A \cdot B$ will be equal. This completes the proof.

Now, in (1), consider the conditions that linearly independent points of R_A will be mapped by M_B into a single point in R_C. First of all, by Lemma 1, the

[1] This version of the proof was suggested by the referee.

matrix B must be singular. Suppose now that the order of B is n, and the rank of B is $n-1$. Let $b_1, b_2, \cdots, b_n$ be the row vectors of B; then there exists an equation

$$\sum_{i=1}^{n} \alpha_i \cdot b_i = 0,$$

where the α_i are real and not all equal to zero. We then may have

$$b_n = \sum_{i=1}^{n-1} \beta_i \cdot b_i.$$

Since B is stochastic,

$$\sum_{i=1}^{n-1} \beta_i = 1.$$

Let

$$e'_n = \sum_{i=1}^{n-1} \beta_i \cdot e_i$$

be called the image of e_n in the hyperplane W_{n-1} determined by $e_1, e_2, \cdots, e_{n-1}$. We then have the following theorem.

THEOREM 2. *With B and W_n as defined above, let L be a line segment in W_n. Then all points in L will be mapped by M_B into a single point if and only if L and the line $\overline{e_n e'_n}$ are coplanar and parallel.*

Proof.[2] If L is coplanar and parallel to $\overline{e_n e'_n}$, then L can be represented as

$$a = a_1 + \lambda(e_n - e'_n),$$

where a_1 is a particular point and λ is a scalar. We then have

$$a \cdot B = a_1 \cdot B + \lambda(e_n - e'_n) \cdot B = a_1 \cdot B = c_1,$$

which is a point.

On the other hand, let the set of points which are mapped by M_B into a single point c_1 be of the form

$$a = a_1 + \delta,$$

where a_1 is a particular solution of the equation

$$a \cdot B = c_1$$

and δ belongs to the null space of B. Since nullity of $B = n - \text{rank}\,(B) = n - n + 1 = 1$, $a = a_1 + \delta$ is a line which is parallel to $\overline{e_n e'_n}$ for $(e_n - e'_n) \cdot B = 0$.

The following two corollaries are immediate consequences of Theorem 2.

COROLLARY 1. *If B is of order n and rank $n-2$, and*

$$b_{n-1} = \sum_{i=1}^{n-2} \gamma_i \cdot b_i,$$

$$b_n = \sum_{i=1}^{n-2} \beta_i \cdot b_i;$$

[2] This version of the proof was suggested by the referee.

thus

$$e'_{n-1} = \sum_{i=1}^{n-2} \gamma_i \cdot e_i,$$

$$e'_n = \sum_{i=1}^{n-2} \beta_i \cdot e_i.$$

Let H be a plane segment in W_n; then all points in H will be mapped by M_B into a single point in W_n if and only if H is parallel to the lines $\overline{e_n e'_n}$ and $\overline{e_{n-1} e'_{n-1}}$.

COROLLARY 2. *Let B be of order n, rank $n - m$, and H be a hyperplane segment in W_n. If $M_B(H)$ is a single point, then the dimension of H is at most m.*

THEOREM 3. *Let Ω be a stochastic finite automaton with n internal states. If Ω is definite and has maximum transition matrix rank of $n - m$, then Ω is k-definite, where k is an integer and $k \geqq (n - m)/m$.*

Proof. The proof follows directly from Lemma 3 and Corollary 2.

3. Decision procedures. By Lemma 1, if a stochastic automaton Ω is definite, then every element of its set of stochastic transition matrices must be singular. The most straightforward decision procedure is to form the matrix products of all possible combinations of all elements of the set $\{A_{\sigma i}\}$ and see whether all these products are stable. If not, then repeat the same procedure after increasing the number of component matrices from two to three, and finally, to $n - 1$, if n is the number of the internal states of Ω. This seems tedious for larger n and for large numbers of elements of Σ. A much simpler testing procedure can be obtained by the following argument.

First, for any matrix A, let A' be a matrix containing only those rows of A which are linearly independent. The inverse operation of this operation is characterized by a coefficient matrix T such that

$$T \cdot A' = A.$$

For example, if A is of order n and rank m, then A' is of order $m \times n$ and T is of order $n \times m$. When $n - m$ zero columns are added to the right of T, the augmented matrix is called a *minimum coefficient matrix* by Liu [2].

For a set of matrices $\{A_{\sigma_i} : \sigma_i \in \Sigma\}$, we construct a set of matrices called the Q-matrices as follows:

$$Q_{\sigma_i} = A_{\sigma_i} \quad \text{for all } \sigma_i \in \Sigma,$$

$$Q_{\sigma_i \sigma_j} = Q'_{\sigma_i} \cdot T_{\sigma_j} \quad \text{for all } \sigma_i, \sigma_j \in \Sigma,$$

and, in general, if

$$x = \sigma_1 \sigma_2 \cdots \sigma_{n-1} \sigma_n,$$

$$Q_x = Q'_{\sigma_1 \sigma_2 \cdots \sigma_{n-1}} \cdot T_{\sigma_2 \sigma_3 \cdots \sigma_{n-1} \sigma_n},$$

and

$$Q_x = T_x \cdot Q'_x \quad \text{for all } x \in \Sigma^+.$$

The Q-matrix obtained above is quite similar to the *minimum reduced matrix* defined by Liu [2].

THEOREM 4. *For any tape x, A_x is stable if and only if Q_x is stable.*

Proof. Assume, without loss of generality, that

$$x = \sigma_1\sigma_2 \cdots \sigma_{n-1}\sigma_n.$$

Then

$$\begin{aligned} A_x &= A_{\sigma_1} \cdot A_{\sigma_2} \cdot A_{\sigma_3} \cdots A_{\sigma_n} \\ &= Q_{\sigma_1} \cdot Q_{\sigma_2} \cdot Q_{\sigma_3} \cdots Q_{\sigma_n} \\ &= T_{\sigma_1} \cdot Q'_{\sigma_1} \cdot T_{\sigma_2} \cdot Q'_{\sigma_2} \cdots T_{\sigma_n} \cdot Q'_{\sigma_n} \\ &= T_{\sigma_1} \cdot Q_{\sigma_1\sigma_2} \cdot Q_{\sigma_2\sigma_3} \cdots Q_{\sigma_{n-1}\sigma_n} Q'_{\sigma_n} \\ &= T_{\sigma_1} \cdot T_{\sigma_1\sigma_2} \cdot Q'_{\sigma_1\sigma_2} \cdot T_{\sigma_2\sigma_3} \cdot Q'_{\sigma_2\sigma_3} \cdots T_{\sigma_{n-1}\sigma_n} \cdot Q'_{\sigma_{n-1}\sigma_n} \cdot Q'_{\sigma_n} \\ &\quad \vdots \\ &= T_{\sigma_1} \cdot T_{\sigma_1\sigma_2} \cdot T_{\sigma_1\sigma_2\sigma_3} \cdots T_{\sigma_1\sigma_2\cdots\sigma_n} \cdot Q'_{\sigma_1\sigma_2\cdots\sigma_n} \\ &\quad \cdot Q'_{\sigma_2\sigma_3\cdots\sigma_n} \cdots Q'_{\sigma_{n-1}\sigma_n} \cdot Q'_{\sigma_n}. \end{aligned} \tag{2}$$

Obviously, if Q_x is stable, then Q'_x is stable, and from (2), since all component matrices are stochastic, A_x will be stable also.

On the other hand, let

$$A = U \cdot V,$$

where

$$\begin{aligned} U &= T_{\sigma_1} \cdot T_{\sigma_1\sigma_2} \cdot T_{\sigma_1\sigma_2\sigma_3} \cdots T_{\sigma_1\sigma_2\cdots\sigma_n}, \\ V &= Q'_{\sigma_1\sigma_2\sigma_3\cdots\sigma_n} \cdot Q'_{\sigma_2\sigma_3\cdots\sigma_n} \cdots Q'_{\sigma_{n-1}\sigma_n} \cdot Q'_{\sigma_n}. \end{aligned}$$

Since all rows of Q'_x and $Q'_{\sigma_2\sigma_3\cdots\sigma_n}$ are respectively linearly independent, by Lemma 2,

$$\operatorname{rank}(Q'_x \cdot Q'_{\sigma_2\sigma_3\cdots\sigma_n}) = \operatorname{rank}(Q'_x).$$

Similarly,

$$\begin{aligned} \operatorname{rank}(V) &= \operatorname{rank}(Q'_x \cdot Q'_{\sigma_2\sigma_3\cdots\sigma_n} \cdots Q'_{\sigma_{n-1}\sigma_n} \cdot Q'_{\sigma_n}) \\ &= \operatorname{rank}(Q'_x) \end{aligned}$$

and

$$\#\text{ of rows of } V = \operatorname{rank}(V) = \operatorname{rank}(Q'_x). \tag{3}$$

Now

$$\operatorname{rank}(T_{\sigma_1}) = \operatorname{rank}(Q_{\sigma_1}) = \operatorname{rank}(Q'_{\sigma_1}),$$

$$\operatorname{rank}(T_{\sigma_1\sigma_2}) = \operatorname{rank}(Q_{\sigma_1\sigma_2}) = \operatorname{rank}(Q'_{\sigma_1\sigma_2}).$$

But

$$Q_{\sigma_1\sigma_2} = Q'_{\sigma_1} \cdot T_{\sigma_2}. \tag{4}$$

Thus

$$\operatorname{rank}(Q_{\sigma_1\sigma_2}) \leqq \operatorname{rank}(Q'_{\sigma_1}),$$
$$\operatorname{rank}(T_{\sigma_1\sigma_2}) \leqq \operatorname{rank}(T_{\sigma_1}).$$

Similarly, we have

$$\begin{aligned}\operatorname{rank}(T_{\sigma_1}) &\geqq \operatorname{rank}(T_{\sigma_1\sigma_2}) \geqq \operatorname{rank}(T_{\sigma_1\sigma_2\sigma_3}) \cdots \\ &\geqq \operatorname{rank}(T_{\sigma_1\sigma_2\sigma_3\cdots\sigma_{n-1}}) \geqq \operatorname{rank}(T_{\sigma_1\sigma_2\sigma_3\cdots\sigma_n}) \\ &= \operatorname{rank}(Q_x) = \operatorname{rank}(Q'_x).\end{aligned}$$

From (4), we have

$$\begin{aligned}\#\text{ of rows of } T_{\sigma_1\sigma_2} &= \#\text{ of rows of } Q_{\sigma_1\sigma_2} = \#\text{ of rows of } Q'_{\sigma_1} \\ &= \operatorname{rank}(Q'_{\sigma_1}) = \operatorname{rank}(T_{\sigma_1}).\end{aligned}$$

By Sylvester's inequality,

$$\operatorname{rank}(T_{\sigma_1} \cdot T_{\sigma_1\sigma_2}) \geqq \operatorname{rank}(T_{\sigma_1\sigma_2}).$$

Similarly,

$$\operatorname{rank}(U) = \operatorname{rank}(T_{\sigma_1} T_{\sigma_1\sigma_2} T_{\sigma_1\sigma_2\sigma_3} \cdots T_{\sigma_1\sigma_2\cdots\sigma_n}) \geqq \operatorname{rank}(T_x).$$

Now, since

$$A_x = U \cdot V,$$

from Sylvester's inequality and (3),

$$\operatorname{rank}(A_x) \geqq \operatorname{rank}(U) \geqq \operatorname{rank}(T_x) = \operatorname{rank}(Q_x).$$

Therefore, if Q_x is not stable, then neither will be A_x.

From the above argument, we have the following corollary.

COROLLARY 3.

$$\operatorname{rank}(A_x) = \operatorname{rank}(Q_x).$$

Thus in order to check whether A_x is stable, we need only check whether Q_x is stable. The advantage of introducing the Q-matrices is that they are simpler in manipulation, for

$$\begin{aligned}Q_x &= Q'_{\sigma_1\cdots\sigma_{n-1}} \cdot T_{\sigma_2\cdots\sigma_{n-1}\sigma_n} \\ &= (Q'_{\sigma_1\cdots\sigma_{n-2}} \cdot T_{\sigma_2\cdots\sigma_{n-1}})' \cdot T_{\sigma_2\cdots\sigma_n} \\ &\;\;\vdots \\ &= (\cdots((Q'_{\sigma_1} \cdot T_{\sigma_2})' \cdot T_{\sigma_2\sigma_3})' \cdots)' \cdot T_{\sigma_2\cdots\sigma_n},\end{aligned}$$

and the T-matrices are much simpler than the A-matrices.

Given a stochastic automaton Ω with a set of transition matrices $\{A_{\sigma_i} : \sigma_i \in \Sigma\}$, the decision procedure for the definiteness of Ω is:

(i) Set $k = 1$.

(ii) Construct Q_x and T_x for all $x \in \Sigma^k$.
Determine if all of the Q_x are stable.
If so, Ω is k-definite.
If not, then go to step (iii).
(iii) Determine if every matrix T_x is an identity matrix I.
If so, Ω is not definite.
If not, determine if $k = n - 1$, where n is the number of internal states of Ω.
If so, Ω is not definite.
If not, put $k = k + 1$; go back to (ii).

4. Example. Let $\Sigma = \{0, 1\}$. Let a stochastic automaton Ω have a set of transition matrices $\{A_0, A_1\}$, where

$$A_0 = \begin{bmatrix} .25 & .225 & .225 & .3 \\ .1 & .25 & .45 & .2 \\ .2 & .2 & .2 & .4 \\ .2 & .225 & .275 & .3 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} .2 & .3 & .1 & .4 \\ .1 & .35 & .35 & .2 \\ .3 & .35 & .15 & .2 \\ .2 & .325 & .175 & .3 \end{bmatrix}.$$

Following the decision procedure described in § 3, we first have $Q_0 = A_0, Q_1 = A_1$, where

$$Q_0 = T_0 \cdot Q_0' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/4 & 1/4 \end{bmatrix} \cdot \begin{bmatrix} .25 & .225 & .225 & .3 \\ .1 & .25 & .45 & .2 \\ .2 & .2 & .2 & .4 \end{bmatrix},$$

$$Q_1 = T_1 \cdot Q_1' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/4 & 1/4 \end{bmatrix} \cdot \begin{bmatrix} .2 & .3 & .1 & .4 \\ .1 & .35 & .35 & .2 \\ .3 & .35 & .15 & .2 \end{bmatrix}.$$

Next we have

$$Q_{00} = Q_0' \cdot T_0 = \begin{bmatrix} .4 & .3 & .3 \\ .2 & .3 & .5 \\ .4 & .3 & .3 \end{bmatrix},$$

$$Q_{01} = Q_0' \cdot T_1 = \begin{bmatrix} .4 & .3 & .3 \\ .2 & .3 & .5 \\ .4 & .3 & .3 \end{bmatrix} = Q_{00},$$

$$Q_{10} = Q_1' \cdot T_0 = \begin{bmatrix} .4 & .4 & .2 \\ .2 & .4 & .4 \\ .4 & .4 & .2 \end{bmatrix},$$

$$Q_{11} = Q_1' \cdot T_1 = \begin{bmatrix} .4 & .4 & .2 \\ .2 & .4 & .4 \\ .4 & .4 & .2 \end{bmatrix} = Q_{10}.$$

Thus

$$T_{00} = T_{01} = T_{10} = T_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Now

$$Q_{000} = Q_{00}' \cdot T_{00} = \begin{bmatrix} .4 & .3 & .3 \\ .2 & .3 & .5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} .7 & .3 \\ .7 & .3 \end{bmatrix},$$

$$Q_{001} = Q_{00}' \cdot T_{01} = Q_{00}' \cdot T_{00} = Q_{000},$$

$$Q_{010} = Q_{01}' \cdot T_{10} = Q_{00}' \cdot T_{00} = Q_{000},$$

$$Q_{011} = Q_{01}' \cdot T_{11} = Q_{00}' \cdot T_{00} = Q_{000},$$

$$Q_{100} = Q_{10}' \cdot T_{00} = \begin{bmatrix} .4 & .4 & .2 \\ .2 & .4 & .4 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} .6 & .4 \\ .6 & .4 \end{bmatrix},$$

$$Q_{101} = Q'_{10} \cdot T_{01} = Q_{100},$$
$$Q_{110} = Q'_{11} \cdot T_{10} = Q_{100},$$
$$Q_{111} = Q'_{11} \cdot T_{11} = Q_{100}.$$

Since all Q_x, for $x \in \Sigma^3$, are stable, Ω is 3-definite.

REFERENCES

[1] S. C. KLEENE, *Representation of events in nerve nets and finite automata*, Automata Studies, Annals of Mathematics Studies, vol. 34, Princeton University Press, Princeton, 1956, pp. 3–41.

[2] C. L. LIU, *A note on definite stochastic sequential machines*, Information and Control, 14 (1969), pp. 407–421.

[3] A. PAZ, *Definite and quasidefinite sets of stochastic matrices*, Proc. Amer. Math. Soc., 16 (1965), pp. 634–641.

[4] A. PAZ AND M. REICHAW, *Ergodic theorems for sequences of infinite stochastic matrices*, Proc. Cambridge Philos. Soc., 63 (1967), pp. 777–784.

[5] M. PERLES, M. O. RABIN AND E. SHAMIR, *The theory of definite automata*, IEEE Trans. Electronic Computers, EC-12 (1963), pp. 233–243.

[6] M. O. RABIN AND D. SCOTT, *Finite automata and their decision problems*, IBM J. Res. Develop., 3 (1959), pp. 114–125.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

NUMERICAL SOLUTION OF DYNAMICAL OPTIMIZATION PROBLEMS*

S. DE JULIO†

1. Introduction. The problem of computing optimal controls is of great interest today as is shown by the number of works dealing with this matter. But while there has been in the literature a large harvest of papers suggesting methods for the computation of optimal controls for finite-dimensional systems, this is not the case for systems with infinite-dimensional state space.

From a theoretical standpoint, there have been some attempts to generalize Pontryagin's maximum principle to distributed parameter systems, as for example in a paper by Butovskii [1], in which the description of the system is in the form of an integral equation, or in a paper by Egorov [2], which deals with systems governed by a particular type of partial differential equations.

As far as the computational aspect of optimal controls for systems governed by partial differential equations is concerned, only a few papers have appeared in the literature. These papers usually deal with particular examples and no attempt has been made to give a general computing technique, not even for the linear case, except for the recent papers by Balakrishnan [3] and the author [4], [5].

For example, Sakawa [6], [7] has considered the boundary control problem for distributed parameter systems with one-dimensional space variable and whose representation is in integral form. Yeh and Tou [8] have studied systems of the form

$$a_0 \frac{D^n}{Dt^n} x(\xi; t) + a_1 \frac{D^{n-1}}{Dt^{n-1}} x(\xi; t) + \cdots + a_n x(\xi; t) = u(\xi; t),$$

where

$$\frac{D^j}{Dt^j} = \left(\frac{\partial}{\partial t} + v \frac{\partial}{\partial \xi} \right)^j, \qquad j = 1, 2, \cdots, n.$$

Yavin and Sivan [9] have dealt with the boundary control problem for the wave equation. In all three cases the authors, assuming a quadratic performance index, have reduced the optimization problem to the solution of a Fredholm integral equation of the second kind.

The determination of closed loop optimal controls has been studied by Kim and Erzberger [10]. They have applied the dynamic programming technique to systems described by wave equations with control on the boundary. Using a quadratic performance index, they have obtained a set of Riccati equations whose solution gives the closed loop optimal control. They have also shown how in some particular cases, namely when the time and the space variables can be separated, an approximate solution of the Riccati equation can be computed.

* Received by the editors December 12, 1968, and in revised form September 16, 1969.

† Istituto di Automatica, Universita di Roma, Rome, Italy. This research was supported in part by the United States Air Force Office of Scientific Research, Applied Mathematics Division, under Grants 68-1408 and 700-69 and in part by the Consiglio Nazionale delle Ricerche.

A technique for the computation of time-optimal controls for systems governed by diffusion equations with one-dimensional space variable has been proposed by Goldwyn, Sriram and Graham [11]. These authors seek the time-optimal control in the class of bang-bang controls. Making use of the Laplace transform they obtain a method for approximate computation of the switching times.

A more general computational technique has been obtained by Axelband [12], which may be applied to systems whose state evolution is defined as the solution of a Cauchy problem. The optimization problem is that of minimizing a continuous convex cost functional, and the author suggests to resort to convex programming for its solution. The shortcoming of the Axelband method seems to be that one has to compute and store as many responses of the system as there are grid points corresponding to the discretization of the space and the time variables.

We also have to mention the work of Lions [13] who has tackled the optimization problem for parabolic and hyperbolic systems. The optimal control is given by the solution of a two-point boundary value problem for which the author proposes a numerical algorithm.

This paper is a study of the technique proposed by Balakrishnan [3] as applied to infinite-dimensional linear systems. This technique is completely different from any of the ones discussed above and seems to have two major advantages over them. First of all, it is more general in that it applies to a broad class of systems which contains the previous ones as particular cases. The second advantage, but not the least, is that the computational algorithm is by far simpler than any other.

The next section is devoted to defining briefly the optimization problem and the new computational technique (the ε-method). The relevant theorems from a previous paper by the author [5] are also reported, while the main concern of the remaining part of the paper is with the numerical solution of a class of optimization problems using the technique treated there in detail. More specifically, in § 3 we give some general concepts of approximation theory and direct reference is made to the work of Trotter [14].

The general concepts of approximation of spaces and operators have been investigated by other authors as well. In particular, we may mention the papers by Aubin [15], [16], although Aubin's work develops in different directions, such as finding the truncation error due to a certain approximation or the best approximation of an operator.

In § 4 we tackle the problem of the approximate minimization of functionals as applied to the solution of the ε-problem.

The ε-method approach presents some resemblance to the Tikhonov method of regularization which has been investigated by some Russian authors [17]–[20], although the goals of the two methods are different. In fact, both methods resort to penalty functions, but while in the method of regularization the scope of the penalty function added to the cost functional is to force a minimizing sequence to converge to the optimum, in the ε-method the penalty function accounts for the dynamics of the system.

2. The optimization problem. In this work we consider systems governed by equations of the type

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad x(0) = 0, \tag{2.1}$$

where for each $t \in [0, T]$ the state $x(t)$ and the control $u(t)$ are elements of the Hilbert spaces H_1 and H_2 respectively, A is an (unbounded) linear operator mapping a domain $D(A)$ dense in H_1 into H_1, and B is a bounded linear operator mapping H_2 into H_1.

The solution of (2.1) will be taken in the weak sense, i.e., instead of (2.1) we consider the following mathematical model for the system

$$\bar{S}x = Bu, \tag{2.2}$$

where the bar denotes closure and S is an operator defined over a suitable dense subset D of $L^2(T; H_1)$ (for more details the interested reader is referred to S. De Julio [5]) given by

$$(Sx)(t) = \frac{\partial x(t)}{\partial t} - Ax(t), \qquad x(0) = 0. \tag{2.3}$$

The optimization problem can be formulated as follows.

The optimization problem. Given the system governed by (2.1), a set $U \subset L^2(T; H_2)$ of admissible controls and a cost functional $J(u; x)$, find $u^0 \in U$, $x^0 \in D(\bar{S})$ such that

$$J(u^0; x^0) = \inf_{\substack{u \in U \\ x \in D \\ \bar{S}x = Bu}} J(u; x).$$

The basic idea underlying the new computing method is that of reducing the dynamical optimization problem into a nondynamical one. This is achieved by defining a new problem, which we shall call the ε-problem, in which a penalty function accounts for the dynamics of the system.

The ε-problem. Let a functional J_ε be defined by

$$J_\varepsilon(u; x) = J(u; x) + \frac{1}{\varepsilon}\|\bar{S}x - Bu\|^2, \tag{2.4}$$

where $\varepsilon > 0$, and the norm is the $L^2(T; H_1)$ norm. Determine $u_\varepsilon \in U$, $x_\varepsilon \in D(\bar{S})$, such that

$$J_\varepsilon(u_\varepsilon; x_\varepsilon) = j_\varepsilon = \inf_{\substack{u \in U \\ x \in D}} J_\varepsilon(u; x).$$

It is intuitive that as ε becomes smaller and smaller, u_ε and x_ε come closer and closer to u^0 and x^0. In fact, this has been demonstrated by the author [5] and we shall here report the relevant theorems.

THEOREM 2.1. *Let* (2.1) *be the system equation with A such that the operator S is closable. Let the set U of admissible controls be closed and convex. Let the cost functional J enjoy the following properties*:

(p_1) $J(u; x) \geqq 0$ *for all* u, x,

(p_2) J *is weakly lower semicontinuous*,

(p_3) J *is radially unbounded*.[1]

[1] Property (p_3) means that if $\{u_n, x_n\}$ is a sequence such that $\lim_{n\to\infty}(\|u_n\| + \|x_n\|) = \infty$ then $\lim_{n\to\infty} J(u_n; x_n) = \infty$.

Then there exist $u_\varepsilon \in U$, $x_\varepsilon \in D(\bar{S})$ *which solve the* ε*-problem. Namely,*

$$J_\varepsilon(u_\varepsilon; x_\varepsilon) = j_\varepsilon = \inf_{\substack{u\in U \\ x\in D}} J_\varepsilon(u; x),$$

where J_ε *is given by* (2.4).

Proof. Let $\{u_n, x_n\}$, $u_n \in U$, $x_n \in D$, be a sequence realizing the infimum of J_ε:

$$J_\varepsilon(u_n; x_n) \downarrow j_\varepsilon.$$

Being a decreasing sequence, $\{J_\varepsilon(u_n; x_n)\}$ is obviously bounded. Recalling the definition (2.4) of J_ε, we see that, because of the property (p_1) of J, J_ε is the sum of two nonnegative quantities. Therefore, $\{J(u_n; x_n)\}$ also is bounded. Hence, because of property (p_3), there will exist constants C_1 and C_2 such that

$$\|u_n\| \leqq C_1, \qquad \|x_n\| \leqq C_2$$

for all n. Since u_n is bounded in norm and B is a bounded operator, there will also exist a constant C_3 such that

$$\|Sx_n\| \leqq C_3.$$

Now, in a Hilbert space every bounded set is weakly compact. Therefore, there exist subsequences (relabel them $\{u_n\}$, $\{x_n\}$) and functions $u_\varepsilon \in L^2(T; H_2)$, x_ε, $y_\varepsilon \in L^2(T; H_1)$ such that

$$w\text{-}\lim_{n\to\infty} u_n = u_\varepsilon,$$

$$w\text{-}\lim_{n\to\infty} x_n = x_\varepsilon,$$

$$w\text{-}\lim_{n\to\infty} Sx_n = y_\varepsilon.$$

Since U is closed and convex, it is also weakly closed. Hence, $u_\varepsilon \in U$.

Let ϕ be any function in the domain of S^*. Then, using the definition of weak convergence, we have

$$[y_\varepsilon, \phi] = \lim_{n\to\infty} [Sx_n, \phi] = \lim_{n\to\infty} [x_n, S^*\phi] = [x_\varepsilon, S^*\phi].$$

Since $D(S^*)$ is dense in $L^2(T; H_1)$ and $S^{**} = \bar{S}$, the equality $[y_\varepsilon, \phi] = [x_\varepsilon, S^*\phi]$ for all $\phi \in D(S^*)$ implies $x_\varepsilon \in D(\bar{S})$ and $\bar{S}x_\varepsilon = y_\varepsilon$.

Finally, exploiting the weak lower semicontinuity of J (property (p_2)) and of the norm, we have

$$j_\varepsilon = \lim_{n\to\infty} J_\varepsilon(u_n; x_n) = \lim_{n\to\infty} J(u_n; x_n) + \lim_{n\to\infty} \frac{1}{\varepsilon}\|Sx_n - Bu_n\|^2$$

$$\geqq J(u_\varepsilon; x_\varepsilon) + \frac{1}{\varepsilon}\|\bar{S}x_\varepsilon - Bu_\varepsilon\|^2 = J_\varepsilon(u_\varepsilon; x_\varepsilon)$$

for which only equality can hold. This completes the proof of the theorem.

THEOREM 2.2. *Let the hypotheses of Theorem* 2.1 *be satisfied and let* $\{\varepsilon_n\}$ *be a sequence of positive real numbers such that*

$$\lim_{n\to\infty} \varepsilon_n = 0.$$

Then there exist $u^0 \in U$, $x^0 \in D(\bar{S})$, *and a subsequence (relable it* $\{\varepsilon_n\}$*) such that*

$$w\text{-}\lim_{n\to\infty} u_{\varepsilon_n} = u^0, \qquad w\text{-}\lim_{n\to\infty} x_{\varepsilon_n} = x^0, \tag{2.5}$$

$$\lim_{n\to\infty} J_\varepsilon(u_{\varepsilon_n}; x_{\varepsilon_n}) = J(u^0; x^0), \tag{2.6}$$

where u^0 *and* x^0 *are optimal. Namely,*

$$J(u^0; x^0) = j_0 = \inf_{\substack{u\in U \\ x\in D \\ \bar{S}x = Bu}} J(u; x), \tag{2.7}$$

where J_ε *is given by* (2.4).

Proof. Let us first of all notice that under condition (2.2), $J_\varepsilon = J$. Therefore, j_0 can be considered the infinum of J_ε subject to the additional condition (2.2). Hence,

$$j_0 \geqq j_{\varepsilon_n} = J_{\varepsilon_n}(u_{\varepsilon_n}; x_{\varepsilon_n})$$

and j_0 is well-defined since the set $\{(u, x): u \in U, \bar{S}x = Bu\}$ is nonvoid because at least the $(0, 0)$ element is in it.

As in Theorem 2.1 we then infer that u_{ε_n}, x_{ε_n} and $\bar{S}x_{\varepsilon_n}$ are bounded in norm, and that there exist subsequences (relable them $\{u_{\varepsilon_n}\}$, $\{x_{\varepsilon_n}\}$) and functions $u^0 \in L^2(T; H_2)$, x^0, $y^0 \in L^2(T; H_1)$ such that

$$w\text{-}\lim_{n\to\infty} u_{\varepsilon_n} = u^0,$$

$$w\text{-}\lim_{n\to\infty} x_{\varepsilon_n} = x^0,$$

$$w\text{-}\lim_{n\to\infty} \bar{S}x_{\varepsilon_n} = y^0,$$

and that

$$u^0 \in U, \quad x^0 \in D(\bar{S}), \quad \bar{S}x^0 = y^0.$$

Moreover there also exists a positive constant k such that

$$\frac{1}{\varepsilon_n}\|\bar{S}x_{\varepsilon_n} - Bu_{\varepsilon_n}\|^2 \leqq k^2$$

for all n, or

$$\|\bar{S}x_{\varepsilon_n} - Bu_{\varepsilon_n}\| \leqq k\sqrt{\varepsilon_n}$$

for all n, whence

$$\lim_{n\to\infty} \|\bar{S}x_{\varepsilon_n} - Bu_{\varepsilon_n}\| = 0. \tag{2.8}$$

We now show that u^0 and x^0 satisfy (2.2). Indeed consider the following inequality:

$$\begin{aligned} 0 &\leqq \|(\bar{S}x_{\varepsilon_n} - Bu_{\varepsilon_n}) - (\bar{S}x^0 - Bu^0)\|^2 \\ &= \|\bar{S}x_{\varepsilon_n} - Bu_{\varepsilon_n}\|^2 + \|\bar{S}x^0 - Bu^0\|^2 - 2[\bar{S}x_{\varepsilon_n} - Bu_{\varepsilon_n}, \bar{S}x^0 - Bu^0]. \end{aligned}$$

Passing to the limit as n goes to infinity and taking (2.8) and the definition of weak convergence into account, we get

$$0 \leqq -\|\bar{S}x^0 - Bu^0\|^2,$$

which implies

$$\bar{S}x^0 = Bu^0. \tag{2.9}$$

Finally, to prove (2.6) and (2.7), we use the weak lower semicontinuity of J and the fact that $(1/\varepsilon_n)\|\bar{S}x_{\varepsilon_n} - Bu_{\varepsilon_n}\|^2$ is nonnegative:

$$j_0 \geqq \overline{\lim_{n\to\infty}} J_{\varepsilon_n}(u_{\varepsilon_n}; x_{\varepsilon_n}) \geqq \underline{\lim_{n\to\infty}} J_{\varepsilon_n}(u_{\varepsilon_n}; x_{\varepsilon_n})$$

$$\geqq \underline{\lim_{n\to\infty}} J(u_{\varepsilon_n}; x_{\varepsilon_n}) + \underline{\lim_{n\to\infty}} \frac{1}{\varepsilon_n}\|\bar{S}x_{\varepsilon_n} - Bu_{\varepsilon_n}\|^2 \geqq J(u^0; x^0),$$

which would contradict the definition of the infimum unless equality holds throughout. The equality between the limit superior and the limit inferior implies the existence of the limit, thus showing (2.6), while the equality

$$j_0 = J(u^0; x^0)$$

together with (2.9) proves (2.7).

So far we have been dealing with the optimization problem for abstract distributed control systems. As far as boundary control systems are concerned the optimization problem can be formulated in a similar manner and theorems analogous to Theorems 2.1 and 2.2 hold, although under somewhat more restrictive conditions [5].

It is clear that (2.1) allows us to treat systems governed by partial differential equations. For instance, let Ω be an open set in the n-dimensional Euclidean space R^n, and ξ (the space variable) a vector in Ω, whose components we denote $\xi_1, \cdots, \xi_n$. Then A may be the partial differential operator

$$A = \sum_{|j| \leqq k} a_j(\xi) D^j,$$

where

$$D^j = \frac{\partial^{|j|}}{\partial \xi_1^{j_1} \cdots \partial \xi_n^{j_n}}, \quad |j| = \sum_{i=1}^{n} j_i, \qquad j = (j_1, \cdots, j_n).$$

Moreover the conditions that the operator A must satisfy for Theorems 2.1 and 2.2 to apply are quite mild, so that the classes of systems treated in the references mentioned in the introduction are surely included in our formulation.

3. General concepts of approximation theory. The practical implication of Theorems 2.1 and 2.2 is that they suggest a simple algorithm for the computation of optimal controls. Indeed, since as we have shown, the solution of the ε-problem can be made arbitrarily close (possibly in the weak sense) to the solution of the optimization problem by means of a suitable choice of ε, a satisfactory approximation to the solution of the latter can be obtained by solving the former.

The advantages of solving the ε-problem instead of the optimization problem are clear. In fact, the ε-problem consists of the straightforward minimization of a functional without differential constraints, and the well-known methods of mathematical programming can be used.

If for the solution of the ε-problem one resorts to a digital computer, some considerations concerning the convergence of the numerical solution are in order.

We begin with the fundamentals of approximation theory. Some of the following definitions concerning the approximation of a Banach space by means of sequences of Banach spaces were suggested by Trotter [14].

DEFINITION 3.1. Let X, X_m, $m = 1, 2, \cdots$, be Banach spaces with norms $\|\cdot\|$, $\|\cdot\|_m$ respectively. Let $P^m : X \to X_m$, $m = 1, 2, \cdots$, be linear operators having the following properties:

(H_1) $$\|P^m\| \leqq N$$

for all m, with N a positive constant independent of m;

(H_2) $$\lim_{m \to \infty} \|P^m x\|_m = \|x\| \qquad \text{for all } x \in X;$$

(H_3) There exists a positive constant M independent of m such that, for each $x^m \in X_m$, there is an $x \in X$ such that $P^m x = x^m$ and $\|x\| \leqq M\|x^m\|_m$.

Then we say that the *sequence of Banach spaces* $\{X_m\}$ *approximates* X.

From now on the superscript m will be used to denote vectors belonging to X_m, whereas the subscript m will be used to denote sequences of vectors in X. Moreover, the sequence $\{X_m\}$ will always be a sequence of Banach spaces approximating X.

DEFINITION 3.2. A sequence of vectors $\{x^m\}$, $x^m \in X_m$, is said to *converge* to $x \in X$ if

$$\lim_{m \to \infty} \|P^m x - x^m\|_m = 0,$$

and we write $\lim_{m \to \infty} x^m = x$.

DEFINITION 3.3. A sequence of operators $\{A^m\}$, $A^m : X_m \to Y (X_m \to X_m)$, with T a Banach space, is said to *converge* to the operator $A : X \to Y\ (X \to X)$ if

$$\lim_{m \to \infty} A^m P^m x = Ax$$

for all $x \in D(A)$, and we write $\lim_{m \to \infty} A^m = A$.

The next theorem shows the close relationship between the convergence of a sequence of vectors, as defined in Definition 3.2, and the usual convergence.

THEOREM 3.1. *A sequence* $\{x^m\}$, $x^m \in X_m$, *converges to* $x \in X$ *if and only if there exists a sequence* $\{x_m\}$, $x_m \in X$, *with the properties*

(3.1) $$x^m = P^m x_m,$$

(3.2) $$\lim_{m \to \infty} x_m = x.$$

Proof. Sufficiency. Given the sequence $\{x^m\}$, suppose there is a sequence $\{x_m\}$ satisfying (3.1) and (3.2). Then, due to property (H_1) we have

$$\lim_{m \to \infty} \|x^m - P^m x\|_m = \lim_{m \to \infty} \|P^m x_m - P^m x\|_m \leqq \lim_{m \to \infty} N\|x_m - x\| = 0,$$

whence $\lim_{m \to \infty} x^m = x$.

Necessity. Suppose that the sequence $\{x^m\}$ converges to x. Put $y^m = x^m - P^m x$. Then $\lim_{m\to\infty} y^m = 0$. From property (H_3) we infer that we can construct a sequence $\{y_m\}$, $y_m \in X$, such that $P^m y_m = y^m$ and $\|y_m\| \leqq M\|y^m\|_m$. Hence, $\lim_{m\to\infty} \|y_m\| \leqq \lim_{m\to\infty} M\|y^m\|_m = 0$ or $\lim_{m\to\infty} y_m = 0$. If we put $x_m = y_m + x$, then we have $P^m x_m = P^m y_m + P^m x = y^m + P^m x = x^m$ and obviously $\lim_{m\to\infty} x_m = x$.

Keeping this theorem in mind we now generalize the concept of weak convergence to sequences of vectors belonging to approximating spaces.

DEFINITION 3.4. A sequence $\{x^m\}$ is said to *converge weakly* to x, and we write $w\text{-}\lim_{m\to\infty} x^m = x$, if and only if there exists a sequence $\{x_m\}$, $x_m \in X$, with the properties

$$P^m x_m = x^m,$$

$$w\text{-}\lim_{m\to\infty} x_m = x.$$

The following is a general approximation theorem concerning the evaluation of the minimum of a functional over X in the approximating spaces X_m. It will be specialized to the approximate solution of the ε-problem in the next section.

THEOREM 3.2. *Let f be a real functional, $f : X \to R$, with unique minimum*[2] *at x^0:*

$$f(x^0) = \inf_{x\in X} f(x).$$

Suppose that X is a reflexive Banach space. Let $\{f^m\}$, $f^m : X_m \to R$, be a sequence of functionals converging to f, with the following properties:

(i) *f^m has a unique minimum at x_0^m,*

(ii) $\lim_{m\to\infty} \|x^m\|_m = \infty$ *implies* $\lim_{m\to\infty} f^m(x^m) = \infty$,

(iii) $w\text{-}\lim_{m\to\infty} x^m = x$ *implies* $\lim_{m\to\infty} f^m(x^m) \geqq f(x)$.[3]

Then we have

$$w\text{-}\lim_{m\to\infty} x_0^m = x^0 \quad \textit{and} \quad \lim_{m\to\infty} f^m(x_0^m) = f(x^0).$$

Proof. From the definition of convergence, given a $\delta > 0$, there exists an integer K such that

$$f^m(P^m x^0) \leqq f(x^0) + \delta$$

for all $m \geqq K$. Also,

$$f^m(x_0^m) \leqq f^m(P^m x^0),$$

whence

$$f^m(x_0^m) \leqq f(x^0) + \delta$$

for all $m \geqq K$. Because of (ii), the sequence $\{x_0^m\}$ is bounded in norm, i.e.,

$$\|x_0^m\|_m \leqq K_1$$

for some $K_1 > 0$. Recalling (H_3) of Definition 3.1, we can construct a sequence $\{x_m^0\}$ in X, such that

$$P^m x_m^0 = x_0^m$$

[2] Strict convexity of f is sufficient to guarantee uniqueness of the minimum.

[3] This can be considered a generalization of the definition of weak lower semicontinuity.

and

$$\|x_m^0\| \leqq M\|x_0^m\|_m.$$

Hence, also the sequence $\{x_m^0\}$ is bounded in norm. Therefore, we can extract a subsequence, which we relable $\{x_m^0\}$, weakly converging to some $\bar{x}^0 \in X$ and, by definition,

$$w\text{-}\lim_{m\to\infty} x_0^m = \bar{x}^0.$$

Applying (iii), we get

$$f(x^0) + \delta \geqq \overline{\lim_{m\to\infty}} f^m(x_0^m) \geqq \underline{\lim_{m\to\infty}} f^m(x_0^m) \geqq f(\bar{x}^0).$$

Since δ is arbitrary, it follows that $f(x^0) = f(\bar{x}^0)$ and $\bar{x}^0 = x^0$ because of the uniqueness of the minimum of f. Moreover,

$$\lim_{m\to\infty} f^m(x_0^m) = f(x^0).$$

Remark. It can be easily verified that we need not evaluate the minimum of f^m exactly, though preserving convergence. Namely, it suffices to construct a sequence $\bar{x}^m$, such that

$$f^m(\bar{x}^m) \leqq f^m(x^m) + \alpha_m \qquad \text{for all } x^m \in X_m,$$

where $\lim_{m\to\infty} \alpha_m = 0$.

4. An approximation theorem for the numerical solution of the ε-problem. To make things more concrete, we now consider the case where both u and x take their values in an open subset Ω of the n-dimensional Euclidean space R^n. Both of the spaces H_1 and H_2 will be identified with the Hilbert space $L^2(\Omega)$ of Lebesgue square-integrable functions with values in Ω endowed with the usual norm. The space $L^2(T; H_1)$ will therefore coincide with the space $L^2(Q)$ of obvious interpretation, Q being the cylinder $\Omega \times [0, T]$.

Taking $X = L^2(Q)$ we shall show how the approximation theory developed in § 3 can be applied to the numerical solution of the ε-problem relative to distributed control systems. We shall assume that all the hypotheses of § 2 are satisfied, without explicitly mentioning them.

Having in mind the numerical solution of the problem, it is natural to take for the X_m, subspaces of X consisting of simple functions. More precisely, let $\{Q_j^m\}, j = 1, 2, \cdots, J_m$, be a partition of Q, Q_j^m being disjoint Lebesgue-measurable subsets of Q, with the property

$$\bigcup_{j=1}^{J_m} Q_j^m = Q.$$

Suppose that $\{Q_j^{m'}\}$ is a refinement of $\{Q_j^m\}$ for $m' > m$, and that $\mu(Q_j^m)$ goes to zero as m goes to infinity, μ being the Lebesgue measure over the $(n + 1)$-dimensional Euclidean space. Let X_m be the subspace of $L^2(Q)$ consisting of all functions f^m of the form

$$f^m = \sum_{j=1}^{J_m} q_j^m \chi_j^m,$$

where χ_j^m is the indicator function of Q_j^m, i.e.,

$$\chi_j^m(\xi; t) = \begin{cases} 1 & \text{if } (\xi, t) \in Q_j^m, \\ 0 & \text{if } (\xi, t) \notin Q_j^m, \end{cases}$$

and q_j^m are real numbers.

$P^m: L^2(Q) \to X_m$ will be the projection operator defined by

$$P^m f = f^m = \sum_{j=1}^{J_m} q_j^m \chi_j^m,$$

where

$$q_j^m = \frac{1}{\mu(Q_j^m)} \int_{Q_j^m} f \, d\mu.$$

Then condition (H_1) will always be satisfied and, under certain hypotheses on the partitions $\{Q_j^m\}$ (see [14]), condition (H_2) will also be satisfied.

The following lemma will be used later.

LEMMA 4.1. *Let $\{S^m\}$ be a sequence of operators $X_m \to X_m$ converging to the closable linear operator $S: D \to X$ defined in* (2.3) *and having the following properties*:

(i) $\lim_{m\to\infty} x^m = x \in D$ *implies* $\lim_{m\to\infty} S^m x^m = Sx$;

(ii) *The sequence $\{S^{m*}\}$ converges to S^*.*[4]

Then

$$w\text{-}\lim_{m\to\infty} x^m = x \quad \text{and} \quad w\text{-}\lim_{m\to\infty} S^m x^m = y$$

imply

$$x \in D(\bar{S}) \quad \text{and} \quad y = \bar{S}x.$$

Proof. Let $\{x^m\}$, $\{S^m x^m\}$ converge weakly to x and y respectively. Then for all $\phi \in D(S^*)$ we have

$$[y, \phi] = \lim_{m\to\infty} [S^m x^m, \phi] = \lim_{m\to\infty} [S^m x^m, P^m \phi] = \lim_{m\to\infty} [x^m, S^{m*} P^m \phi].$$

Recalling the definition of convergence of a sequence of operators, we have $\lim_{m\to\infty} S^{m*} P^m \phi = S^* \phi$, whence $\lim_{m\to\infty} [x^m, S^{m*} P^m \phi] = [x, S^* \phi]$. Then the equality $[y, \phi] = [x, S^* \phi]$ for all $\phi \in D(S^*)$ will imply $x \in D(\bar{S})$ and $\bar{S}x = y$.

The next theorem is fundamental in justifying the numerical solution of the ε-problem.

THEOREM 4.1. *Let $J(u; x)$ be a strictly convex continuous functional satisfying conditions* (p_1), (p_2) and (p_3).[5] *Let $\{S^m\}$ be a sequence of operators $X_m \to X_m$, enjoying properties* (i) *and* (ii) *of Lemma* 4.1. *Let $J_\varepsilon^m: X_m \times X_m \to R$ be defined by*

$$J_\varepsilon^m(u^m; x^m) = J(u^m; x^m) + \frac{1}{\varepsilon} \| S^m x^m - B u^m \|$$

[4] It is well known that for partial differential operators, conditions (i) and (ii) are usually satisfied if the ratios between the discretization intervals are suitably chosen.

[5] Actually, the condition that J be continuous and strictly convex can be substituted for the condition (p_2) since the former implies the latter.

and let

$$J_\varepsilon^m(u_\varepsilon^m; x_\varepsilon^m) = \inf_{\substack{u^m \in U_m \\ x^m \in X_m}} J_\varepsilon^m(u^m; x^m),$$

where $U_m = P^m U$. *Then we have*

$$w\text{-}\lim_{m\to\infty} u_\varepsilon^m = u_\varepsilon, \qquad w\text{-}\lim_{m\to\infty} x_\varepsilon^m = x_\varepsilon \tag{4.1}$$

and

$$\lim_{m\to\infty} J_\varepsilon^m(u_\varepsilon^m; x_\varepsilon^m) = J_\varepsilon(u_\varepsilon; x_\varepsilon). \tag{4.2}$$

Proof. First of all we notice that, because of the strict convexity of J, u_ε and x_ε, as well as u_ε^m, x_ε^m are unique.

Since J is continuous, $\{S^m\}$ converges to S, and B is continuous, it is easily seen that the sequence $\{J_\varepsilon^m\}$ approximates J_ε in the sense of Definition 3.3. Hence, with the aid of Lemma 4.1, the proof of Theorem 4.1 follows rather directly from the proof of Theorem 3.2.

An analogous result can be proved for the case of boundary control.

From a computational viewpoint, it may be interesting to investigate the convergence of $u_{\varepsilon_n}^m$ and $x_{\varepsilon_n}^m$ when both m and n go simultaneously to infinity, $\{\varepsilon_n\}$ being a decreasing sequence going to zero.

Let $\{\delta_n\}$ be a decreasing sequence of positive numbers such that $\lim_{n\to\infty} \delta_n = 0$. Then, for each n and each $\phi \in X$ there exist positive integers K_n, H_n such that

$$|[u_{\varepsilon_n}^m - u_{\varepsilon_n}, \phi]| < \delta_n \qquad \text{for all } m \geqq K_n,$$

$$|[x_{\varepsilon_n}^m - x_{\varepsilon_n}, \phi]| < \delta_n \qquad \text{for all } m \geqq H_n.$$

Construct the sequences $\{u_{\varepsilon_n}^{m_n}\}$, $\{x_{\varepsilon_n}^{m_n}\}$, where $m_n < m_{n+1}$ and $m_n \geqq \max(K_n, H_n)$. Then we have the following theorem.

THEOREM 4.2.

$$w\text{-}\lim_{n\to\infty} u_{\varepsilon_n}^{m_n} = u^0, \qquad w\text{-}\lim_{n\to\infty} x_{\varepsilon_n}^{m_n} = x^0$$

and

$$\lim_{n\to\infty} J_\varepsilon^{m_n}(u_{\varepsilon_n}^{m_n}; x_{\varepsilon_n}^{m_n}) = J(u^0; x^0).$$

Proof. We confine ourselves to proving the first equality. Let $\delta > 0$ be arbitrary. Let N and M be such that

$$\delta_n \leqq \delta/2 \qquad \text{for all } n \geqq N,$$

$$|[u_{\varepsilon_n} - u^0, \phi]| \leqq \delta/2 \qquad \text{for all } n \geqq M.$$

Consequently

$$|[u_{\varepsilon_n}^{m_n} - u^0, \phi]| \leqq |[u_{\varepsilon_n}^{m_n} - u_{\varepsilon_n}, \phi]| + |[u_{\varepsilon_n} - u^0, \phi]| < \delta$$

for all $n \geqq K = \max(N, M)$, whence $w\text{-}\lim_{n\to\infty} u_{\varepsilon_n}^{m_n} = u^0$.

Loosely speaking, Theorem 4.2 justifies what we would do in writing a computer program. Namely, we would first choose a value for ε and a certain approximation scheme (e.g., a discretization interval), i.e., a value for m, and minimize J_ε^m. We would then repeatedly increase m until no appreciable improvement of the solution was detected. Next we would reduce the value of ε and proceed as before until neither reduction of ε nor increase of m would produce appreciable changes of the solution.

5. Concluding remarks. After having defined the optimization problem for a class of distributed control systems, we have shown how an approximate solution can be obtained by solving a nondynamical optimization problem.

If, as is usual, one resorts to a digital computer for the computation of the optimal control, the question that naturally arises is: how well does the numerical solution approximate the real solution or, to put it differently, does the numerical solution converge to the continuous solution as the discretization intervals go to zero? It is well known that this is not in general the case. The main objective of this paper has been to give the conditions under which such convergence takes place, devoting particular attention to the solution of the ε-problem.

REFERENCES

[1] A. G. BUTOVSKII, *The maximum principle for optimum systems with distributed parameters*, Avtomat. i Telemekh., 22 (1961), pp. 1288–1301.

[2] A. I. EGOROV, *Optimal processes in distributed parameter systems and certain problems in invariance theory*, this Journal, 4 (1966), pp. 601–661.

[3] A. V. BALAKRISHNAN, *On a new computing technique in optimal control*, this Journal, 6 (1968), pp. 149–173.

[4] S. DE JULIO, *Computation of optimal controls for infinite dimensional systems*, Proc. Second Annual Princeton Conference on Information Sciences and Systems, Princeton University, Princeton, New Jersey, 1968.

[5] ———, *On the optimization of infinite dimensional linear systems*, Paper presented at the 2nd International Conference on Computing Methods in Optimization Problems, San Remo, Italy, 1968.

[6] Y. SAKAWA, *Solution of an optimal control problem in a distributed-parameter system*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 420–426.

[7] ———, *Optimal control of a certain type of linear distributed-parameter system*, Ibid., AC-11 (1966), pp. 35–41.

[8] H. H. YEH AND J. T. TOU, *Optimal control of a class of distributed parameter systems*, Ibid., AC-12 (1967), pp. 29–37.

[9] Y. YAVIN AND R. SIVAN, *The optimal control of distributed parameter systems*, Ibid., AC-12 (1967), pp. 758–761.

[10] M. KIM AND H. ERZBERGER, *On the design of distributed parameter systems with boundary control functions*, Ibid., AC-12 (1967), pp. 22–28.

[11] R. M. GOLDWYN, K. P. SRIRAM AND M. GRAHAM, *The optimal control of a linear diffusion process*, this Journal, 5 (1967), pp. 295–308.

[12] ELLIOT IRA AXELBAND, *The optimal control of certain classes of linear distributed parameter systems*, Doctoral thesis, University of California, Los Angeles, 1966.

[13] J. L. LIONS, *Control problems in systems described by partial differential equations*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, London, 1967, pp. 251–271.

[14] H. F. TROTTER, *Approximation of semi-group of operators*, Pacific J. Math., 8 (1968), pp. 887–919.

[15] J. P. AUBIN, *Evaluation des erreurs de troncature des approximations des espaces de Sobolev*, J. Math. Anal. Appl., 21 (1968), pp. 356–368.

[16] ———, *Best approximation of linear operators in Hilbert space*, SIAM J. Numer. Anal., 5 (1968), pp. 518–521.

[17] A. N. TIKHONOV, *Methods for the regularization of optimal control problems*, Soviet Math. Dokl., 6 (1965), pp. 761–763.

[18] ———, *On the stability of the functional optimization problem*, U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), pp. 28–33.

[19] E. S. LEVITIN AND B. T. POLJAK, *Convergence of minimizing sequences in conditional extremum problems*, Soviet Math. Dokl., 7 (1966), pp. 764–767.

[20] B. T. POLJAK, *Existence theorems and convergence of minimizing sequences in extremum problems with restrictions*, Ibid., 7 (1966), pp. 72–75.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

CHARACTERIZATION OF THE SETS OF CONSTRAINTS FOR WHICH THE NECESSARY CONDITIONS FOR OPTIMIZATION PROBLEMS HOLD*

JEAN PIERRE AUBIN†

Introduction. Many papers are devoted to the problem of proving the necessary conditions for optimization problems. (See for instance [2] to [5], [10], [11], [12], [14].) In this paper, we shall put the problem this way: Let $\mathscr{B}$ be a convex subset of a vector space $\mathscr{X}$, j a functional on $\mathscr{X}$ satisfying the property

$$\text{(A)} \qquad \chi = \inf_{\xi \in \mathscr{B}} j(\xi) > -\infty.$$

On the other hand, there are many ways to represent a closed convex set by a set of constraints. For instance, if $\pi_{\mathscr{B}}$ is the "gauge" of $\mathscr{B}$,

$$\mathscr{B} = \{\xi \in \mathscr{X} \text{ such that } 1 - \pi_{\mathscr{B}}(\xi) \geqq 0\},$$

or, if $\sigma_{\mathscr{B}}(f)$ is the support functional of $\mathscr{B}$,

$$\mathscr{B} = \{\xi \in \mathscr{X} \text{ such that } (f, \xi) - \sigma_{\mathscr{B}}(f) \geqq 0 \text{ for any } f \in \mathscr{L}(\mathscr{B})\},$$

where

$$\mathscr{L}_{\mathscr{B}} = \{f \in \mathscr{X}' \text{ such that } \sigma_{\mathscr{B}}(f) = \inf_{\xi \in \mathscr{B}} (f, \xi) > -\infty\}.$$

Usually, a concrete convex $\mathscr{B}$ is represented by a set $\mathscr{L}$ of concave functionals f in this way:

$$\mathscr{B} = \{\xi \in \mathscr{X} \text{ such that } f(\xi) \geqq 0 \text{ for any } f \in \mathscr{L}\}.$$

When such a representation of $\mathscr{B}$ by a set $\mathscr{L}$ of constraints is given, we shall say that j satisfies the "optimality conditions" ($B(\mathscr{L})$) if and only if:

($B(\mathscr{L})$) There exists $\chi > -\infty$ and $f_0 \in \mathscr{L}$ such that $j(\xi) - \chi \geqq f_0(\xi)$ for any $\xi \in \mathscr{X}$,

which can be useful in several purposes. Clearly, (B($\mathscr{L}$)) implies (A) and the so-called "necessary conditions" hold when the converse is true. This is the Kuhn-Tucker theorem when we translate the problem in terms of "convex programming" or the maximum principle when we look at it as an optimal control problem (see § 2.4, for instance).

We shall solve the following problem: *Characterize the representations $\mathscr{L}$ of $\mathscr{B}$ for which the conditions* (A) *and* (B($\mathscr{L}$)) *are equivalent for a given class $\mathscr{C}$ of functionals j.* The answer lies in Theorem 1.1 and some particular cases are given in Theorems 1.2, 1.3 and 1.4. By differentiation, we shall extend the results to the case where the functional j is directionally differentiable (Corollaries 1.2 and 1.3

* Received by the editors February 5, 1969.

† Division of Mathematical Sciences, Purdue University, Lafayette, Indiana 47907. This work was supported by the Mathematics Research Center, U.S. Army, under Contract DA-31-124-ARO-D-462.

of § 1.3). But we do not know yet if the method used here can again solve the problem under decent assumptions when the constraints are nonconcave differentiable functionals.

In the second part of this paper, we show how the property (B($\mathscr{L}$)) leads to the maximum principle for optimal control problems (Theorem 2.2) when the "state u" and the "control k" are linked by linear differential equations (where the control appears both in the partial differential equation, the boundary conditions and/or the initial conditions) and when the "cost function" is a Gâteaux-differentiable functional (not necessarily quadratic) (see [7] in this case). To illustrate this rather abstract result, we shall construct in § 2.4 the optimality equations for the following simple problem: The constraints are defined by:

(i) $-\Delta u(x) + u(x) = f(x)$ when x belongs to a bounded open set Ω of R^n;

(ii) $u(x) = \varphi(x) + k(x)$ on the boundary Γ of Ω;

(iii) $k(x)$ is positive on Γ;

and the cost function is given by

$$j(u, k) = \int_\Gamma |u(x) - \psi(x)|^p \, d\sigma(x) + \int_\Gamma |k(x)|^p \, d\sigma(x), \qquad 1 < p < +\infty,$$

where f, φ and ψ are given functions. In this case, the property (B($\mathscr{L}$)) amounts to saying that there exist $u_0(x)$, $k_0(x)$, the optimal state and control, and $v_0(x)$, the solution of the adjoint equation satisfying the relations:

(i) $-\Delta u_0 + u_0 = f$ and $-\Delta v_0 + v_0 = 0$ on Ω;

(ii) $k_0 = u_0 - \varphi$ on Γ, $v_0 = |\partial u_0/\partial n - \psi|^{p-2}(\partial u_0/\partial n - \psi)$ on Γ;

(iii) $k_0|_\Gamma \geqq 0$ and $k_0(x)^{p-1} + \partial v_0/\partial n \geqq 0$ on Γ;

(iv) $k_0(x) \cdot (k_0(x)^{p-1} + \partial v_0/\partial n) = 0$ a.e. on Γ.

1. Sets of constraints and necessary conditions for optimization problems.

1.1. Statement of the results. Let $\mathscr{A}$ and $\mathscr{B}$ be two convex subsets of a real space vector $\mathscr{X}$ (without topology for the time) such that

$$\mathscr{B} \subset \mathscr{A} \subset \mathscr{X}; \qquad (\mathscr{A} \text{ can be equal to } \mathscr{X}). \tag{1.1}$$

We shall say that a convex functional j finite on $\mathscr{A}$ satisfies the property (A) if and only if

$$\chi = \inf_{\xi \in \mathscr{B}} j(\xi) > -\infty. \tag{A}$$

Let us assume now that the convex set $\mathscr{B}$ is defined by a set $\mathscr{L}$ of "concave constraints" f defined and finite on $\mathscr{A}$:

$$\mathscr{B} = \{\xi \in \mathscr{A} \text{ such that } f(\xi) \geqq 0 \text{ for any } f \in \mathscr{L}\}. \tag{1.2}$$

Naturally, this can be done in many ways and we shall say that $\mathscr{L}$ is a representation of $\mathscr{B}$. With such a representation $\mathscr{L}$ of $\mathscr{B}$, we shall say that a convex functional j on $\mathscr{A}$ satisfies the property (B($\mathscr{L}$)) ("the $\mathscr{L}$-optimality property") if

and only if:

(B($\mathscr{L}$)) There exists $\chi > -\infty$ and $f_0 \in \mathscr{L}$ such that $j(\xi) - \chi \geqq f_0(\xi)$ for any $\xi \in \mathscr{A}$.

Obviously, the property (B($\mathscr{L}$)) implies the property (A). We shall look for the representations $\mathscr{L}$ such that the properties (A) and (B($\mathscr{L}$)) are equivalent for a given class of functionals j.

Before stating the theorem, we need the following notations. $\mathscr{F}(\mathscr{A})$ is the real vector space of the (finite) functionals on $\mathscr{A}$ equipped with the pointwise convergence topology. $\mathscr{P} \subset \mathscr{F}(\mathscr{A})$ is the closed convex cone of positive functionals. If $\mathscr{T} \subset \mathscr{F}(\mathscr{A})$, $\bar{\mathscr{T}}$ denotes the closure of $\mathscr{T}$ and $\hat{\mathscr{T}}$ the closed convex cone (with vertex 0) spanned by $\mathscr{T}$. If $\mathscr{C}$ is a set of functionals, $\mathscr{C}_c$ denotes the set of functionals $f_c(\xi) = f(\xi) + \lambda$ when $f \in \mathscr{C}$ and $\lambda \in R$ (the field of real numbers).

THEOREM 1.1. *The above conditions* (A) *and* (B($\mathscr{L}$)) *are equivalent for any functional j of a set $\mathscr{C}$ of convex functionals finite on $\mathscr{A}$ if and only if*

$$\mathscr{C}_c \cap (\mathscr{P} + \mathscr{L}) = \mathscr{C}_c \cap \widehat{(\mathscr{P} + \mathscr{L})}. \tag{1.3}$$

Remark 1.1. It is not a restriction to assume that $\mathscr{L}$ is actually a closed convex cone (in $\mathscr{F}(\mathscr{A})$) of constraints, since we can obviously replace $\mathscr{L}$ in (1.2) by the closed convex cone $\bar{\mathscr{L}}$ spanned by $\mathscr{L}$. In particular, we deduce the following corollary.

COROLLARY 1.1. *If the cone $\mathscr{P} + \mathscr{L}$ is closed in $\mathscr{F}(\mathscr{A})$, the conditions* (A) *and* (B($\mathscr{L}$)) *are equivalent for any convex functionals $j(\xi)$.*

This assumption holds for instance when $\mathscr{L}$ is spanned by a finite number of constraints $f_1(\xi), \cdots, f_m(\xi)$. Namely, let us set

$$\mathscr{B}_k = \{\xi \in \mathscr{A} \text{ such that } f_j(\xi) \geqq 0 \text{ for any } j \neq k\} \tag{1.4}$$

and let us assume that for any k,

$$\text{there exists a point } \xi_k \in \mathscr{B}_k \quad \text{such that } f_k(\xi_k) > 0. \tag{1.5}$$

This assumption amounts to saying that there are only useful constraints. We shall give another proof of the following well-known theorem.

THEOREM 1.2. *If $\mathscr{L}$ is spanned by concave constraints $f_1, \cdots, f_m$ such that* (1.5) *holds, the cone $\mathscr{P} + \mathscr{L}$ is closed and the property* (A) *is equivalent to: There exists a sequence $\lambda_1, \cdots, \lambda_m$ of nonnegative scalars such that*

$$j(\xi) - \chi \geqq \sum \lambda_j f_j(\xi) \quad \textit{for any } \xi \textit{ in } \mathscr{A}. \tag{1.6}$$

(See [5], [12] for instance.)

Let us consider now the important case where the constraints are affine functionals.

Let $\mathscr{X}$ and $\mathscr{X}'$ be two paired real vector spaces for the duality pairing (f, ξ) on $\mathscr{X}' \times \mathscr{X}$, equipped with the weak topologies.

We associate with any continuous functional of a convex cone $\mathscr{L}$ of $\mathscr{X}'$ a scalar functional $\sigma(f)$, the affine functional

$$\tilde{f}(\xi) = (f, \xi) - \sigma(f), \tag{1.7}$$

and the closed convex subset $\mathscr{B}$ of $\mathscr{X}$ defined by

$$\mathscr{B} = \{\xi \in \mathscr{X} \text{ such that } (f, \xi) - \sigma(f) \geqq 0 \text{ for any } f \in \mathscr{L}\}. \tag{1.8}$$

Let us notice that we can replace $\sigma(f)$ in (1.8) by the support functional $\sigma_{\mathscr{B}}(f)$ of $\mathscr{B}$ defined by

$$\sigma_{\mathscr{B}}(f) = \inf_{\xi \in \mathscr{B}} (f, \xi) \tag{1.9}$$

since if $f \in \mathscr{L}$ we have $\sigma_{\mathscr{B}}(f) \geqq \sigma(f)$. Let us recall that $\sigma_{\mathscr{B}}$ is an upper semi-continuous concave and positively homogeneous functional of $f \in \mathscr{X}'$. (See [10], [14].)

On the other hand, using this support functional, we can always define an affine representation of a closed convex subset $\mathscr{B}$ of $\mathscr{X}$. Let us summarize these two remarks in the following lemma.

LEMMA 1.1. *The closed convex subset $\mathscr{B}$ defined by* (1.8) *is equal to*

$$\mathscr{B} = \{\xi \in \mathscr{X} \text{ such that } (f, \xi) - \sigma_{\mathscr{B}}(f) \geqq 0 \text{ for any } f \in \mathscr{L}\}. \tag{1.10}$$

More generally, if $\mathscr{L}(\mathscr{B}) = \{f \in \mathscr{X}' \text{ such that } \sigma_{\mathscr{B}}(f) > -\infty\}$, this closed convex set $\mathscr{B}$ is also defined by

$$\mathscr{B} = \{\xi \in \mathscr{X} \text{ such that } (f, \xi) - \sigma_{\mathscr{B}}(f) \geqq 0 \text{ for any } f \in \mathscr{L}_{\mathscr{B}}\}. \tag{1.11}$$

We shall deduce from Theorem 1.1 the following theorems.

THEOREM 1.3. *Let us assume $\mathscr{X}$ barreled and $\mathscr{B}$ defined by* (1.10). *If $\mathscr{L}$ is a weakly closed convex cone of $\mathscr{X}'$ and if j is any finite convex functional, the condition* (A) *is equivalent to the condition* (B($\mathscr{L}$)): *There exists $\chi > -\infty$ and $f_0 \in \mathscr{L}$ such that*

$$j(\xi) - \chi \geqq (f_0, \xi) - \sigma_{\mathscr{B}}(f_0) \quad \text{for any } \xi \in \mathscr{X}. \tag{1.12}$$

THEOREM 1.4. *If $\mathscr{B}$ is defined by* (1.10), *if $\mathscr{L}$ is a weakly closed convex cone of $\mathscr{X}'$ and if $j(\xi) = (g, \xi) - \lambda$ is any (continuous) affine functional, the condition* (A) *is equivalent to* (B($\mathscr{L}$)): *There exists $\xi_0 \in \mathscr{B}$ and $f_0 \in \mathscr{L}$ such that*

$$g = f_0 \quad \text{and} \quad \sigma_{\mathscr{B}}(f_0) = (f_0, \xi_0).$$

Remark 1.2. In Theorem 1.3, we can replace the support functional $\sigma_{\mathscr{B}}(f)$ by the initial functional $\sigma(f)$ if it is also an upper semicontinuous functional which is concave and positively homogeneous.

Remark 1.3. Theorems 1.3 and 1.4 include the case where the number of independent constraints is infinite and the case where the constraints are equality constraints (f is an "equality constraint" if f and $-f$ belong to $\mathscr{L}$).

1.2. Proofs of the results of § 1.1. First of all, the convexity assumptions will imply the following lemma.

LEMMA 1.2. *The property* (A) *is equivalent to*:

For any finite sequence of positive scalars α_i and of points $\xi_i \in \mathscr{A}$, there exists $\chi > -\infty$ such that

$$\text{(A')} \quad \sum \alpha_i f(\xi_i) \geqq 0 \quad \text{for any } f \in \mathscr{L} \quad \text{implies} \quad \sum \alpha_i (j(\xi_i) - \chi) \geqq 0.$$

Obviously, (A') implies (A). Conversely, let us set $\bar{\xi} = \sum \alpha_i \xi_i / \sum \alpha_i$. Since the functionals f of $\mathscr{L}$ are concave, the condition $\sum \alpha_i f(\xi_i) \geqq 0$ for any $f \in \mathscr{L}$ implies that $\bar{\xi}$ belongs to $\mathscr{B}$. Therefore, $j(\bar{\xi}) - \chi \geqq 0$ and since j is convex, we deduce that $\sum \alpha_i (j(\xi_i) - \chi) \geqq 0$.

Let us now recall the following consequence of the Hahn-Banach theorem.

LEMMA 1.3. *Let $\mathcal{F}$ and $\mathcal{F}'$ be two paired real vector spaces, and let $\langle \varphi, f \rangle$ be the duality pairing on $\mathcal{F}' \times \mathcal{F}$. Let $\mathcal{P}^+$ denote the polar cone of a set $\mathcal{P}$:*

$$\mathcal{P}^+ = \{\varphi \in \mathcal{F}' \text{ such that } \langle \varphi, f \rangle \geqq 0 \text{ for any } f \in \mathcal{P}\}. \tag{1.13}$$

The weakly closed convex cone $\widehat{\mathcal{P} + \mathcal{L}}$ spanned by $\mathcal{P} + \mathcal{L}$ in $\mathcal{F}$ is equal to

$$\widehat{\mathcal{P} + \mathcal{L}} = (\mathcal{P}^+ \cap \mathcal{L}^+)^+. \tag{1.14}$$

Obviously, $\widehat{\mathcal{P} + \mathcal{L}}$ is contained in $(\mathcal{P}^+ \cap \mathcal{L}^+)^+$. Let us assume that there exists $f \in (\mathcal{P}^+ \cap \mathcal{L}^+)^+$ such that $f \notin \widehat{\mathcal{P} + \mathcal{L}}$. By the Hahn-Banach theorem, there exists $\varphi \in \mathcal{F}'$ such that $\langle \varphi, f \rangle < 0$ and $\langle \varphi, g \rangle \geqq 0$ for any $g \in \mathcal{P} + \mathcal{L}$. Therefore $\varphi \in (\mathcal{P}^+ \cap \mathcal{L}^+)$ and since f belongs to $(\mathcal{P}^+ \cap \mathcal{L}^+)^+$, we deduce $\langle \varphi, f \rangle \geqq 0$. This is a contradiction.

We shall now apply Lemma 1.2 when $\mathcal{F} = \mathcal{F}(\mathcal{A})$ is the real vector space of functionals defined on $\mathcal{A}$, when $\mathcal{F}' = \mathcal{F}'(\mathcal{A})$ is the space of finite sequences $\varphi = \{(\alpha_i, \xi_i)\}_i$ of pairs of scalars α_i and of points $\xi_i \in \mathcal{A}$ and when the duality pairing $\langle \varphi, f \rangle$ is defined by

$$\langle \varphi, f \rangle = \sum \alpha_i f(\xi_i). \tag{1.15}$$

Therefore the weak topology of $\mathcal{F}(\mathcal{A})$ is nothing else than the pointwise convergence topology, and $\mathcal{F}'(\mathcal{A})$ is the dual of $\mathcal{F}(\mathcal{A})$ for this topology.

We shall choose for $\mathcal{P}$ the (closed convex) cone of positive functionals defined on $\mathcal{A}$.

LEMMA 1.4. *If $\mathcal{P} \subset \mathcal{F}(\mathcal{A})$ is the closed convex cone of positive functionals defined on $\mathcal{A}$, its polar $\mathcal{P}^+ \subset \mathcal{F}(\mathcal{A})$ is the closed convex cone of sequences $\varphi = \{(\alpha_i, \xi_i)\}_i$ with positive coefficients α_i.*

Indeed, if the coefficients α_i of $\varphi = \{(\alpha_i, \xi_i)\}$ are positive, $\langle \varphi, f \rangle = \sum \alpha_i f(\xi_i) \geqq 0$ for any positive functional f of $\mathcal{P}$. Conversely, if $\varphi \in \mathcal{P}^+$, its coefficients α_i are positive since $\langle \varphi, \delta\xi_i \rangle = \alpha_i \geqq 0$ when $\delta\xi_i$ is the positive functional equal to 0 for $\xi \neq \xi_i$ and to 1 when $\xi = \xi_i$. We thus deduce the following lemma.

LEMMA 1.5. *If the convex functional j satisfies* (A), *then*

$$g(\xi) = j(\xi) - \chi \tag{1.16}$$

belongs to the closed convex cone $\widehat{\mathcal{P} + \mathcal{L}}$ spanned by $\mathcal{P} + \mathcal{L}$.

By Lemma 1.2, $g(\xi) = j(\xi) - \chi$ belongs to $(\mathcal{P}^+ \cap \mathcal{L}^+)^+$. Indeed, let φ belong to $(\mathcal{P}^+ \cap \mathcal{L}^+)$. Then the coefficients α_i of φ are positive and $\sum \alpha_i f(\xi_i) \geqq 0$ for all $f \in \mathcal{L}$. Therefore $\langle \varphi, g \rangle = \sum \alpha_i (j(\xi_i) - \chi) \geqq 0$. This amounts to saying that g belongs to $(\mathcal{P}^+ \cap \mathcal{L}^+)^+ = \widehat{\mathcal{P} + \mathcal{L}}$ by Lemma 1.3.

Proof of Theorem 1.1. The sufficiency is a consequence of Lemma 1.5. Conversely, let us assume that (A) and (B($\mathcal{L}$)) are equivalent for any j of $\mathcal{C}$. Let $g = j + \lambda$ belong to $\mathcal{C}_c \cap \widehat{\mathcal{P} + \mathcal{L}}$. Therefore, by Lemmas 1.3 and 1.4, $g(\xi)$ is positive on $\mathcal{B}$, and thus $0 \leqq \chi_0 = \inf_{\xi \in \mathcal{B}} g(\xi)$. The functional $j = g - \lambda$ belongs to $\mathcal{C}$ and satisfies the property (A) since $j(\xi) \geqq \chi_0 - \lambda$ on $\mathcal{B}$. By (B($\mathcal{L}$)), there exists $f_0 \in \mathcal{L}$ such that g is greater than f_0 since

$$g(\xi) = j(\xi) + \lambda \geqq j(\xi) + \lambda - \chi_0 \geqq f_0(\xi). \tag{1.17}$$

Proof of Theorem 1.2. It results from the following lemma.

LEMMA 1.6. *Let $\mathscr{P}$ be a closed convex cone and $\mathscr{L}$ be the closed convex cone spanned by a finite number of elements $f_0, \cdots, f_m$. If for any k, $0 \leqq k \leqq m$, there exists $\varphi_k \in \mathscr{F}'$ such that*

$$\varphi_k \in \mathscr{P}^+, \quad \langle \varphi_k, f_j \rangle \geqq 0 \quad \text{for } j \neq k \quad \text{and} \quad \langle \varphi_k, f_k \rangle > 0; \tag{1.18}$$

then $\mathscr{P} + \mathscr{L}$ is closed. The assumption (1.5) *of Theorem* 1.2 *implies* (1.18) *when we take $\langle \varphi_k, f \rangle = f(\xi_k)$.*

Let us prove Lemma 1.6. Let $j_\nu = h_\nu + \sum_{k=0}^m \lambda_\nu^k f_k$ be a generalized sequence of $\mathscr{P} + \mathscr{L}$ converging to j. Applying φ_k to j_ν, we deduce the estimates

$$0 \leqq \lambda_\nu^k \leqq \frac{\langle \varphi_k, j_\nu \rangle}{\langle \varphi_k, f_k \rangle}, \qquad 0 \leqq k \leqq m. \tag{1.19}$$

Therefore a subsequence λ_μ^k converges to a positive number λ^k, and thus h_μ converges to an element h of $\mathscr{P}$ (which is closed). Then j_μ converges to $j = h + \sum_{k=0}^m \lambda^k f_k$ which belongs to $\mathscr{P} + \mathscr{L}$.

Proof of Theorem 1.3. Let $\tilde{\mathscr{P}}$ be the set of affine functionals $\tilde{f}(\xi) = (f, \xi) - \sigma_{\mathscr{B}}(f)$ when $f \in \mathscr{L}$. Since $\sigma_{\mathscr{B}}$ is concave and positively homogeneous, $\mathscr{P} + \tilde{\mathscr{P}}$ is a convex cone. We have to prove that $\mathscr{P} + \tilde{\mathscr{P}}$ is closed. Let $j(\xi)$ be the limit of a generalized sequence j_ν,

$$j_\nu(\xi) = h_\nu(\xi) + (f_\nu, \xi) - \sigma_{\mathscr{B}}(f_\nu), \tag{1.20}$$

which belongs to $\mathscr{P} + \tilde{\mathscr{P}}$.

Let $\hat{\xi}$ belong to $\mathscr{B}$. Since

$$\sigma_{\mathscr{B}}(f_\nu) = \inf_{\xi \in \mathscr{B}} (f_\nu, \xi) \leqq (f_\nu, \hat{\xi}) \quad \text{and} \quad h_\nu(\xi) \geqq 0, \tag{1.21}$$

we deduce the following estimates:

$$j_\nu(\hat{\xi} - \xi) \leqq (f_\nu, \xi) \leqq j_\nu(\hat{\xi} + \xi) \qquad \text{for any } \xi \in \mathscr{X}. \tag{1.22}$$

Since j_ν converges pointwise to j, the sequence f_ν is weakly bounded and since $\mathscr{X}$ is barreled, this sequence is actually weakly compact in $\mathscr{X}'$. Therefore, a subsequence f_μ converges to an element f of $\mathscr{L}$ (since it is closed) and

$$\begin{aligned} \limsup h_\mu(\xi) &= \limsup (j_\mu(\xi) - (f_\mu, \xi) + \sigma_{\mathscr{B}}(f_\mu)) \\ &= j(\xi) - (f, \xi) + \sigma_{\mathscr{B}}(f) \geqq 0, \end{aligned} \tag{1.23}$$

since $\sigma_{\mathscr{B}}$ is upper semicontinuous. Thus $j(\xi)$ actually belongs to $\mathscr{P} + \tilde{\mathscr{P}}$.

Proof of Theorem 1.4. We shall deduce it from Theorem 1.1 when $\mathscr{C} = \mathscr{C}_c$ is the set of (continuous) affine functionals. Let $j(\xi) = (g, \xi) - \lambda$ belong to the closure of $\mathscr{P} + \tilde{\mathscr{P}}$. By Lemmas 1.3 and 1.4, this implies that $j(\xi)$ is positive on $\mathscr{B}$ and achieves its minimum on a point ξ_0 of $\mathscr{B}$.

Let η belong to $\mathscr{L}^+$. Thus $\eta + \xi_0$ belongs to $\mathscr{B}$ (since $(f, \eta + \xi_0) \geqq (f, \xi_0) \geqq \sigma_{\mathscr{B}}(f)$ for any $f \in \mathscr{L}$). Therefore $(g, \eta) = j(\eta + \xi_0) - j(\xi_0)$ is positive on $\mathscr{L}^+$. This amounts to saying that g belongs to $\mathscr{L}^{+\,+}$. But since $\mathscr{L}$ is closed, $\mathscr{L} = \mathscr{L}^{+\,+}$ (by Lemma 1.3 with $\mathscr{F} = \mathscr{X}$ and $\mathscr{P} = 0$). Therefore $g = f_0$ belongs to $\mathscr{L}$. On the other hand,

$$0 \leqq \inf_{\xi \in \mathscr{B}} j(\xi) = \inf_{\xi \in \mathscr{B}} [(f_0, \xi) - \lambda] = \sigma_{\mathscr{B}}(f_0) - \lambda. \tag{1.24}$$

Therefore j belongs to $\mathscr{P} + \mathscr{L}$ since

$$j(\xi) = (g, \xi) - \lambda \geqq (f_0, \xi) - \sigma_{\mathscr{B}}(f_0) \qquad \text{for any } \xi \in \mathscr{L}. \tag{1.25}$$

1.3. Applications to the characterization of stationary solutions. Let us recall that a functional j is directionally differentiable at ξ_0 if the following limit exists for any $\xi \in \mathscr{X}$:

$$Dj(\xi_0)(\xi) = \lim_{\theta \to 0+} \frac{j(\xi_0 + \theta\xi) - j(\xi_0)}{\theta}. \tag{1.26}$$

A convex functional is directionally differentiable and $Dj(\xi_0)(\xi)$ is a convex positively homogeneous functional of ξ.

If $\mathscr{X}$ is a topological vector space, we say that j is Gâteaux differentiable if $Dj(\xi_0)$ is actually a continuous linear functional. Let us assume now that $\mathscr{B}$ is convex and that

$$\xi_0 \in \mathscr{B}, \qquad j(\xi_0) = \inf_{\xi \in \mathscr{B}} j(\xi). \tag{1.27}$$

It is classical to check that ξ_0 is a "stationary solution," i.e., satisfies

$$\xi_0 \in \mathscr{B}, \qquad 0 = \inf_{\xi \in \mathscr{B}} Dj(\xi_0)(\xi - \xi_0). \tag{1.28}$$

(The converse is true if j is convex.) Let $\mathscr{L}$ be a representation of $\mathscr{B}$. We deduce the following corollary from Theorem 1.1.

COROLLARY 1.2. *Let us assume that j is directionally differentiable on $\mathscr{B}$, and that $Dj(\xi_0)(\eta)$ is convex with respect to η. Let us assume that* (1.3) *holds when $\mathscr{C}$ is the cone of convex and positively homogeneous functionals. Therefore* (1.28) *is equivalent to: There exists $f_0 \in \mathscr{L}$ such that*

$$Dj(\xi_0)(\xi - \xi_0) \geqq f_0(\xi) \qquad \textit{for any } \xi \in \mathscr{X}. \tag{1.29}$$

If f_0 is directionally differentiable at ξ_0, this implies that

$$Dj(\xi_0) \geqq Df_0(\xi_0), \qquad f_0(\xi_0) = 0. \tag{1.30}$$

In particular, Theorem 1.3 implies the above corollary whenever $\mathscr{X}$ is a barreled space and $\mathscr{L}$ an affine representation of $\mathscr{B}$. On the other hand, we deduce the following corollary from Theorem 1.4.

COROLLARY 1.3. *Let us assume j Gâteaux differentiable on $\mathscr{B}$ and $\mathscr{B}$ defined by constraints $(f, \xi) - \sigma_{\mathscr{B}}(f)$, where f belongs to a closed convex cone $\mathscr{L}$ of $\mathscr{X}'$. Therefore,* (1.28) *is equivalent to: There exists $f_0 \in \mathscr{L}$ such that*

$$Dj(\xi_0) = f_0, \qquad \sigma_{\mathscr{B}}(f_0) = (f_0, \xi_0). \tag{1.31}$$

We shall devote the second part of this paper to the application of this corollary.

2. The maximum principle for optimal control problems.

2.1. Abstract boundary value operators and their adjoints. The differential problems with boundary conditions and (or) initial conditions can be embedded in the following abstract framework.

We introduce Banach spaces U, E and Φ and two operators $\Lambda \in \mathscr{L}(U, E)$ and $\alpha \in \mathscr{L}(U, \Phi)$ such that

(2.1) $\Lambda \times \alpha$ is an isomorphism from U onto $E \times \Phi$.

Example. For instance, Λ is a differential operator

$$\Lambda u = \sum_{|p|, |q| \leqq m} (-1)^{|q|} D^q (a_{pq}(x) D^p u)$$

defined on a space U of functions $u(x)$ on a bounded open set Ω or R^n. If $B^j u$ is a differential operator of order m_j, $0 \leqq m_j \leqq 2m - 1$, on the boundary Γ of Ω, we shall take

$$\alpha u = (B^0 u, \cdots, B^j u, \cdots, B^{m-1} u).$$

Under convenient assumptions (ellipticity), the assumption (2.1) is satisfied for several choices of spaces U, E and Φ of functions or distributions (see § 2.4 for a precise example and [8]). Another example is given by "parabolic" equations. We take

$$\Lambda u(x, t) = \frac{\partial u}{\partial t} - \sum_{|p|, |q| \leqq m} (-1)^{|q|} D^q (a_{pq}(x) D^p u)$$

and α will map $u(x, t)$ into the sequence

$$\alpha u = (u(x, 0), B^0 u(x, t), \cdots, B^{m-1} u(x, t)).$$

Here again, (2.1) is satisfied for several choices of spaces U, E, Φ under suitable assumptions. (See [8].)

We shall need the construction of an adjoint operator $\Lambda^* \times \beta^*$. To fulfill this purpose, we assume the following:

(i) There exists an operator β mapping U onto a Banach space Ψ such that $\alpha \times \beta$ is a right-invertible operator from U onto $\Phi \times \Psi$.

Let us set

(2.2) $$U_0 = \{u \in U \text{ such that } \alpha u = \beta u = 0\}.$$

We finally assume:

(ii) There exists a Banach space H in which U and U_0 are dense with a stronger topology.

Let us denote by $\Lambda^* \in \mathscr{L}(E', U_0')$ the "formal adjoint" of Λ defined by

(2.3) $$(\Lambda^* u, v) = (u, \Lambda v) \quad \text{for any } u \in E' \quad \text{and } v \in U_0$$

and by U^* its "domain",

(2.4) $$U^* = \{u \in E' \text{ such that } \Lambda^* u \in H'\}.$$

(Indeed, by (ii), H' can be identified with a subspace of U_0'.) We thus can prove the existence of unique operators $\alpha^* \in \mathscr{L}(U^*, \Phi')$ and $\beta^* \in \mathscr{L}(U', \Psi')$ such that the following "Green's formula" holds:

(2.5) $$(\Lambda^* u, v) - (u, \Lambda v) = \langle \beta^* u, \beta v \rangle - \langle \alpha^* u, \alpha v \rangle \quad \text{for any } u \in U^* \text{ and } v \in U.$$

We then state the following theorem (see [1]).

THEOREM 2.1. *Let f belong to H' and ψ belong to Ψ'. Assume* (i) *and* (ii). *Therefore any solution $(u, \varphi) \in E' \times \Phi'$ of the transposed equation*

$$\Lambda' u + \alpha' \varphi = (\Lambda \times \alpha)'(u, \varphi) = f + \beta' \psi \tag{2.6}$$

is a solution of the "adjoint problem"

$$u \in U^*, \quad \Lambda^* u = f, \quad \beta^* u = \psi \quad \text{and} \quad \varphi = \alpha^* u, \tag{2.7}$$

and conversely. Moreover, if (2.1) *holds, $\Lambda^* \times \beta^*$ is an isomorphism from U^* onto $H' \times \Psi'$.*

Let us sketch in a few words the proof of (2.5) and of Theorem 2.1. Let $\Lambda' \in \mathscr{L}(E', U')$ be the transpose of Λ. Therefore, by (2.3), when u belongs to U^*, $\Lambda^* u - \Lambda' u$ belongs to the orthogonal $U_0^\perp$ of U_0 in U'. On the other hand, by (i) and (2.2), $\alpha' + \beta' = (\alpha \times \beta)'$ is an isomorphism from $\Phi' \times \Psi'$ onto its closed range $U_0^\perp = \alpha' \Phi' + \beta' \Psi'$. Thus we can write this formula in a unique way:

$$\Lambda^* u - \Lambda' u = \alpha' \alpha^* u - \beta' \beta^* u, \tag{2.8}$$

which is equivalent to (2.5). Using (2.5), we see that (2.7) implies (2.6). On the other hand, let u be a solution of (2.6), or equivalently, of

$$(u, \Lambda v) + \langle \varphi, \alpha v \rangle = (f, v) + \langle \psi, \beta v \rangle \quad \text{for any } v \text{ in } U. \tag{2.9}$$

When v ranges over U_0, we deduce that $\Lambda^* u = f$ and that u belongs to U^* (since f belongs to H'). Therefore we can use (2.5) and we deduce that $\langle \alpha^* u - \varphi, \alpha v \rangle - \langle \beta^* u - \psi, \beta v \rangle = 0$ for any v in U. Therefore (i) implies that $\alpha^* u = \varphi$ and $\beta^* u = \psi$.

2.2. Optimal control problems and the maximum principle. We shall deduce from Corollary 1.3 the construction of the optimality equations for a problem studied (in the quadratic case) in [7], [6] by other methods.

Let us consider an operator $\Lambda \times \alpha$ mapping a space U of "states" into a space $E \times \Phi$ and let us assume (2.1), (i) and (ii). We introduce a space K of "controls" k and we consider the equations

$$\Lambda u = \bar{f} + Bk, \tag{2.10a}$$

$$\alpha u = \bar{\varphi} + Ck, \tag{2.10b}$$

where $\bar{f}$ is given in E, $\bar{\varphi}$ in Φ and where B and C are linear operators mapping K into E and Φ respectively. Actually, the controls k are required to obey the following affine constraints: Let D be a linear operator from K into a Banach space Z, P be a closed convex cone of Z, $\bar{z}$ a given element of Z and

$$K_{ad} = \{k \in K \text{ such that } Dk - \bar{z} \in P\}. \tag{2.11}$$

Finally, we introduce the following "cost function":

$$j(u, k) = j_1(u - \bar{u}) + j_2(\beta u - \bar{\psi}) + j_2(k - \bar{k}), \tag{2.12}$$

where

(2.13a) j_1 is a Gâteaux differentiable functional defined on $H \supset U$, $\bar{u} \in U$;

(2.13b) j_2 is a Gâteaux differentiable functional defined on Ψ, $\bar{\psi} \in \Psi$;

(2.13c) j_3 is a Gâteaux differentiable functional defined on K, $\bar{k} \in K$.

Let us set

$$J_i = Dj_i \qquad i = 1, 2, 3, \tag{2.14}$$

where J_1 maps H into H', J_2 maps Ψ into Ψ' and J_3 maps K into K',

$$\sigma \quad \text{is the support functional of } K_{ad}, \tag{2.15}$$

and let us denote by $\mathscr{B}$ the subset of elements $(u, k) \in U \times K$ satisfying

$$k \in K_{ad}, \quad \Lambda u = \bar{f} + Bk, \quad \alpha u = \bar{\varphi} + Ck. \tag{2.16}$$

We say that $(u_0, k_0) \in U \times K$ is a solution of the optimal control problem if and only if

$$(u_0, k_0) \in \mathscr{B} \quad \text{and} \quad j(u_0, k_0) \leqq j(u, k) \quad \text{for any } (u, k) \in \mathscr{B} \tag{2.17}$$

and that $(u_0, v_0, k_0) \in U \times U^* \times K$ is a solution of the maximum principle problem (or of the optimality equations) if and only if

$$\Lambda u_0 = \bar{f} + Bk_0, \qquad \Lambda^* v_0 = J_1(u_0 - \bar{u}); \tag{2.18a}$$

$$\alpha u_0 = \bar{\varphi} + Ck_0, \qquad \beta^* v_0 = J_2(\beta u_0 - \bar{\psi}); \tag{2.18b}$$

$$Dk_0 - \bar{z} \in P, \qquad J_3(k_0 - \bar{k}) + (B' + C'\alpha^*)v_0 \in D'P^+; \tag{2.18c}$$

$$\sigma(J_3(k_0 - \bar{k}) + (B' + C'\alpha^*)v_0) = (J_3(k_0 - \bar{k}) + (B' + C'\alpha^*)v_0, k_0). \tag{2.18d}$$

From Corollary 1.3 and Theorem 2.1 we deduce the following theorem.

THEOREM 2.2. *Let us assume* (2.1), (i), (ii), (2.13) *and*

$$C'\Phi' + B'E' - DP^+ \quad \textit{is weakly closed in } K'. \tag{2.19}$$

If (u_0, k_0) *is a solution of the optimal control problem* (2.17), *there exists* $v_0 \in U^*$ *such that* (u_0, v_0, k_0) *is a solution of the maximum principle problem* (2.18). *Conversely, if the functionals* j_1, j_2 *and* j_3 *are convex and if* (u_0, v_0, k_0) *is a solution of* (2.18), (u_0, k_0) *is a solution of* (2.17).

We can give several examples where (2.19) is fulfilled. For instance, if one of the operators C' and B' is surjective, (2.13) holds. If the Banach spaces are reflexive, if C, B and D have a closed range and if P is actually a subspace, (2.19) holds by the closed range theorem.

The case where $K_{ad} = K$ (case without "constraints") occurs when we choose $Z = K = P$, $D = 1$, $\bar{z} = 0$. Therefore, we replace (2.18c) and (2.18d) by

$$J_3(k_0 - \bar{k}) + (B' + C'\alpha^*)v_0 = 0. \tag{2.20}$$

2.3. Proof of Theorem 2.2. Let us set $\mathscr{X} = U \times K$, $\xi = (u, k)$. The convex set $\mathscr{B}$ is defined by the constraints

$$(\Lambda u - Bk, v) - (\bar{f}, v) = 0 \quad \text{for any } v \text{ in } E', \tag{2.21a}$$

$$\langle \alpha u - Ck, \varphi \rangle - \langle \bar{\varphi}, \varphi \rangle = 0 \quad \text{for any } \varphi \text{ in } \Phi', \tag{2.21b}$$

$$(Dk, z) - (\bar{z}, z) \geqq 0 \quad \text{for any } z \in P^+. \tag{2.21c}$$

We can write (2.21) in the form

$$(\Lambda' E' + \alpha' \Phi') \times (D' P^+ - B' E' - C' \Phi') = U' \times (D' P^+ - B' E' - C' \Phi') \tag{2.22}$$

by (2.1). Therefore the assumption (2.19) implies that the cone $\mathscr{L}$ is weakly closed in

$$U' \times K' = \mathscr{X}'. \tag{2.23}$$

On the other hand, if $(u_0, k_0) \in U \times K$, the derivative of j is equal to

$$Dj(u_0, k_0) = (J_1(u_0 - \bar{u}) + \beta' J_2(\beta u_0 - \bar{\psi}), J_3(k_0 - \bar{k})) \in U' \times K'. \tag{2.24}$$

By Corollary 1.3, we know that there exists

$$f_0 = (\Lambda' v_0 + \alpha' \varphi_0, D' z_0 - C' \varphi_0 - B' v_0) \in \mathscr{L} \tag{2.25}$$

such that

$$\Lambda' v_0 + \alpha' \varphi_0 = J_1(u_0 - \bar{u}) + \beta' J_2(\beta u_0 - \bar{\psi}), \tag{2.26a}$$

$$D' z_0 - B' v_0 - C' \varphi_0 = J_3(k_0 - \bar{k}). \tag{2.26b}$$

Since $J_1(u_0 - \bar{u})$ belongs to H' and $J_2(\beta u_0 - \bar{\psi})$ belongs to Ψ', we deduce from Theorem 2.1 that

$$v_0 \in U^*, \quad \Lambda^* v_0 = J_1(u - \bar{u}), \quad \beta^* v_0 = J_2(\beta u_0 - \bar{\psi}) \tag{2.27}$$

and that $\varphi_0 = \alpha^* v_0$. Therefore, since $z_0 \in P^+$, we deduce that

$$D' z_0 = J_3(k_0 - \bar{k}) + B' v_0 + C' \alpha^* v_0 \in D' P^+. \tag{2.28}$$

Let us now prove (2.18d).

If k belongs to K_{ad}, there exists a unique solution u of (2.10) (by (1.1)) and therefore $\xi = (u, k)$ belongs to $\mathscr{B}$. Then $(Dj(\xi_0), \xi - \xi_0) \geqq 0$ for any $k \in K_{ad}$. Writing this inequality and using (2.27) and (2.28) we obtain

$$\begin{aligned} (\Lambda' v_0 + \alpha' \alpha^* v_0, u - u_0) - (B' v_0 + C' \alpha^* v_0, k - k_0) + (D' z_0, k - k_0) \\ = (D' z_0, k - k_0) \geqq 0 \qquad \text{for any } k \in K_{ad}. \end{aligned} \tag{2.29}$$

We thus deduce that

$$\sigma(D' z_0) = (D' z_0, k_0) = \inf_{k \in K_{ad}} (D' z_0, k), \tag{2.30}$$

where σ is the support functional of K_{ad}.

2.4. Example. Let us consider the following optimal control problem where the state u is given through an elliptic equation and where the control k appears in the boundary conditions.

Namely, let Ω be a smooth bounded open set of R^n, Γ its boundary. If $\vec{f}$ is a given function of Ω, $\bar{\varphi}$ a given function on Γ and if the controls k are functions defined on Γ, the constraints of the optimal control problem will be defined by

$$-\Delta u + u = \vec{f} \quad \text{on } \Omega, \tag{2.31a}$$

$$u|_\Gamma = \bar{\varphi} + k \quad \text{on } \Gamma, \tag{2.31b}$$

$$k(x) \geqq 0 \quad \text{on } \Gamma. \tag{2.31c}$$

We shall minimize the following functional:

$$j(u_0, k_0) \leqq j(u, k) \tag{2.32}$$

where

$$j(u, k) = \frac{1}{p}\left[\int_\Gamma \left|\frac{\partial u}{\partial n} - \bar{\psi}\right|^p d\sigma + \int_\Gamma |k|^p \, d\sigma\right], \qquad 1 < p < +\infty,$$

where $d\sigma$ is the superficial measure on Γ.

We shall deduce from Theorem 2.2 that when $\vec{f}$, $\bar{\varphi}$, $\bar{\psi}$ and $\bar{k}$ belong to suitable spaces of functions, the optimality equations are

$$-\Delta u_0 + u_0 = \vec{f}, \quad -\Delta v_0 + v_0 = 0, \tag{2.33a}$$

$$k_0 = u_0|_\Gamma - \bar{\varphi}, \quad v_0|_\Gamma = \left|\frac{\partial u_0}{\partial n} - \bar{\psi}\right|^{p-2}\left(\frac{\partial u_0}{\partial n} - \bar{\psi}\right), \tag{2.33b}$$

$$u_0|_\Gamma \geqq \bar{\varphi}, \quad (u_0|_\Gamma - \bar{\varphi})^{p-1} + \frac{\partial v_0}{\partial n} \geqq 0, \tag{2.23c}$$

$$(u_0 - \bar{\varphi})\left((u_0 - \bar{\varphi})^{p-1} + \frac{\partial v_0}{\partial n}\right) = 0 \quad \text{a.e. on } \Gamma. \tag{2.33d}$$

To prove the equivalence of those two problems, we have to define the spaces U, E, H, Φ and Ψ and to verify that the assumptions (2.1), (i) and (ii) are satisfied.

We shall choose, for instance (see [16], [8] and the references of [8]),

$$E = H = L^p(\Omega), \qquad U = W^{2,p}(\Omega), \tag{2.34}$$

where $1 < p < +\infty$ and where $W^{2,p}(\Omega)$ denotes the subspace of functions u of $L^p(\Omega)$ such that the weak derivatives of order $\leqq 2$ belong to $L^p(\Omega)$. Let us set

$$\alpha u = u|_\Gamma \quad \text{and} \quad \beta u = \frac{\partial u}{\partial n}. \tag{2.35}$$

These two operators map $W^{2,p}(\Omega)$ into $L^p(\Gamma)$ and the assumption (i) is satisfied when we take

$$\Phi = W^{2-1/p,p}(\Gamma), \qquad \Psi = W^{1-1/p,p}(\Gamma). \tag{2.36}$$

If we choose $\Lambda u = -\Delta u + u$, it is known that (2.1) is satisfied.

When Ω is smooth enough, the space U_0 coincides with the closure $W_0^{2,p}(\Omega)$ in $W^{2,p}(\Omega)$ of the functions with compact support. Therefore the assumption (ii)

is satisfied. By using (2.5), the usual Green's formula

$$\int_\Omega (-\Delta u)v\,dx - \int_\Omega u\cdot(-\Delta v)\,dx = \int_\Gamma u\cdot\frac{\partial v}{\partial n}\,d\sigma - \int_\Gamma \frac{\partial u}{\partial n}\cdot v\,d\sigma \tag{2.37}$$

can be extended when v belongs to $W^{2,p}(\Omega)$ and u to the space

$$U^* = \{u\in L^{p'}(\Omega) \text{ such that } \Delta u\in L^{p'}(\Omega)\}, \tag{2.38}$$

where $1/p + 1/p' = 1$. Therefore we can set

$$\Lambda^* u = -\Delta u + u,\quad \alpha^* u = \partial u/\partial n,\quad \beta^* u = u|_\Gamma. \tag{2.39}$$

We now take

$$K = Z = \Phi = W^{2-1/p,p}(\Gamma), \tag{2.40}$$

P the cone of positive functions on Γ, and

$$B = 0,\quad C = 1,\quad D = 1,\quad \bar{k} = 0. \tag{2.41}$$

Therefore the assumption (2.19) is satisfied.

Finally, let us recall that the differential of

$$\frac{1}{p}\int_\Gamma |u(x)|^p\,d\sigma$$

is the operator which associates with u the function $|u|^{p-2}u$. Therefore Theorem 2.1 implies the following corollary.

COROLLARY 2.1. *Let us assume that*

$$f\in L^p(\Omega),\quad \bar{\varphi} \text{ and } k\in W^{2-1/p,p}(\Gamma) \quad \text{and} \quad \bar{\psi}\in L^p(\Gamma). \tag{2.42}$$

Therefore if (u_0, k_0) *is a solution of* (2.31) *and* (2.32), (u_0, v_0, k_0) *is a solution of* (2.33) *and conversely.*

It is classical to check that the solution of (2.31), (2.32) is unique.

2.5. Some remarks on the existence of optimal solutions. We can deduce the existence of an optimal solution of a control problem from the Weirstrass theorem stating that a lower semicontinuous functional achieves its minimum on a compact set.

In the first place, if a pair (u_1, k_1) satisfies (2.16), we have to look for a solution (u_0, k_0) minimizing $j(u, k)$ on the set defined by

$$k\in K_{ad},\quad \Lambda u = \bar{f} + Bk,\quad \alpha u = \bar{\varphi} + Ck, \tag{2.43a}$$

$$j(u,k) \leqq j(u_1, k_1) = r. \tag{2.43b}$$

We thus deduce the following proposition.

PROPOSITION 2.1. *Let us assume the following*:

(i) $j(u, k)$ *is weakly lower semicontinuous on* $U \times K$;

(ii) $U \times K$ *is a reflexive Banach space*; *and*

(iii) *the set of elements defined by* (2.43) *is bounded in* $U \times K$.

Then there exists a solution (u_0, k_0) *of the optimal control problem* (2.17). *If moreover*

the assumptions of Theorem 2.2 *hold, there exists a solution* (u_0, v_0, k_0) *of the maximum principle problem* (2.18).

Indeed, the set defined by (2.43) is bounded and weakly closed, and then, weakly compact.

The assumption (iii) holds if, for instance

(iv) the subset K_{ad} is bounded

or, if

(v) the functional $j_3(k)$ is the norm of the space K, j_1 and j_2 being positive.

Another set of conditions implying (iii) is the following:

(vi) the functional $j_2(\psi)$ is the norm of ψ, j_1 and j_3 being positive;

(vii) $\Lambda \times \alpha$ is an isomorphism from U onto $E \times \Psi$;

(viii) the subset of $k \in K_{ad}$, such that Bk is bounded in E, Ck is bounded in ψ and $j_3(k)$ is bounded, is itself bounded.

Example. Let us consider the example given in § 2.4 where we replace $j(u, k)$ defined by (2.32) by the following functional:

$$j(u,k) = \left\| \frac{\partial u}{\partial n} - \bar{\psi} \right\|^p_{L^p(\Gamma)} + \|k\|^p_{W^{2-1/p,p}(\Gamma)}. \tag{2.44}$$

Then, if we choose $U = W^{2,p}(\Omega)$, $K = \varphi = W^{2-1/p,p}(\Gamma)$, the assumption (v) is satisfied and we obtain existence of a solution (u_0, k_0) of the optimal control problem (2.31), (2.32).

For the sake of simplicity, we gave only the simplest choice of spaces U, E, φ for which the operator $(-\Delta + 1) \times \alpha$ is an isomorphism from U onto $E \times \varphi$. Using the theory of interpolation of Banach spaces (see [8] and the references of this book), we have many other choices. If we associate with any of such choices of a cost function of the following form:

$$j(u,k) = j_2\left(\frac{\partial u}{\partial n} - \bar{\psi}\right) + \|k\|^p_{\Phi}, \tag{2.45}$$

where j_2 is Gâteaux differentiable and weakly lower semicontinuous on Ψ, we shall obtain existence of a solution of the optimal control problem.

If we denote by J_2 the Gâteaux derivative of the functional j_2 and by J_3 the Gâteaux derivative of $\|k\|^p_{\Phi}$, we shall deduce the existence of a solution (u_0, v_0) of the following problem:

$$-\Delta u_0 + u_0 = \bar{f}, \quad -\Delta v_0 + v_0 = 0, \quad v_0|_\Gamma = J_2(\partial u_0/\partial n - \bar{\psi}), \tag{2.46a}$$

$$u_0|_\Gamma - \bar{\varphi} \geqq 0 \quad \text{on } \Gamma, \qquad J_3(u_0|_\Gamma - \bar{\varphi}) + \partial v_0/\partial n \geqq 0 \quad \text{on } \Gamma, \tag{2.46b}$$

$$(u_0|_\Gamma - \bar{\varphi})(J_3(u_0|_\Gamma - \bar{\varphi}) + \partial v_0/\partial n) = 0 \quad \text{on } \Gamma. \tag{2.46c}$$

REFERENCES

[1] J. P. AUBIN, *Abstract boundary value operators and their adjoints*, to appear.

[2] V. F. DEM'YANOV AND A. M. RUBINOV, *Minimization of functionals in normed spaces*, this Journal, 6 (1968), pp. 73–88.

[3] A. Y. DUBOVITSKII AND A. A. MILYUTIN, *Extremum problems with constraints*, Zh. Vychisl. Mat. i Mat. Fiz. 0 (1965), pp. 395–453.

[4] R. V. GAMKRELIDZE, *On some extremal problems in the theory of differential equations*, this Journal, 3 (1965), pp. 106–128.

[5] H. HALKIN AND L. W. NEUSTADT, *General necessary conditions for optimization problems*, Proc. Nat. Acad. Sci. USA, 56 (1966), pp. 1066–1071.

[6] J. L. LIONS, *Sur le contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[7] ———, *Sur le côntrole optimal de systèmes décrits par des équations aux dérivées partielles linéaires. I, II, III*, C. R. Acad. Sci. Paris, 263 (1966), pp. 661–663, pp. 713–715, pp. 776–779.

[8] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes*, Dunod, Paris, 1968.

[9] J. J. MOREAU, *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.

[10] L. W. NEUSTADT, *A general theory of extremals*, U.S. C.E.E. Rep. 297, 1968.

[11] ———, *An abstract variational theory with applications to a broad class of optimization problems*, this Journal, 4 (1966), pp. 505–525.

[12] R. T. ROCKAFELLAR, *Duality and stability in extremum problems involving convex functions*, Pacific J. Math., 21 (1967), pp. 167–187.

[13] ———, *Convex Analysis*, Princeton University Press, Princeton, 1968.

[14] D. L. RUSSELL, *The Kuhn–Tucker conditions in Banach spaces with an application to control theory*, J. Math. Anal. Appl., 15 (1966), pp. 200–212.

[15] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control*, W. B. Saunders, Philadelphia, 1969.

[16] S. AGMON, A. DOUGLIS AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

DAVIDON'S METHOD FOR MINIMIZATION PROBLEMS IN HILBERT SPACE WITH AN APPLICATION TO CONTROL PROBLEMS*

H. TOKUMARU, N. ADACHI AND K. GOTO†

1. Introduction. Many methods are known for finding minimum values of functions on a finite-dimensional space. In the case when we can make use of the gradient of the given function, the steepest descent method and Newton-Raphson method are well known and they are also the foundations for many other modified methods. The steepest descent method has a very simple algorithm and the convergence of the method is assured from any initial approximation. But, the convergence is slow, especially in the neighborhood of the extremum point of the function. On the other hand, the Newton-Raphson method has the quadratic convergence property but the method may not converge at all.

In recent years, several rapidly convergent methods have been proposed. Among these are the conjugate gradient method [1] and Davidon's method [2]. In both methods the direction of the search is determined by the gradient of the function. The stability property of the method is analogous to that of the steepest descent method. The conjugate gradient method is very simple. Davidon's method is more complex, but it is known by experience that its convergence is superior to that of the conjugate gradient method.

The steepest descent method and Newton–Raphson method have been extended to function spaces and applied to control problems by authors such as H. J. Kelly, A. E. Bryson, R. McGill and R. E. Kopp [3]–[6]. L. S. Lasdon, S. K. Mitter and A. D. Waren have extended the conjugate gradient method to function spaces and have applied it to optimum control problems [7].

In this paper, an extension of Davidon's method to Hilbert space is presented. The stability and convergence of the method are shown in the case when the functionals to be minimized are quadratic. The method is applied to optimum control problems and numerical examples are given.

2. Formulation of the problem. Let H be a (real) separable Hilbert space with inner product (f, g), $f, g \in H$. The norm of an element $f \in H$ is defined as $\|f\| = (f, f)^{1/2}$. Let G be a linear self-adjoint operator on H such that

$$m\|f\|^2 \leqq (f, Gf) \leqq M\|f\|^2, \tag{2.1}$$

where

$$M = \sup_{\|f\| \neq 0} \frac{(f, Gf)}{\|f\|^2}, \qquad m = \inf_{\|f\| \neq 0} \frac{(f, Gf)}{\|f\|^2} \tag{2.2}$$

and $0 < m \leqq M$. Then the norm of G is equal to M:

$$\|G\| = M. \tag{2.3}$$

* Received by the editors August 21, 1968, and in revised form March 20, 1969.

† Department of Applied Mathematics and Physics, Faculty of Engineering, Kyoto University, Kyoto, Japan.

Since M is finite, G is a continuous operator. From the condition (2.1), an inequality

$$\|Gf\| \geqq m\|f\| \tag{2.4}$$

holds. The inequality is a necessary and sufficient condition for the inverse operator G^{-1} of the self-adjoint operator G to be defined. The inverse G^{-1} satisfies the inequalities

$$\begin{aligned} &\|G^{-1}g\| \leqq \frac{1}{m}\|g\|, \\ &\|G^{-1}g\| \geqq \frac{1}{M}\|g\|, \qquad g \in H. \end{aligned} \tag{2.5}$$

Using Schwarz's inequality, we prove that G^{-1} satisfies the inequalities

$$\frac{1}{M}\|g\|^2 \leqq (g, G^{-1}g) \leqq \frac{1}{m}\|g\|^2. \tag{2.6}$$

Since G is self-adjoint, G^{-1} is also a self-adjoint operator. Let $S(f)$ be a Fréchet differentiable functional on H. We call the operator F, which is defined by the formula

$$\lim_{t \to 0} \frac{1}{t}\{S(f + th) - S(f)\} = (F(f), h) \tag{2.7}$$

for any $h \in H$, the gradient of the functional S: $F = \operatorname{grad} S$ (see [8]).

PROBLEM. Let $S(u) = \frac{1}{2}(u - u^*, G(u - u^*))$ be a quadratic form on H, where G is a linear self-adjoint operator satisfying the condition (2.1). Find the u^* which minimizes the functional $S(u)$.

By the definition of the gradient of functionals, grad $S(u)$ exists and is defined by

$$\operatorname{grad} S(u) = G(u - u^*). \tag{2.8}$$

The gradient of $S(u)$ is denoted by $g(u)$:

$$g(u) = G(u - u^*). \tag{2.9}$$

Then the solution of the above problem is given by

$$u^* = u - G^{-1}g(u). \tag{2.10}$$

In other words, u^* is obtained directly if we can make use of the gradients $g(u)$ and the operator G^{-1}. But, in this problem, we assume that G^{-1} cannot be evaluated directly.

3. Algorithm of Davidon's method. Let the ith approximation of the solution of the problem be u^i; then the $(i + 1)$st approximation is determined as follows using the gradient at u^i.

Define $p^i \in H$ as

$$p^i = -K^i g^i,$$

where g^i is the gradient at u^i;

$$g^i = G(u^i - u^*),$$

and K^i is an operator from H to H such that

$$(f, K^i f) > 0 \quad \text{for } f \in H \quad \text{and } f \neq 0.$$

Then u^{i+1} is given by

$$u^{i+1} = u^i + \alpha^i p^i,$$

where α^i is a constant which minimizes a function of α;

$$S(u^i + \alpha p^i) = S(u^i) + \alpha(p^i, g^i) + \frac{\alpha^2}{2}(p^i, Gp^i).$$

By the definition,

$$\alpha^i = -\frac{(p^i, g^i)}{(p^i, Gp^i)}. \tag{3.1}$$

The operator K^i is modified at each step so that p^i becomes an eigenelement of $K^{i+1}G$. The algorithm of the computation is given as follows:

(i) Choose an initial estimation u^0 and identify K^0 with an identity operator I; $K^0 = I$.

(ii) Evaluate the gradient g^i at u^i.

(iii) Set $p^i = -K^i g^i$.

(iv) Set $u^{i+1} = u^i + \alpha^i p^i$.

Here, α^i is a constant such that

$$S(u^i + \alpha^i p^i) = \min_{\alpha} S(u^i + \alpha p^i).$$

(v) Set $y^i = (g^{i+1} - g^i)/\alpha^i$.

(vi) Set $q^i = K^i y^i$.

(vii) Set

$$p_N^i = \frac{p^i}{\sqrt{(p^i, y^i)}}, \qquad q_N^i = \frac{q^i}{\sqrt{(q^i, y^i)}}.$$

(viii) Define an operator K^{i+1} as follows:

$$K^{i+1} f = K^i f + (f, p_N^i) p_N^i - (f, q_N^i) q_N^i,$$

where f is an arbitrary element of H.

(ix) Set $i = i + 1$ and repeat (ii)–(viii).

By the definition of g^i,

$$\begin{aligned} g^{i+1} &= G(u^{i+1} - u^*) \\ &= G(u^i - u^* + \alpha^i p^i) \\ &= g^i + \alpha^i G p^i \end{aligned} \tag{3.2}$$

Hence, from (v),

$$y^i = Gp^i. \tag{3.3}$$

Substituting (3.3) into (p^i, y^i), and considering positivity of G, we show that $(p^i, y^i) > 0$. In the following section we shall show that $(q^i, y^i) = (K^i y^i, y^i) > 0$. Therefore the vectors p_N^i, q_N^i are well-defined, so that the operator K^i is also defined.

Now, suppose that $S(u)$ is not necessarily quadratic. By the definition of α^i,

$$(p^i, g^{i+1}) = \frac{\partial}{\partial \alpha} S(u^i + \alpha p^i)\Big|_{\alpha=\alpha^i} = 0. \tag{3.4}$$

Then,

$$(p^i, y^i) = \frac{1}{\alpha^i}\{(p^i, g^{i+1}) - (p^i, g^i)\} = \frac{1}{\alpha^i}(K^i g^i, g^i).$$

On assuming the positivity of K^i, we have

$$(p^i, g^i) = -(K^i g^i, g^i) < 0 \quad \text{for } g^i \neq 0,$$

so that $\alpha^i > 0$ and $(p^i, y^i) > 0$ if $g^i \neq 0$. Hence, if K^i is positive also for nonquadratic functionals, vectors p_N^i and q_N^i are defined as well. The positivity of K^i in the case of the nonquadratic form is noted in Remark 1 in the following section.

4. Stability of the scheme. In this section, we shall show that the value of the functional to be minimized decreases at each step with this method.

The following two lemmas (Lemmas 1–2) are direct extensions of the results in [2], and the proofs formally follow proofs in the reference.

LEMMA 1. *K^i is a linear self-adjoint, positive operator and $(f, K^i f) = 0$ only if $f = 0$, $i = 1, 2, \cdots$.*

Proof. We shall prove the lemma by induction. Since $K^0 = I$, the assertion is trivial for $i = 0$. Assume that the lemma is valid for $i = 1, 2, \cdots, n$; we shall now prove that the statement holds for $i = n + 1$. From (viii) it is clear that K^{n+1} is a linear self-adjoint operator. Hence, it is sufficient to show positivity of K^{n+1}. From the relations (vi)–(viii),

$$(f, K^{n+1} f) = (f, K^n f) + (f, p_N^n)^2 - (f, q_N^n)^2 = \frac{(f, K^n f)(y^n, K^n y^n) - (f, K^n y^n)^2}{(y^n, K^n y^n)} + (f, p_N^n)^2.$$

Since K^n is a positive operator, inequalities

$$(f, K^n f)(y^n, K^n y^n) \geqq (f, K^n y^n)^2$$

hold by Schwarz's inequality. Therefore the first term of the right-hand side of the above equality is nonnegative, and the second term is clearly nonnegative. The

first term becomes zero only if f is a scalar multiple of y^n:

$$f = ay^n$$
$$= \frac{a}{\alpha^n}(g^{n+1} - g^n),$$

where a is an arbitrary constant. From this fact and the relation (3.4), $(f, K^{n+1}f)$ vanishes if and only if $(g^n, p^n) = 0$. But this contradicts the positiveness of K^n. Hence, K^{n+1} is a positive operator and the lemma is proved.

Remark 1. In the above proof the quadratic property of the functional $S(u)$ is not used. Therefore the assertion of Lemma 1 is valid also in nonquadratic functionals.

LEMMA 2. *The relations*

$$(p_N^i, Gp_N^j) = \delta_{ij}, \qquad i < k, \quad j < k, \tag{4.1}$$

$$K^k G p_N^i = p_N^i, \qquad i < k, \quad i = 1, 2, \cdots, \tag{4.2}$$

hold, where δ_{ij} *is Kronecker's symbol.*

Proof. From (vi) and (viii),

$$K^{i+1}y^i = K^i y^i + (y^i, p_N^i)p_N^i - (y^i, q_N^i)q_N^i$$
$$= K^i y^i + p^i - K^i y^i = p^i.$$

Hence

$$K^{i+1}Gp^i = p^i \tag{4.3}$$

by (3.3). The statement of the lemma is satisfied for $k = 1$ by (4.3). Assume that the relations (4.1) and (4.2) are satisfied for $k = n$. From (2.8),

$$g^n = g^{i+1} + \sum_{j=i+1}^{n-1} \alpha^j G p^j, \qquad 0 \leqq i < n. \tag{4.4}$$

From relations (3.1) and (3.2),

$$(p^i, g^{i+1}) = 0, \qquad i = 0, 1, \cdots, n. \tag{4.5}$$

Hence, from (4.4), (3.3) and (4.1) with $k = n$,

$$(p^i, g^n) = (p^i, g^{i+1}) = 0. \tag{4.6}$$

Therefore

$$(K^n G p^i, g^n) = (Gp^i, K^n g^n) = 0,$$

since (4.2) holds for $k = n$. Substituting $p^i = -K^i g^i$, we obtain a formula

$$(Gp^i, p^n) = 0, \qquad 0 \leqq i < n. \tag{4.7}$$

Now, by the self-adjointness of K^n and G^n,

$$(K^n y^n, Gp^i) = (y^n, K^n G p^i)$$
$$= (Gp^n, p^i), \qquad 0 \leqq i < n, \tag{4.8}$$

taking into consideration relation (3.3), (4.2) and (4.7). By using this result it is

simple to prove the equalities

$$\begin{aligned} K^{n+1}Gp^i &= K^nGp^i \\ &= p^i, \end{aligned} \qquad 0 \leqq i < n, \tag{4.9}$$

by the definition of K^{n+1}. The relations (4.3), (4.7) and (4.9) show that the statement in the lemma holds for $k = n + 1$.

LEMMA 3. *Let $\varphi_i \in H$ be a complete system in H, which satisfies conditions*

(i) $(\varphi_i, G\varphi_j) = \delta_{ij}$,

(ii) $m\|f\|^2 \leqq (f, Gf) \leqq M\|f\|^2, m, M > 0$.

Then, for any element $f \in H$, the following equalities hold:

$$\begin{aligned} f &= \sum_{i=0}^{\infty} (f, \varphi_i)G\varphi_i \\ &= \sum_{i=0}^{\infty} (f, G\varphi_i)\varphi_i. \end{aligned}$$

Proof. Denote $(f, G\varphi_i)$ by d_i; then, we have the inequalities

$$M\|f\|^2 \geqq (f, Gf) \geqq \sum_{1}^{n} d_i^2$$

since

$$\left(f - \sum_{0}^{n} d_i\varphi_i, G\left(f - \sum_{0}^{n} d_i\varphi_i\right)\right) = (f, Gf) - \sum_{1}^{n} d_i^2 \geqq 0.$$

Define f_n as

$$f_n = \sum_{0}^{n} d_i\varphi_i;$$

then

$$\begin{aligned} 0 \leqq m\|f_n - f_l\|^2 &\leqq (f_n - f_l, G(f_n - f_l)) \\ &= \sum_{l}^{n} d_i^2, \end{aligned} \qquad n \geqq l.$$

The right-hand side of the equality tends to zero as l and n tend to infinity. Therefore, there exists an element $\varphi \in H$ such that $f_n \to \varphi$ as $n \to \infty$. The element is expressed as

$$\varphi = \sum_{1}^{\infty} d_i\varphi_i.$$

By the conditions of the lemma,

$$(f - \varphi, G\varphi_i) = d_i - d_i = 0, \qquad i = 0, 1, \cdots.$$

Since $\{\varphi_i\}$ is a complete system, the equalities mean that φ is identical with f. Using G^{-1} in place of G in the above discussions, the last part of the lemma can be proved. This completes the proof.

This lemma asserts that if $\{\varphi_i\}$ is a complete system, then $\{G\varphi_i\}$ is also a complete system.

We introduce a well-known property with respect to an increasing sequence of self-adjoint operators.

LEMMA 4. *Let $\{U_i\}$ be an increasing sequence of positive self-adjoint operators such that*

$$\sup_n \|U_n\| < A < \infty.$$

Then, there is a linear operator U such that $Uf = \lim_{n\to\infty} U_n f$ for any $f \in H$, and $\|U\| \leqq A$.

Remark 2. By an increasing sequence of operators we mean a system of operators $\{U_n; n = 0, 1, 2, \cdots\}$ such that

$$(f, U_n f) \leqq (f, U_{n+1} f)$$

for an arbitrary $f \in H$ and for $n = 0, 1, 2, \cdots$.

We shall prove the following theorem, using the above lemmas.

THEOREM 1. *The sequence of operators $\{K^i\}$ is uniformly bounded and converges on H to a linear operator K.*

Proof. Denote by A_n, $n = 0, 1, 2, \cdots$, an operator such that

$$A_n f = \sum_{i=0}^{n} (f, p_N^i) p_N^i \quad \text{for } f \in H.$$

The elements $p_N^i, i = 1, 2, \cdots, n$, satisfy the conditions of Lemma 3. Add a sequence $r^i, i = -1, -2, \cdots$, to p_N^i so that a system of elements $\{r^i, p_N^j: i = -1, \cdots, -n, \cdots, j = 0, 1, 2, \cdots, n, \cdots\}$ becomes a complete system satisfying the condition of Lemma 3. Then for any $f \in H$,

$$G^{-1} f = \sum_{i=0}^{n} (f, p_N^i) p_N^i + \sum_{i=-\infty}^{-1} (f, r^i) r^i$$

by Lemma 3, so that

$$(f, G^{-1} f) = \sum_{0}^{n} (f, p_N^i)^2 + \sum_{-\infty}^{-1} (f, r^i)^2 \geqq \sum_{0}^{n} (f, p_N^i)^2.$$

The right-hand side of this inequality is equal to $(f, A_n f)$. Therefore,

$$M' \|f\|^2 \geqq (f, A^n f);$$

in other words, $\|A_n\| \leqq M'$ where $M' = 1/m$. $\{A_n\}$ is an increasing sequence of positive self-adjoint operators by the definition of A_n.

Therefore, by Lemma 4, there exists a linear operator A such that

$$Af = \lim_{n\to\infty} A_n f, \qquad f \in H,$$

and $\|A\| \leqq M'$

Now, define operators $B_n, n = 0, 1, \cdots$, as

$$B_n f = \sum_{0}^{n} (f, q_N^i) q_N^i.$$

Then the operator K^{n+1} is expressed as

$$K^{n+1} = I + A_n - B_n.$$

Hence, for an arbitrary $f \in H$,

$$(f, K^{n+1}) = \|f\|^2 + (f, A_n f) - (f, B_n f).$$

Since K^{n+1} is a positive operator,

$$(f, B_n f) \leqq \|f\|^2 \leqq f^2 + (f, A_n f) \leqq (M' + 1) f^2.$$

Hence B_n is bounded;

$$\|B_n\| \leqq (M' + 1), \qquad n = 0, 1, 2, \cdots.$$

Since B_n is also an increasing sequence, by Lemma 4 there exists an operator B such that

$$Bf = \lim_{n \to \infty} B_n f \quad \text{for } f \in H$$

and

$$\|B\| \leqq M' + 1.$$

Let us define an operator K by

$$K = I + A - B.$$

Then it is clear from the above discussions that K is a linear bounded operator such that

$$Kf = \lim_{n \to \infty} K^n f$$

and

$$\|K\| \leqq 2(M' + 1).$$

Hence, the theorem is proved.

Using Theorem 1, we shall show that the values of the given functional decrease with each step.

THEOREM 2. *With the scheme in* §3,

$$S(u^{i+1}) < S(u^i) \quad \textit{for } g^i \neq 0, \quad i = 0, 1, 2, \cdots.$$

Proof. We shall show that the inner product of the direction of search P^i and the gradient g^i is negative and the step size α^i is positive for every i, $i = 0, 1, 2, \cdots$. Since $p^i = -K^i g^i$ and K^i is positive from Lemma 1,

$$(p^i, g^i) = -(K^i g^i, g^i) < 0 \quad \text{for } g^i \neq 0, \quad i = 0, 1, 2, \cdots.$$

By the definition of α^i,

$$\alpha^i = \frac{(K^i g^i, g^i)}{(p^i, Gp^i)},$$

so $\alpha^i > 0$ for $g^i \neq 0$. From these considerations the statement of the theorem is valid.

5. Convergence of the method. It will be shown in this section that u^i converges to u^* as $i \to \infty$ and that there is a subspace of H on which the sequence of operators K^i converges to G^{-1}.

LEMMA 5. *$K^i y^i \in H$ is expressed as a linear combination of Gp^i, $i = 0, 1, 2, \cdots$.*

For $i = 0$, the assertion of the lemma is valid since $K^0 = I$; $K^0 y^0 = y^0 - Gp^0$. It is assumed that the lemma holds for $K = i$. Then from (viii),

$$K^{i+1} y^{i+1} = y^{i+1} - \sum_{j=0}^{i} \frac{(y^{i+1}, K^j y^i)}{(y^j, K^j y^j)} K^j y^j.$$

Since $y^{i+1} = Gp^{i+1}$ from (3.3), the right-hand side of the above equality is a linear combination of the Gp^j, $j = 0, 1, 2, \cdots, i + 1$.

THEOREM 3. *Let u^i, $i = 0, 1, 2, \cdots$, be a sequence of elements as defined in § 3; then, the sequence converges to u^* as $i \to \infty$.*

Proof. From (3.1),

$$\begin{aligned} S(u^{i+1}) &= S(u^i) - \frac{(K^i g^i, g^i)^2}{(p^i, Gy^i)} \\ &= S(u^i) - \frac{(g^i, K^i g^i)^2}{(K^i g^i, GK^i g^i)}. \end{aligned}$$

Since $S(u^i)$ is bounded and monotone decreasing,

$$\frac{(g^i, K^i g^i)^2}{(K^i g^i, GK^i g^i)} \to 0 \quad \text{as } i \to \infty. \tag{5.1}$$

By Schwarz's inequality,

$$\begin{aligned} (K^i g^i, K^i g^i)^2 &\leqq (K^i g^i, g^i)(K^i(K^i g^i), K^i g^i) \\ &\leqq (K^i g^i, g^i)\|K^i\| \cdot \|K^i g^i\|^2 \\ &\leqq 2(M' + 1)\|K^i g^i\|^2 (K^i g^i, g^i). \end{aligned}$$

Hence,

$$(K^i g^i, g^i)^2 \geqq \frac{\|K^i g^i\|^4}{4(M' + 1)^2}.$$

From the condition (2.1) for G,

$$M\|K^i g^i\|^2 \geqq (K^i g^i, GK^i g^i).$$

Combining the above two inequalities, we have

$$\frac{(g^i, K^i g^i)^2}{(K^i g^i, GKg^i)} \geqq \frac{1}{4M(M' + 1)^2} \|K^i g^i\|^2.$$

Since the left-hand side of this inequality tends to zero from (5.1),

$$\|K^i g^i\| \to 0 \quad \text{as } i \to \infty. \tag{5.2}$$

By (2.6),

$$m'\|K^i g^i\|^2 \leqq (K^i g^i, G^{-1} K^i g^i) \leqq M'\|K^i g^i\|^2,$$

where $M' = 1/m$ and $m' = 1/M$. Hence,

$$(K^i g^i, G^{-1} K^i g^i) \to 0 \quad \text{as } i \to \infty. \tag{5.3}$$

From (3.1), (3.2) and (4.1),

$$\begin{aligned} g^i &= g^0 + \sum_0^{i-1} \alpha^j G p^j \\ &= g^0 - \sum_0^{i-1} (g^j, p_N^j) G p_N^j \\ &= g^0 - \sum_0^{i-1} (g^0 + \sum_0^{i-1} \alpha^k G p^k, p_N^j) G p_N^j \\ &= g^0 - \sum_0^{i-1} (g^0, p_N^j) G p_N^j. \end{aligned} \tag{5.4}$$

On the other hand, by (viii) and (4.2),

$$K^i g^i = g^i - \sum_{j=0}^{i-1} \frac{(g^i, K^j g^j)}{(y^j, K^j y^j)} K^j g^j.$$

The second term of the right-hand side of this equality is a linear combination of the $Gp^k, k = 0, 1, \cdots, i-1$. Hence,

$$K^i g^i = g^i - \sum_{j=0}^{i-1} \beta_j G p_N^j, \tag{5.5}$$

where the $\beta_j, j = 0, 1, \cdots, i-1$, are appropriate constants. By Lemma 2,

$$\begin{aligned} (K^i g^i, G^{-1} K^i g^i) &= (g^i - \sum_0^{i-1} \left(\beta_j G p_N^j, G^{-1} \left(g^i - \sum_0^{i-1} \beta_j G p_N^j \right) \right) \\ &= (g^i, G^{-1} g^i) + \sum_0^{i-1} \beta_j^2. \end{aligned}$$

G^{-1} is a positive operator, and by (5.3),

$$(g^i, G^{-1} g^i) \to 0 \quad \text{as } i \to \infty.$$

Therefore, taking into consideration the inequalities

$$m' \|g^i\|^2 \leqq (g^i, G^{-1} g^i),$$

we see that the gradient of $S(u^i)$ tends to zero as $i \to \infty$:

$$\|g^i\| \to 0 \quad \text{as } i \to \infty.$$

By definition of g^i, this means that the sequence $u^i, i = 0, 1, 2, \cdots$, converges to u^*.

THEOREM 4. *There is a subspace $\overline{M}$ in H such that*

$$K^i f \to G^{-1} f \quad \text{as } i \to \infty, \qquad \text{for any element } f \in \overline{M}.$$

Proof. By Theorem 1, K^i converges to an operator K on H. Operate G on the equality (4.2) from the right-hand side; then

$$GKGp_N^i = Gp_N^i. \tag{5.6}$$

Let M be a subset of H which consists of linear combinations of Gp^i, $i = 0, 1, 2, \cdots, n, \cdots$. Then the closure of M is clearly a subspace of H. The subspace is denoted by $\overline{M}$. We shall show that p^i, $i = 1, 2, \cdots$, is an element of $\overline{M}$. From (5.4),

$$g^0 = \sum_0^\infty (g^0, p_N^j) G p_N^j. \tag{5.7}$$

Hence,

$$g^i = \sum_{j=i}^\infty (g^0, p_N^j) G p_N^j.$$

Substituting this into (5.5), we have

$$\begin{aligned} p^i &= -K^i g^i \\ &= -\sum_{j=i}^\infty (g^i, p_N^j) G p_N^j + \sum_0^{i-1} \beta_j G p_N^j. \end{aligned} \tag{5.8}$$

This expression of p^i means that p^i is an element of $\overline{M}$. By Lemma 2 an element $f \in \overline{M}$ has the expression

$$f = \sum_0^\infty (f, p_N^i) G p_N^i. \tag{5.9}$$

From (5.9) we see that if an element $f \in \overline{M}$ is orthogonal to every p_N^i, $i = 0, \cdots$, then $f = 0$. Hence p_N^i, $i = 0, 1, \cdots$, is a complete system on $\overline{M}$. Then,

$$f = \sum_0^\infty (f, G p_N^i) p_N^i \tag{5.10}$$

by Lemma 3. Substitute (5.9) and (5.10) into (4.1) and (5.6); then

$$KGf = f, \qquad f \in \overline{M}, \tag{5.11}$$

$$GKf = f, \qquad f \in \overline{M}. \tag{5.12}$$

Let $K_{\overline{M}}$ and $G_{\overline{M}}$ be operators on $\overline{M}$ such that

$$\begin{aligned} K_{\overline{M}} f &= Kf, \\ G_{\overline{M}} f &= Gf \quad \text{for } f \in \overline{M}. \end{aligned}$$

Then, (5.11) and (5.12) show that

$$K_{\overline{M}} = G_{\overline{M}}^{-1}. \tag{5.13}$$

In other words,

$$\lim_{i \to \infty} K^i f = G^{-1} f \quad \text{for } f \in \overline{M}. \tag{5.14}$$

This completes the proof.

Let V be a sphere on $\overline{M}$, i.e.,

$$V = \{f : f \in \overline{M}, \| f \| \leqq 1\}.$$

If the convergence of (5.14) is uniform on V, the direction of the search in this method converges to that of Newton's method;

$$\frac{p^i}{\|g^i\|} = -K^i \frac{g^i}{\|g^i\|} \to \frac{-G^{-1}g^i}{\|g^i\|}.$$

6. Applications to optimal control problems. A control system is described by a system of ordinary differential equations

$$\dot{x} = f(x, t, u), \tag{6.1}$$

where $x \in R^n$ is a state vector and $u \in R^r$ is a control vector. Then the problem is to find a control function $u = u^*(t)$ which minimizes the value of the function

$$P(x(t_f)), \tag{6.2}$$

subject to (6.1) with an initial condition $x(t_0) = x^0$. The following conditions are assumed:

(i) $f(x, u, t)$ and $P(x)$ have continuous partial derivatives of at least third order in all variables.

(ii) There are no constraints for x and u.

Let H be a space of r-dimensional control vector functions such that

$$\int_{t_0}^{t_f} \sum_{i=1}^{r} u_i^2(\tau)\, d\tau < +\infty. \tag{6.3}$$

Then the space H is a Hilbert space with inner product

$$(u, v) = \int_{t_0}^{t_f} \sum_{1}^{r} u_i(\tau) v_i(\tau)\, d\tau. \tag{6.4}$$

Now, introduce an auxiliary vector $\psi = (\psi_1, \cdots, \psi_n)$ and a Hamiltonian $\mathscr{H}(x, \psi, u, t)$ defined as follows:

$$\mathscr{H}(x, \psi, u, t) = \sum_{1}^{n} \psi_i f_i(x, u, t), \tag{6.5}$$

$$\dot{\psi}_i = -\sum_{1}^{n} \frac{\partial f_i(x, u, t)}{\partial x_i} \psi_i, \qquad i = 1, \cdots, n, \tag{6.6}$$

$$\psi_i(t_f) = \frac{\partial p(x(t_f))}{\partial x_i(t_f)}, \qquad i = 1, 2, \cdots, n. \tag{6.7}$$

The equations (6.1) and (6.6) can be written with the Hamiltonian in canonical form:

$$\dot{x} = \frac{\partial \mathscr{H}(x, \psi, u, t)}{\partial \psi}, \qquad x(t_0) = x^0, \tag{6.8}$$

$$\dot{\psi} = -\frac{\partial \mathscr{H}(x, \chi, u, t)}{\partial x}, \qquad \psi(t_f) = \frac{\partial p}{\partial x}. \tag{6.9}$$

Let $\chi(t)$ and $\psi(t)$ be a solution of the equations (6.8), (6.9) corresponding to a certain control $u(t)$. The performance index $P(x(t_f))$ is a functional of $u(\cdot) \in H$. We denote

this functional by $J(u)$:

$$J(u) = P(x(t_f)).$$

Let $g(t)$ be the gradient of $J(u)$; then

$$g(t) = \frac{\partial \mathscr{H}(x(t), \psi(t), u(t), t)}{\partial u}. \tag{6.10}$$

If the gradient is computed according to (6.10), p_N^i and q_N^i can be constructed following the algorithm in § 3. The determination problem of the step length α^i is called the problem of linear search, and several schemes are known [1].

7. Examples.

Example 1. Consider a control system

$$\dot{x}_1 = x_2,$$

$$\dot{x}_2 = -x_1 + (1 - x_1^2)x_2 + u, \quad x_1(0) = x_{10}, \quad x_2(0) = x_{20}(0),$$

with a performance index

$$J = \int_0^5 (x_1^2 + x_2^2 + u^2)\, dt.$$

The control time is fixed as $t_0 = 0$, $t_f = 5$. We introduce the third coordinate x_3 such that

$$\dot{x}_3 = x_1^2 + x_2^2 + u^2, \qquad x_3(0) = 0.$$

The canonical equation then becomes:

$$\begin{aligned}
\dot{x}_1 &= x_2, & x_1(0) &= x_{10},\\
\dot{x}_2 &= -x_1 + (1 - x_1^2) + u, & x_2(0) &= x_{20},\\
\dot{x}_3 &= x_1^2 + x_2^2 + u^2, & x_3(0) &= 0,\\
\dot{\psi}_1 &= (1 + 2x_1x_2)\psi_2 - 2x_1\psi_3, & \psi_1(5) &= 0,\\
\dot{\psi}_2 &= -\psi_1 - (1 - x_1^2)\psi_2 - 2x_2\psi_3, & \psi_2(5) &= 0,\\
\dot{\psi}_3 &= 0, & \psi_3(5) &= 1.
\end{aligned}$$

The computed results for $x_{10} = 3.0$, $x_{20} = 0.0$ are shown in Fig. 1 and Fig. 2. The results obtained by the steepest descent method and the conjugate gradient method are also shown.

Example 2.

$$\dot{x}_1 = x_2,$$

$$\dot{x}_2 = -0.2x_2 + 2.0x_3 - 0.2x_2x_3^2,$$

$$\dot{x}_3 = -5x_2 + u,$$

$$J = \int_0^5 (x_1^2 + x_2^2 + x_3^2 + u^2)\, dt.$$

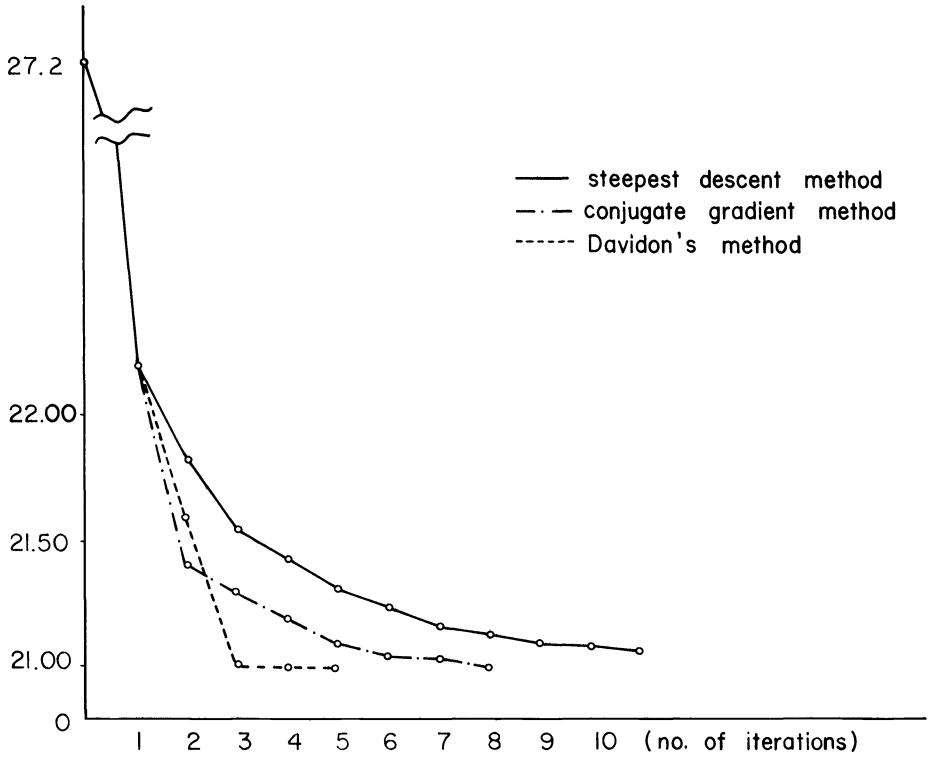


FIG. 1. *Values of performance index*

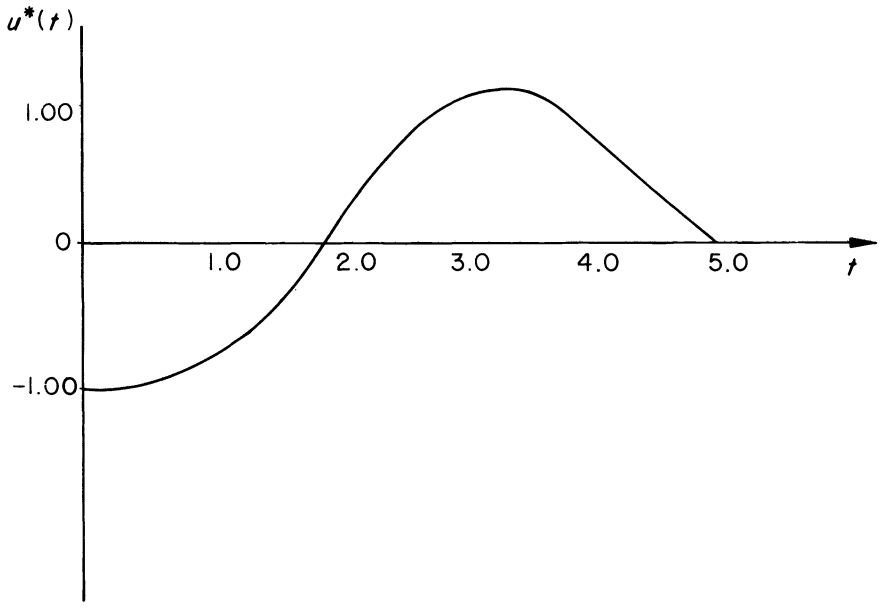


FIG. 2. *Optimal control* $u^*(t)$

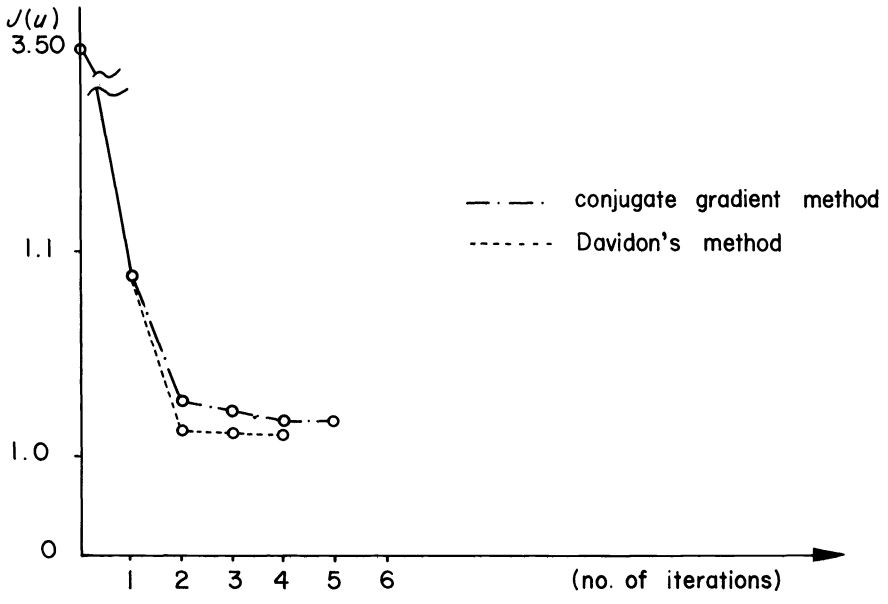


FIG. 3. *Values of performance index*

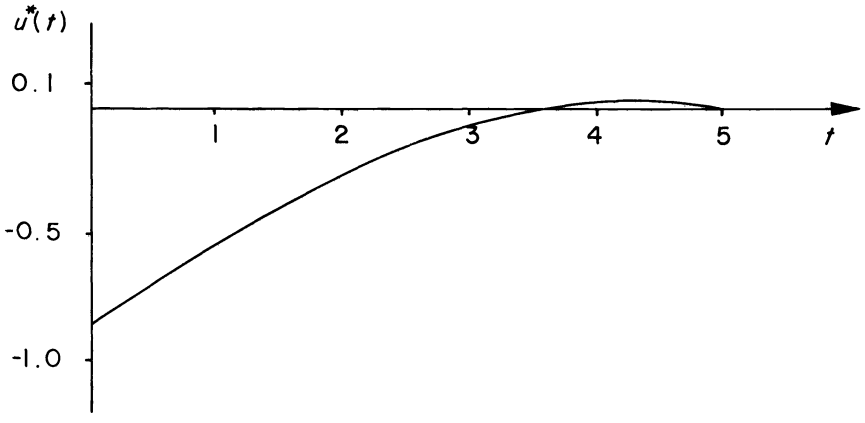


FIG. 4. *Optimal control* $u^*(t)$

The numerical results are shown in Fig. 3 and Fig. 4 with $x_{10} = 0.25$, $x_{20} = 0.25$, $x_{30} = 0.1$.

From these examples we can say that Davidon's method proposed here is applicable to nonlinear control problems and the rapid convergence is assured for these examples. These results also show that the conjugate gradient method is also a very useful scheme.

8. Conclusion. Minimization problems in Hilbert spaces are discussed. Davidon's method in finite-dimensional spaces is extended to the problems in Hilbert spaces. The stability of the method is studied for the case of quadratic

functionals. And it is proved that the method is stable from any initial approximation. In § 4, it is proved for quadratic functionals that the sequences of iterations converge to the true solution of the problem. In § 5 it is shown that the direction of the search converges to that of the Newton–Raphson method. Hence the scheme has the analogous property in the neighborhood of the extremal point. Stability is also assured for nonquadratic problems, and so this method can be applied to such problems. From these discussions for quadratic problems, we can say that Davidon's method has stability properties like those of the steepest descent method and that the convergence property in the vicinity of the extremum point is similar to that of the Newton-Raphson method.

This method is applied to optimal control problems and two numerical examples are shown. These examples show the superiority of this method compared with the steepest descent method or the conjugate gradient method.

The disadvantage of this method is that the information to be stored in the computer increases with the number of iterations. So, if convergence is slow, computing will be difficult.

REFERENCES

[1] R. FLETCHER AND C. M. REEVES, *Functional minimization by conjugate gradients*, Comput. J., 7 (1964), pp. 149–154.

[2] R. FLETCHER AND M. POWELL, *A rapidly convergent descent method for minimization*, Ibid., 6 (1963, 1964), pp. 163–168.

[3] H. J. KELLEY, *Method of gradients*, Optimization Techniques, G. Leitmann, ed., Academic Press, New York, 1962, Chap. 6, pp. 206–252.

[4] A. E. BRYSON AND W. F. DENHAM, *A steepest ascent method for solving optimal programming problems*, J. Appl. Mech., 29 (1962), pp. 247–257.

[5] R. MCGILL, *Optimal control, inequality state constraints, and the generalized Newton-Raphson algorithm*, this Journal, 3 (1965), pp. 291–293.

[6] R. E. KOPP AND R. MCGILL, *Several trajectory optimization techniques, Part I, Discussion*, Computing Methods in Optimization Problems, Academic Press, New York, 1964, pp. 65–91.

[7] L. S. LASDON, S. K. MITTER AND A. D. WAREN, *The conjugate gradient method for optimal control problems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 132–138.

[8] M. M. VAINBERG, *Variational Methods for the Study of Nonlinear Operators*, Holden-Day, Inc., San Francisco, 1964.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

EXISTENCE OF OPTIMAL STRATEGIES BASED ON SPECIFIED INFORMATION, FOR A CLASS OF STOCHASTIC DECISION PROBLEMS*

V. E. BENEŠ†

Abstract. The existence of admissible strategies $\gamma(\cdot,\cdot)$ minimizing a function $E\int_0^1 c(t, x, \gamma(t, x))\,dt$ is studied, with $x = x(\cdot, \omega)$ a continuous stochastic process, and admissible strategies defined as those utilizing a specified pattern of information about x, here described by σ-algebras G_t included in the "past" $\{x(s), s \leqq t\}$. Conclusion: Moment conditions on x, growth conditions on c, and continuity of c in its third variable (the control) ensure that an optimal admissible strategy exists. The method of proof depends on properties of conditional expectations, and on a variant of a general Filippov (or implicit functions) lemma due to McShane and Warfield.

1. Introduction. We prove an existence theorem for optimal strategies based on prescribed information, in the special case in which decisions affect only the performance criterion, and not the trajectory of a random dynamical system. This restriction is offset by the fact that our result is valid for an arbitrary stochastic process with continuous sample paths and an integrable second moment, together with an arbitrary pattern of information about the trajectory, prescribed in advance, on which decisions may be based. The general case of stochastic *control*, involving both an arbitrary information pattern and control-dependent trajectories, is substantially harder [1], [2] and is not discussed.

2. Discussion: strategies and information. A basic problem of stochastic decision theory is to formulate realistic mathematical descriptions of the decision situation. An important feature of this situation is the pattern of information available to the controller or decision-maker at various epochs of time. The possible patterns are many and diverse: he may know the state of the system at all past times, or only at the present; he may know only some functionals of the past; he may or may not remember what he did in the past; etc. It is particularly important to have an exact and useful account of the information available for decisions.

A strategy for a stochastic decision problem must be a recipe which specifies, at each time and for each condition of the controller's knowledge, a mode of action. Logically, then, a strategy should be described by a function which maps information into action, and which can be given independently of any probability spaces or stochastic processes. It is prior, in the good old Aristotelian sense, to any stochastic process which may result from its adoption.

Some previous discussions of stochastic optimal control have featured the description and role of the information available to the controller. For example, Fleming and Nisio [1] consider the case

$$dx = \alpha(t, x_t)\,dU + \beta(t, x_t)\,dw \tag{1}$$

of linear controls, with $U(\cdot)$ (a "control process" which determines the $x(\cdot)$ trajectory) satisfying

(i) $|U(t) - U(s)| \leqq |t - s|$ a.s.

(ii) $U(t)$ is independent of $w(s) - w(t)$ for $s > t$, and $w(\cdot)$ is a Wiener process.

* Received by the editors May 13, 1969.

† Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07974.

As the authors themselves realize, the class of "admissible" controls described by (i) and (ii) is very large. It cannot reflect any limitations of the information available to the controller, since among the $U(\cdot)$ thus envisaged are those depending on *the entire past of* $w(\cdot)$.

Another type of approach appears in a paper [2] of Kushner. He considers first control laws that are functions of the current state and time, and takes as admissible only those that are Lipschitz in the state uniformly in the time. This smoothness assumption is convenient for obtaining compactness in the space of control laws, but it has no basis in the original control problem.

These examples suggest that a careful examination of the role of the information available for decisions and a precise description of its place in the mathematical formulation are necessary for a full understanding of the problems of optimal stochastic decision. A start in this direction has been made in a paper [3] of Fleming on diffusion processes controlled on the basis of knowledge of only some of the components of the diffusing vector.

We shall take the view that a strategy is a function $\gamma(t, \cdot)$ which, at any time t, indicates what point of (some decision space) Γ is to be exercised as control at t as a function of whatever information about the trajectory the controller is allowed to know, remember, and use at time t. It determines the control $u(t)$ at time t in terms of the past x_t (Hale's notation) of the system trajectory, subject to restrictions on what the controller can know about x_t at t, according to the formula

$$u(t) = \gamma(t, x_t).$$

The restrictions on $\gamma(\cdot, \cdot)$ representing the pattern of available information will be described mathematically by the concept of measurability with respect to a σ-algebra. Roughly speaking, if A_1 and A_2 are two σ-algebras over the same space, and if $A_1 \subseteq A_2$, then the functions measurable with respect to A_1 are less complicated (depend on less) than those which are measurable with respect to A_2 but not A_1.

The information pattern at t will be described by specifying a σ-algebra G_t of Borel subsets of $C[0, t]$, and by imposing the condition that for each t,

$$\gamma(t, \cdot)$$

be G_t-measurable.

3. Formulation and principal result. Let $(\Omega, P, \mathscr{B})$ be a probability space, and let $\{x(t, \omega), 0 \leqq t \leqq 1, \omega \in \Omega\}$ be a measurable separable stochastic process with values in R^n, having continuous paths with probability one, and such that

$$E \int_0^1 |x(s)|^2 \, ds < \infty. \tag{2}$$

Let $I = [0, 1]$, let $C(I)$ be the space of continuous functions $x : I \to R^n$, and let S_t be the σ-algebra generated by the sets

$$\{y(\cdot) \in C : y(s) \in A\}, \qquad 0 \leqq s \leqq t \leqq 1, \quad A \text{ Borel in } R^n.$$

This is the σ-algebra representing knowledge of the full past up to time t. Let G_t be a sub-σ-algebra of S_t. Decisions at time t may depend only on information

contained in G_t, that is, as a function of the past trajectory prior to t, control exerted at t must be *measurable* on G_t.

Since $x(t, \omega)$ has continuous sample paths, there is a set $\Omega_0 \in \mathscr{B}$ such that $P(\Omega_0) = 1$ and $x(\cdot, \omega) \in C(I)$ for $\omega \in \Omega_0$. The process induces a measurable map $x: \Omega_0 \to C(I)$ according to the formula $x(\omega) = x(\cdot, \omega)$. The map x is measurable in the sense that $x^{-1}(S_1) \subseteq \mathscr{B}$. This is easily seen as follows: every set $\{y \in C(I): y(t) \in A\}$ for A Borel is in S_1; every set $\{\omega: x(t, \omega) \in A\}$ for A Borel is in $\mathscr{B}$; but

$$\{\omega: x(t, \omega) \in A\} \cap \Omega_0 = x^{-1}(\{y \in C(I): y(t) \in A\}).$$

The classes $\{x^{-1}(G_t), t \in I\}$ of $\mathscr{B}$-sets are all σ-algebras; they will provide us with a means for doing all our work with the probability space Ω, and then returning to the measurable space $C(I)$.

There is given a compact metric space Γ of control points, and a function $c: I \times C(I) \times \Gamma \to R^+$ representing cost per unit time as a function of time, trajectory, and applied control: $c(t, y, u)$ for $t, y, u \in I \times C(I) \times \Gamma$ is the cost rate at t if following trajectory $y \in C(I)$ and applying control $u \in \Gamma$ at t. We assume that $c(t, \cdot, u)$ is, for each $(t, u) \in I \times \Gamma$, measurable on the σ-algebra S_1, that $c(\cdot, y, u)$ is Lebesgue measurable for each $(y, u) \in C(I) \times \Gamma$, that $c(t, y, \cdot)$ is continuous on Γ for each $(t, y) \in I \times C(I)$, and that

$$0 \leqq c(t, y, u) \leqq \text{const.}\left(1 + \int_0^t |y(s)|^2 \, ds\right), \qquad y \in C(I), \quad u \in \Gamma.$$

An *admissible strategy* is a function $\gamma: I \times C(I) \to \Gamma$ such that $\gamma(\cdot, y)$ is Lebesgue measurable for each $y \in C(I)$ and $\gamma(t, \cdot)$ is G_t-measurable for $t \in I$. We now pose this existence problem: is there an admissible strategy that achieves

$$\inf_{\gamma \text{ admissible}} E \int_0^1 c(t, x(\cdot, \omega), \gamma(t, x(\cdot, \omega))) \, dt?$$

An admissible strategy that achieves this infimum is called optimal.

THEOREM. *Let $x(t, \omega)$ have continuous sample paths for $\omega \in \Omega_0$, with $P(\Omega_0) = 1$, let $\int_0^1 E\{|x(s)|^2 \, ds\} < \infty$, and let $c(t, y, u)$ be Lebesgue measurable in t, measurable on the Borel sets of $C(I)$ in y, and continuous in $u \in \Gamma$ compact metric, with*

$$0 \leqq c(t, y, u) \leqq \text{const.}\left(1 + \int_0^t |y(s)|^2 \, ds\right).$$

Then there exists an optimal strategy.

The proof is in § 6, following a series of preliminary results.

The referee has suggested that the situation of the small investor in the stock market can be represented approximately by the kind of setup considered here. In this case the vector of prices of stocks on the market forms the stochastic process $x(\cdot)$ in question; over this the *small* investor has no control. The control variables are the amounts of money the investor has invested in each stock. The performance index is the sum over all the stocks of the integral of the product of the amount invested (in the stock in question) times the rate at which the price is decreasing. The construction to be given would show that an optimal investment policy exists,

and that it is obtained by choosing the control which minimizes the conditional expectation of the performance rate with respect to the investor's information. This corresponds, not surprisingly, to placing money in the stocks with greatest expected growth based on the facts known to the investor, which is what investors generally try to do.

4. A lemma of Filippov's type. We shall prove and use a version of an implicit functions lemma due to McShane and Warfield [4]. As will be seen, our version differs from theirs in several significant features: it requires the range space to be separable metric instead of merely Hausdorff, and it allows simultaneous explicit as well as implicit dependence on the independent variable, provided that this dependence is measurable with respect to the same σ-algebra as is the desired function. (The separable metrizability is an easy way of paying the price of the explicit dependence.) In their paper these authors stated that results like theirs were desirable in the theory of optimal stochastic control, but their own application was only to a control problem in relaxed trajectories. Our results indicate how very fundamental their lemma is for existence theorems in optimal stochastic decision and control.

If $\mathscr{M}$ is a σ-algebra of subsets of a set M, and S is a topological space, we say that a function $g: N \to S$, $N \in \mathscr{M}$, (defined on N) is $\mathscr{M}$-measurable if and only if $g^{-1}(F) \in \mathscr{M}$ for closed $F \subseteq S$.

LEMMA 1. *Let $(M, \mathscr{M})$ be a measure space, A separable metric, and U the union of countably many compact metrizable subsets of itself. Let $k: M \times U \to A$ be continuous in its second argument for each value of the first, and $\mathscr{M}$-measurable in the first for each value of the second. Let $y: M \to A$ be $\mathscr{M}$-measurable, with*

$$y(x) \in k(x, U), \qquad x \in M.$$

Then there exists an $\mathscr{M}$-measurable $u: M \to U$ such that

$$y(x) = k(x, u(x)).$$

Proof. Suppose first that U is L a closed subset of $(0, \infty)$. The first task is to show that for C compact,

$$\zeta(C) = \{x : y(x) \in k(x, L \cap C)\} \in \mathscr{M}.$$

Let π_m be a countable cover of A by open sets of diameter $\leqq (\frac{1}{2})^m$. We claim that

$$\zeta(C) = \bigcap_m \bigcup_{S \in \pi_m} \{x : y(x) \in S \text{ and } k(x, L \cap C) \cap S \neq \varnothing\}.$$

For if $x \in \zeta(C)$, then $y(x) \in k(x, L \cap C)$. For each m there is an open set $S \in \pi_m$ with $y(x) \in S$ so that $y(x) \in k(x, L \cap C) \cap S \neq \varnothing$. Conversely, with x in the set on the right, for every m there is a set S of diameter $\leqq (\frac{1}{2})^m$ and $y(x) \in S$, $k(x, L \cap C) \cap S \neq \varnothing$; thus $y(x)$ is at most $(\frac{1}{2})^m$ away from $k(x, L \cap C)$. Since this is true for each m, $y(x) \in \overline{k(x, L \cap C)} = k(x, L \cap C)$ because $k(x, \cdot)$ is continuous and $L \cap C$ is compact. Thus $x \in \zeta(C)$. It is apparent that for S open,

$$\{x : k(x, L \cap C) \cap S \neq \varnothing\} = \bigcup_{u \in L \cap C} \{x : k(x, u) \in S\}.$$

Each set in the possibly uncountable union on the right belongs to $\mathcal{M}$. We show that a countable union can be used. Let D be countable dense in $L \cap C$, $\{u_n\} \subseteq D$ converge to u, with $k(x,u) \in S$. Then $k(x,u_n) \to k(x,u)$ by continuity of $k(x,\cdot)$. Thus $k(x,u_n) \in S$ eventually because S is open. Hence for some n depending on S, u, x,

$$k(x,u) \in S \Rightarrow x \in \{w : k(w,u_n) \in S\}$$
$$\Rightarrow x \in \bigcup_m \{w : k(w,u_m) \in S\} \subseteq \bigcup_{u \in D} \{x : k(x,u) \in S\}.$$

This proves that $L \cap C$ in the union can be replaced by the countable set D. Hence for S open, $\{x : k(x, L \cap C) \cap S \neq \varnothing\} \in \mathcal{M}$ and $\{x : y(x) \in S\} \in \mathcal{M}$. Thus $\zeta(C) \in \mathcal{M}$.

On the set

$$B_j^q = \{x : y(x) \in k(x, L \cap [0, j \cdot 2^{-q}]) - k(x, L \cap [0, (j-1) \cdot 2^{-q}])\} \in \mathcal{M},$$

set $u_q(x) = \inf\{u : u \in L \cap ((j-1) \cdot 2^{-q}, j \cdot 2^{-q}]\}$, $j_q(x) = j$, noting that $L \cap ((j-1) \cdot 2^{-q}, j \cdot 2^{-q}] = \varnothing$ implies $B_j^q = \varnothing$. u_q and j_q are $\mathcal{M}$-measurable functions:

$$\{x : j_q(x) = j\} = B_j^q \in \mathcal{M}.$$

Note that $j_{q+1}(x) = 2j_q(x)$ or $2j_q(x) - 1$. Thus

$$u_{q+1}(x) = \inf\{u : u \in L \cap ((j_{q+1}(x) - 1) \cdot 2^{-q-1}, j_{q+1} \cdot 2^{-q-1}]\}$$
$$= \begin{cases} \inf : L \cap ((j_q - \tfrac{1}{2}) \cdot 2^{-q}, j_q \cdot 2^{-q}] \\ \text{or} \\ \inf : L \cap ((j_q - 1) \cdot 2^{-q}, (j_q - \tfrac{1}{2}) \cdot 2^{-q}] \end{cases}$$
$$\geqq u_q(x).$$

So $u_q \uparrow$. We show that $u_q(x)$ is bounded. Suppose $x \in B_j^0$, so that $j_0(x) = j$,

$$u_0(x) = \inf\{u : L \cap ((j-1), j]\}$$
$$\leqq j.$$

Then $j_1(x) = 2j$ or $2j - 1$ and

$$u_1(x) = \begin{cases} \inf\{u : L \cap ((2j-1) \cdot 2^{-1}, 2j \cdot 2^{-1}]\} \\ \text{or} \\ \inf\{u : L \cap ((2j-2) \cdot 2^{-1}, (2j-1) \cdot 2^{-1}]\} \end{cases}$$
$$\leqq j.$$

In general $j_{q+1}(x) = 2j_q(x)$ or $2j_q(x) - 1 \leqq 2^{q+1} j$,

$$u_{q+1}(x) = \begin{cases} \inf\{u : L \cap ((2j_q - 1) \cdot 2^{-q-1}, 2j_q \cdot 2^{-q-1}]\} \\ \text{or} \\ \inf\{u : L \cap ((2j_q - 2) \cdot 2^{-q-1}, (2j_q - 1) \cdot 2^{-q-1}]\} \end{cases}$$
$$\leqq 2j_q \cdot 2^{-q-1} \leqq j.$$

So $u_q \uparrow u$ $\mathcal{M}$-measurable. We have $u_q \in L$, so $u \in L$ since L is closed.

We now claim

$$y(x) = k(x, u(x)).$$

If not, there exists an $x \in M$ and a neighborhood V of $k(x, u(x))$ such that $y(x) \notin V$. (A is Hausdorff.) $k(x, \cdot)$ is continuous, so $k(x, \cdot)^{-1} V$ includes a neighborhood of $u(x)$. Thus there exist j and q such that

$$u(x) \in L \cap ((j-1) \cdot 2^{-q}, j \cdot 2^{-q}],$$

$$L \cap [(j-1) \cdot 2^{-q}, j \cdot 2^{-q}] \subseteq k(x, \cdot)^{-1} V.$$

Hence $k(x, L \cap [(j-1) \cdot 2^{-q}, j \cdot 2^{-q}]) \subseteq V$, and so

$$k(x, L \cap [0, j \cdot 2^{-q}]) - k(x, L \cap [0, (j-1) \cdot 2^{-q}]) \subseteq V.$$

We show that for this j and q,

$$u_q(x) = \inf\{u : u \in L \cap ((j-1) \cdot 2^{-q}, j \cdot 2^{-q}]\}.$$

We have $u_q(x) \in L$ and

$$u_q(x) \leqq u(x) \leqq j \cdot 2^{-q},$$

so

$$j_q(x) \leqq j.$$

Suppose

$$j_q(x) \leqq j - 1.$$

Then

$$j_{q+1}(x) \leqq 2 j_q(x) \leqq 2(j-1),$$

$$j_{q+n}(x) \leqq 2^n (j-1).$$

But

$$u_{q+n}(x) \leqq j_{n+q}(x) \cdot 2^{-(n+q)},$$

so

$$u_{q+n}(x) \leqq 2^n (j-1) 2^{-n-q} = (j-1) 2^{-q};$$

hence

$$u(x) \leqq (j-1) 2^{-q}.$$

This contradicts $u(x) \in ((j-1) \cdot 2^{-q}, j \cdot 2^{-q}] \cap L$. Hence $j_q(x) = j$, i.e., $x \in B_j^q$, or

$$y(x) \in k(x, L \cap [0, j \cdot 2^{-q}]) - k(x, [0, (j-1) \cdot 2^{-q}]) \subseteq V,$$

in contradiction with $y(x) \notin V$. The theorem is proved for $U = L$, a closed subset of $(0, \infty)$.

Now let U be such that there is a continuous map φ [5] taking L onto U. By what has been proved there is an $\mathcal{M}$-measurable $T : M \to L$ such that $y(x) = k(x, \varphi(Tx))$. Set $u(x) = \varphi(Tx)$. If F is a closed subset of U, then $\varphi^{-1}(F)$ is closed, and $\{x : Tx \in \varphi^{-1}(F)\} \in \mathcal{M}$. Thus $u(\cdot)$ is $\mathcal{M}$-measurable. The extension to the case where U is the countable union of metrizable subsets of itself is as in [4]; since it is not used, it is omitted.

5. Additional preliminaries. It is convenient to express the property of being an admissible strategy in terms of a single σ-algebra. This is done as follows: consider the class G of measurable subsets E of $I \times C(I)$ such that:

(i) Every t-section of E is a G_t-set, for $t \in I$.

(ii) Every y-section of E is a Lebesgue set, for $y \in C(I)$.

It is easy to verify that G is an algebra; since G is closed under monotone limits, it is a σ-algebra; it can then be proved that a function h on $I \times C(I)$ is G-measurable if and only if $h(t, \cdot)$ is G_t-measurable for fixed $t \in I$ and $h(\cdot, y)$ is Lebesgue measurable for fixed $y \in C(I)$. Thus measurability with respect to G concisely expresses the requirement of admissibility for the present problem.

Let $\varphi: I \times \Omega_0 \to I \times C(I)$ according to the formula $\varphi(t, \omega) = t, x(\omega)$. Define $\mathscr{F} = \varphi^{-1}(G)$. $\mathscr{F}$ is a σ-algebra of $I \times \Omega_0$ sets, and will be used to express the requirement of admissibility in terms of functions of t, ω rather than t, y for $y \in C(I)$. Indeed we shall prove the existence of an optimal control law by first expressing it as an $\mathscr{F}$-measurable t, ω function, and then (properly, now) as a G-measurable t, y function, using the following elementary result.

LEMMA 2. *If $f: I \times \Omega_0 \to \Gamma$ compact metric is $\mathscr{F}$-measurable, then there exists a G-measurable $g: I \times C(I) \to \Gamma$ such that*

$$g(t, x(\omega)) = f(t, \omega), \qquad \omega \in \Omega_0.$$

Proof. Let $\{\pi_m, m \geqq 1\}$ be refining countable partitions of Γ into sets of diameter $\leqq 2^{-m}$, and let $S_1^m, S_2^m, \cdots$ enumerate π_m, m fixed. For each S_n^m there is a set $B_{mn} \in G$ such that $f^{-1}(S_n^m) = \varphi^{-1}(B_{mn})$. Define $A_{m1} = B_{m1}, A_{mn} = B_{mn} - \bigcup_{i=1}^{n-1} A_{mi}$. The $A_{mn}, n \geqq 1$, are disjoint G-sets.

Given $S_{n'}^{m+1}$, there is an S_n^m containing it, so that $f^{-1}(S_{n'}^{m+1}) \subseteq f^{-1}(S_n^m)$ and $\varphi^{-1}(B_{(m+1)n'}) \subseteq \varphi^{-1}(B_{mn})$. Hence $B_{(m+1)n'} \subseteq B_{mn}$. Thus

$$\bigcup_n B_{mn} \downarrow \bigcap_m \bigcup_n B_{mn} = A \in G$$

and $\bigcup_n A_{mn} \downarrow A$. Let γ be a fixed element of Γ, and define $g_m: I \times C(I) \to \Gamma$ by

$$g_m(t, y) = \begin{cases} \text{some element of } S_n^m & \text{if } t, y \in A_{mn} \cap A, \\ \gamma & \text{if } t, y \notin A. \end{cases}$$

The g_m are G-measurable functions, and $I \times x(\Omega_0) = \varphi \bigcup_n f^{-1}(S_n^m) = \varphi \bigcup_n \varphi^{-1}(B_{mn}) = \varphi \bigcup_n \varphi^{-1}(A_{mn}) = \bigcup_n \varphi\varphi^{-1}(A_{mn}) \subseteq \bigcup_n A_{mn}$. Thus $I \times x(\Omega_0) \subseteq A$.

Now

$$\begin{aligned} g_m(t, x(\omega)) \in S_n^m &\Leftrightarrow t, x(\omega) \in A_{mn} \\ &\Leftrightarrow t, x(\omega) \in B_{mn}, \notin \bigcup_{i=1}^{n-1} B_{mi} \\ &\Leftrightarrow t, \omega \in \varphi^{-1}(B_{mn}), \notin \bigcup_{i=1}^{n-1} \varphi^{-1}(B_{mi}) \end{aligned}$$

$$\Leftrightarrow t, \omega \in f^{-1}(S_n^m), \notin \bigcup_{i=1}^{n-1} f^{-1}(S_i^m)$$

$$\Leftrightarrow t, \omega \in f^{-1}(S_n^m)$$

$$\Leftrightarrow f(t, \omega) \in S_n^m.$$

Since

$$\operatorname{dist}\{g_m(t, y), g_{m+p}(t, y)\} \leqq 2^{-m},$$

the g_m form a Cauchy sequence pointwise. Since Γ is complete, there is a G-measurable g such that $g_m \to g$ pointwise. Since for $\omega \in \Omega_0$,

$$\operatorname{dist}\{g_m(t, x(\omega)), f(t, \omega)\} \leqq 2^{-m},$$

we conclude that $g(t, x(\omega)) = f(t, \omega)$ for $t \in I, \omega \in \Omega_0$.

Another preliminary concerns composition of measurable functions with conditional expectations. $(\Omega, P, \mathscr{B})$ is a probability space, and $\mathscr{S}$ a topological space.

LEMMA 3. *Let $k: \Omega \times \mathscr{S} \to R$ be continuous in the second variable for almost every value of the first, with $0 \leqq k(\omega, u) \leqq k(\omega)$ integrable, let $\mathscr{A}$ be a sub-σ-algebra of $\mathscr{B}$, and let $K(\omega, u) = E\{k(\omega, u)|\mathscr{A}\}$. If $f: \Omega \to S$ is $\mathscr{A}$-measurable, then*

$$K(\omega, f(\omega)) = E\{k(\omega, f(\omega))|\mathscr{A}\} \quad a.e.$$

Proof. The functions $\{k(\cdot, u), u \in \mathscr{S}\}$ are uniformly dominated and

$$\sup_{u \in \mathscr{S}} Ek(\omega, u) < \infty.$$

Hence $K(\omega, \cdot)$ is continuous for almost all ω, and $\{K(\cdot, u), u \in \mathscr{S}\}$ are uniformly dominated. We need to show that for an $\mathscr{A}$-set A,

$$\int_A k(\omega, f(\omega))\, dP = \int_A K(\omega, f(\omega))\, dP.$$

Suppose that f is $\mathscr{A}$-simple, indeed that $f = a_i$ on A_i, for $A_i \in \mathscr{A}$, $0 \leqq i \leqq n$. Then

$$\begin{aligned}\int_A k(\omega, f(\omega))\, dP &= \sum_{i=1}^{n} \int_{A \cap A_i} k(\omega, a_i)\, dP \\ &= \sum_{i=1}^{n} \int_{A \cap A_i} K(\omega, a_i)\, dP \\ &= \int_A K(\omega, f(\omega))\, dP.\end{aligned}$$

The lemma thus holds for $\mathscr{A}$-simple f. Let f_n, $\mathscr{A}$-simple, approach f pointwise. Then

$$\int_A k(\omega, f_n(\omega))\, dP = \int_A K(\omega, f_n(\omega))\, dP,$$

and since both k, K are continuous (a.e.) and dominated uniformly in u, we may integrate to the limit.

6. Proof of theorem. With $\varphi(t, \omega) = t, x(\omega)$ as in § 5, set $\mathscr{F} = \varphi^{-1}(G)$, and let $\bar{\mathscr{F}}$ be the completion of $\mathscr{F}$ with respect to the probability measure $\lambda \times P$,

with λ = Lebesgue measure on the Lebesgue sets of the unit interval. Let

$$K(t, \omega, u) = E\{c(t, x(\omega), u) | \mathscr{F}\}.$$

For $u \in \Gamma, K(\cdot, \cdot, u)$ is an $\bar{\mathscr{F}}$-measurable function, and $K(t, \omega, \cdot)$ is a.s. continuous, because $c(t, x(\omega), \cdot)$ is, and the $\{c(t, x(\omega), u), u \in \Gamma\}$ are dominated by an integrable function.

Let ψ be a continuous map of the Cantor set C *onto* Γ. Such a map exists by the theorem of Hocking and Young [5] because Γ is compact metric; use of ψ was prompted by a similar use in [4].

The r.v.

$$A_h(t, \omega) = \sup_{\substack{s, s' \in C \\ |s - s'| \leqq h}} |c(t, x(\omega), \psi(s)) - c(t, x(\omega), \psi(s'))|$$

is well-defined; $A_h \to 0$ as $h \downarrow 0$ a.s., and $\{A_h, h > 0\}$ are dominated by an integrable function.

$$|K(t, \omega, \psi(s)) - K(t, \omega, \psi(s'))| \leqq E\{|c(t, x(\omega), \psi(s)) - c(t, x(\omega), \psi(s'))| | \mathscr{F}\} \text{ a.e.}$$

$$\sup_{\substack{s, s' \in C \\ |s - s'| \leqq h}} |K(t, \omega, \psi(s)) - K(t, \omega, \psi(s'))| \leqq E\{A_h(t, \omega) | \mathscr{F}\} \quad \text{a.e.}$$

The right side approaches 0 a.s. as $h \downarrow 0$, since $\{A_h\}$ are dominated. This shows that the process

$$\{K(t, \omega, \psi(s)), s \in C\}$$

has continuous sample paths on some $\bar{\mathscr{F}}$-set Λ of full measure. Hence

$$\inf_{u \in \Gamma} K(t, \omega, u) \in K(t, \omega, \psi(C))$$

on Λ.

For $\omega \in \Omega_0$ and any t, $c(t, x(\omega), \cdot)$ is a continuous function on Γ. So there exists a set $M_0 \in \bar{\mathscr{F}}$ such that for $t, \omega \in M_0$, $c(t, x(\omega), \cdot)$ is continuous, and M_0 has $(\lambda \times P)$-measure one. Let ρ be the metric on Γ. We have

$$\sup_{\rho(u_1, u_2) \leqq h} |E\{c(t, x(\omega), u_1 c(t, x(\omega), u_1) | \mathscr{F}\} - E\{c(t, x(\omega), u_2) | \mathscr{F}\}|$$
$$\leqq E\left\{ \sup_{\rho(u_1, u_2) \leqq h} |c(t, x(\omega), u_1) - c(t, x(\omega), u_2)| \, | \mathscr{F} \right\}.$$

The sup on the right decreases to 0 as $h \downarrow 0$, on a t, ω set of full measure, and it is dominated uniformly in h. Hence there is a set M of full measure, $M \in \bar{\mathscr{F}}$, such that $K(t, \omega, u)$ is continuous in u for $t, \omega \in M$. Γ is separable, so there is a countable set $\{u_n, n \geqq 1\} = S$ dense in Γ such that

$$\inf_{u \in \Gamma} K(t, \omega, u) = \inf_{u_n \in S} K(t, \omega, u_n), \quad \text{whenever } t, \omega \in M.$$

Clearly

$$\left\{ t, \omega : \inf_{u \in \Gamma} K(t, \omega, u) < a \right\} \cap M = M \cap \bigcup_n \{t, \omega : K(t, \omega, u_n) < a\}.$$

Since the sets in the union belong to $\bar{\mathscr{F}}$, $\inf_{u \in \Gamma} K(t, \omega, u)$ is equal a.e. $\lambda \times P$ to an $\bar{\mathscr{F}}$-measurable function.

Let

$$y(t,\omega) = \begin{cases} \inf_{u\in\Gamma} K(t,\omega,u) & \text{on } \Lambda \cap M, \\ 1 & \text{on } \Omega - (\Lambda \cap M), \end{cases}$$

$$A(t,\omega,s) = \begin{cases} K(t,\omega,\psi(s)) & \text{on } \Lambda \cap M, \\ 1 & \text{on } \Omega - (\Lambda \cap M). \end{cases}$$

It is clear that $y(t,\omega) \in A(t,\omega,C)$ and that $A(t,\omega,\cdot)$ is continuous on C for *every* (t,ω), and also that $A(\cdot,\cdot,s)$ is $\bar{\mathscr{F}}$-measurable for each $s \in C$, and that $y(\cdot,\cdot)$ is $\bar{\mathscr{F}}$-measurable.

By Lemma 1, there is an $\bar{\mathscr{F}}$-measurable function $\xi:[0,1] \times \Omega \to C$ such that

$$y(t,\omega) = A(t,\omega,\xi(t,\omega)).$$

Hence on $\Lambda \cap M$,

$$K(t,\omega,\psi(\xi(t,\omega))) = \inf_{u\in\Gamma} K(t,\omega,u).$$

Let $\bar{\eta} = \varphi(\xi)$; $\bar{\eta}(\cdot)$ is then an $\bar{\mathscr{F}}$-measurable function such that a.e.,

$$K(t,\omega,\bar{\eta}(t,\omega)) = \inf_{u\in\Gamma} K(t,\omega,u).$$

We can change $\bar{\eta}$ on a set of measure zero to an $\mathscr{F}$-measurable function η.

By Lemma 2, there is a G-measurable function $\gamma^*:[0,1] \times C[0,1] \to \Gamma$ such that a.e.

$$\gamma^*(t,x(\omega)) = \eta(t,\omega).$$

We claim that γ^* is an optimal strategy. γ^* is admissible, and for any other admissible law γ,

$$K(t,\omega,\gamma^*(t,x(\omega))) \leqq K(t,\omega,\gamma(t,x(\omega))) \quad \text{a.e.}$$

By Lemma 3, since γ, γ^* are $\bar{\mathscr{F}}$-measurable, this is equivalent to

$$E\{c(t,x(\omega),\gamma^*(t,x(\omega)))|\mathscr{F}\} \leqq E\{c(t,x(\omega),\gamma(t,x(\omega)))|\mathscr{F}\}.$$

Integrating, we see that γ^* is optimal.

7. Acknowledgment. The author is deeply indebted to H. S. Witsenhausen for calling his attention to the existence problems for optimal strategies in stochastic decision situations, for extensive clarifying discussions, and for encouragement during rough going. Many of the ideas and methods used here were suggested to the author by him.

REFERENCES

[1] W. H. Fleming and M. Nisio, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777–794.

[2] H. J. Kushner, *On the existence of optimal stochastic controls*, this Journal, 3 (1965), pp. 463–474.

[3] W. H. Fleming, *Optimal control of partially observable diffusions*, this Journal, 6 (1968), pp. 194–214.

[4] E. J. McShane and R. B. Warfield, *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.

[5] J. G. Hocking and G. S. Young, *Topology*, Addison-Wesley, Reading, Mass., 1961.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

RECOVERABILITY FOR PROCESSES WITH BOUNDED CONTROL AMPLITUDES AND RATES*

J. S. SHAFRAN† AND J. Y. S. LUH‡

1. Introduction. In many processes, the solution to a specified control problem may not exist. As an example, for a given system with a particular set of control constraints it may not be possible to drive a certain set of initial values to the origin. Therefore in this case there is no solution to the minimal time problem for these initial values. The subject of recoverability is concerned with determining the conditions which guarantee the existence of solutions for all possible initial values.

The recoverability of linear systems with bounded control amplitudes has been investigated by, among others, LaSalle [1]. Later, his results were extended to linear time-varying processes by LeMay [2]. This paper investigates the subject of recoverability for linear time-varying processes with control inputs bounded in both amplitude and rate. The results are summarized in two theorems, which provide a necessary and a sufficient condition for complete recoverability. Also, the results are compared to those obtained by the previously mentioned authors.

2. Statement of the problem. The control process under investigation is described by the linear time-varying differential system:

$$\dot{y}(t) = E(t)y(t) + F(t)u(t) \tag{1}$$

on $t \in [t_0, \infty)$. The n-dimensional vector y describes the state of the system. u is an m-dimensional vector which represents the control input, $m \leqq n$, and $E(t)$ and $F(t)$ are bounded and continuous real matrices of dimensions $n \times n$ and $n \times m$, respectively. It is assumed that (1) is completely controllable, for general initial time t_0, in the sense defined by LaSalle [1]. Let G and Ω be two polyhedrons in R^m, given by

$$G = \{u \,|\, |u_i| \leqq 1, i = 1, 2, \cdots, m\}, \tag{2}$$

and

$$\Omega = \{\dot{u} \,|\, |\dot{u}_i| \leqq h_i, h_i > 0, i = 1, 2, \cdots, m\}. \tag{3}$$

The control vector u of (1) is said to be admissible if, for all $t \in [t_0, \infty)$,

(a) $u(t)$ is continuous and has a piecewise continuous derivative, and

(b) $u(t) \in G$ and $\dot{u}(t) \in \Omega$.

The problem is to establish, for (1), the conditions under which every initial value can be driven to the origin using admissible controls. This is termed the problem of determining the conditions for complete recoverability.

* Received by the editors February 28, 1969, and in revised form September 16, 1969. This work was supported in part by the Jet Propulsion Laboratory, California Institute of Technology, sponsored by the National Aeronautics and Space Administration under Contract NAS 7-100.

† TRW Systems Group, Inc., Redondo Beach, California 90268.

‡ School of Electrical Engineering, Purdue University, Lafayette, Indiana 47077.

From condition (a) it is evident that the admissible controls must certainly be continuous at $t = t_0$, i.e., $\lim_{\varepsilon \to 0} u(t_0 - \varepsilon) = u(t_0), \varepsilon > 0$. Therefore, the admissible controls for each initial condition $y(t_0)$ are dependent on the choice of the initial value of the control vector. Hence, for completeness, the recoverability of each initial condition $y(t_0)$ must be determined for all possible values of the initial control. In order to accomplish this, the above control process will be reformulated as a bounded state variable process. For $t \in [t_0, \infty)$ and $i = 1, 2, \cdots, m$, let

$$\begin{aligned} x_{n+i}(t) &= u_i(t), \\ v_i(t) &= \dot{u}_i(t). \end{aligned} \tag{4}$$

Then (1) can be augmented as

$$\dot{x}(t) = A(t)x(t) + Bv(t) \tag{5}$$

with

$$x(t_0) = \begin{bmatrix} y(t_0) \\ u(t_0) \end{bmatrix},$$

where

$$x(t) = \begin{bmatrix} y_1(t) \\ \vdots \\ y_n(t) \\ \vdots \\ u_m(t) \end{bmatrix} = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \\ \vdots \\ x_{n+m}(t) \end{bmatrix}, \quad A(t) = \begin{bmatrix} E(t) & F(t) \\ 0_1 & 0_2 \end{bmatrix}, \quad B = \begin{bmatrix} 0_3 \\ I_3 \end{bmatrix},$$

and 0_1 is an $m \times n$ zero matrix, 0_2 is an $m \times m$ zero matrix, 0_3 is an $n \times m$ zero matrix and I_3 is an $m \times m$ identity matrix. System (5) is in the bounded state variable formulation since $x(t) \in \hat{G}$ for every $t \in [t_0, \infty)$, where $\hat{G} = \{x \| x_j| \leqq 1, j = n + 1, n + 2, \cdots, n + m\}$. Notice that matrix B always has the same form for any given (1). For $t \in [t_0, \infty)$, the class of admissible controls for (5) may be written in terms of the new control vector v as

$$\begin{aligned} \Lambda(t) = \{v | v &\text{ is piecewise continous} \\ &v(s) \subset \Omega \text{ and } x(s) \subset \hat{G} \text{ for all } s \in [t_0, t]\}. \end{aligned} \tag{6}$$

The problem for (5) can be stated as follows. Under what conditions can every initial condition $x(t_0) = \begin{bmatrix} y(t_0) \\ u(t_0) \end{bmatrix}$ be driven to the origin, using admissible controls? Note that in this formulation the value of the initial control vector is indeed included as part of the problem of complete recoverability. It should be pointed out that the requirement that the augmented state vector x be driven to the origin implies that the original control vector u be driven to $u = 0$. This additional constraint was introduced to simplify the development of the conditions for

complete recoverability. It may also be argued that this is a reasonable physical constraint since it is desirable to avoid disengaging the control input at $y = 0$. Therefore it is desirable that $u = 0$, whenever $y = 0$.

In order to analyze the above problem, it is necessary to examine the various properties of the set of recoverability and the set of recoverable initial conditions. These two sets are precisely defined below.

3. Set definitions. From the variation of parameters formula, the solution to (5) may be written as

$$x(t) = \phi(t, t_0)x(t_0) + \int_{t_0}^{t} \phi(t, s)Bv(s)\,ds, \qquad v \in \Lambda(t), \tag{7}$$

where $\phi(t, t_0)$ is a fundamental solution matrix of the homogeneous differential equations corresponding to (5) with $\phi(t_0, t_0) = I$, the identity matrix. The set of recoverability, $K(t, t_0)$, is defined as the set of initial conditions at time t_0 that can be brought to the origin at time t, using admissible controls. By setting $x(t) = 0$ and by solving $x_0 = x(t_0)$ in (7), the following expression for the set of recoverability is obtained:

$$K(t, t_0) = \left\{ x | x = -\int_{t_0}^{t} \phi(t_0, s)Bv(s)\,ds, v \in \Lambda(t) \right\}. \tag{8}$$

It can be shown (Appendix A.1) that the set of recoverability is compact, convex and varies continuously with t. Moreover, the set $K(t, t_0)$ is nondecreasing in t.

The set of all recoverable initial conditions is defined, using (8), as

$$S(t_0) = \bigcup_{t \geqq t_0} K(t, t_0) \subset \hat{G}, \tag{9}$$

where the set $\hat{G}$ was defined in § 2. In Appendix A.2, it is shown that the set $S(t_0)$ is convex.

With the aid of the above two set definitions, it is seen that the problem of complete recoverability for (5) (and hence, (1)) is equivalent to determining the conditions under which $S(t_0) = \hat{G}$.

4. A necessary condition for complete recoverability. The following definitions are useful in the development. Let $\psi(t)$, an $(n + m)$-dimensional row vector, be a nontrivial solution to the adjoint equations $\dot{\psi}(t) = -\psi(t)A(t)$ of (5), corresponding to some initial condition $\psi(t_0)$. Let $\psi(t)$ be partitioned as $\psi(t) = [\psi^1(t), \psi^2(t)]$, where $\psi^1(t) = [\psi_1(t), \cdots, \psi_n(t)]$ and $\psi^2(t) = [\psi_{n+1}(t), \cdots, \psi_{n+m}(t)]$. Also, let $\|\psi^2(t)\| = \sum_{i=1}^{m} |\psi_{n+i}(t)|$.

THEOREM 1. *If $S(t_0) = \hat{G}$, then for every choice $\psi(t_0)$, with $\psi^1(t_0)$ nonzero, of the initial value of the adjoint vector, the following relation must be satisfied*:

$$\lim_{T \to \infty} \int_{t_0}^{T} \|\psi^2(s)\|\,ds = \infty. \tag{10}$$

Theorem 1 gives a necessary condition that the entire original state space be recoverable, subject to the restraint sets Ω, G and $\hat{G}$. The proof of this theorem is straightforward and is presented in [3]. This condition is equivalent to the results derived by LeMay [2] for processes with no bound on the control rate.

5. A sufficient condition for complete recoverability. For convenience, in the following discussion a class of vector functions $\psi^2(t)$, called Class Γ, is introduced. The definition of this class of functions is developed below.

The scalar function $\psi_{n+j}(t), j = 1, 2, \cdots, m, t \geqq t_0$, is said to belong to Class Γ_1 if for any number $H > 0$, there exists a time $\tilde{T} > t_0$ such that

$$|\psi_{n+j}(t)| > H \qquad \text{for all } t \geqq \tilde{T}. \tag{11}$$

The scalar function $\psi_{n+j}(t), j = 1, 2, \cdots, m, t \geqq t_0$, is said to belong to Class Γ_2 if it does not belong to Class Γ_1 and if it satisfies all of the following three relations:

$$\text{(i)} \qquad |d\psi_{n+j}(t)/dt| \leqq r, \qquad r > 0, \quad \text{for all } t \geqq t_0, \tag{12}$$

$$\text{(ii)} \qquad \Delta t_i = (t_{i+1} - t_i) \leqq \beta, \quad \beta > 0, \quad i = 1, 2, \cdots, N-1, \tag{13}$$

where $t_0 < t_1 < t_2 < \cdots < t_N < T, N = N(T)$ and $T \in (t_0, \infty)$, are defined to be the time instants at which $\psi_{n+j}(t)$ is at a local maximum or minimum value; and no local extremum exists at any other time instants between t_0 and T. Also, β and r are both independent of T.

(iii) For any number $\tilde{H} > 0$, there exists a time $T > t_0$ and a corresponding integer $N = N(T) > 0$, such that

$$\sum_{i=1}^{N-1} (\Delta\psi_i)^2 > \tilde{H}, \tag{14}$$

where $\Delta\psi_i = |\psi_{n+j}(t_{i+1}) - \psi_{n+j}(t_i)|$.

The m-dimensional function $\psi^2(t), t \geqq t_0$, is said to be in Class Γ of vector functions if at least one component $\psi^2(t)$ belongs to either Class Γ_1 or Class Γ_2.

THEOREM 2. *If $\psi^2(t)$ of the adjoint vector $\psi(t)$, $t \geqq t_0$, belongs to Class Γ for every choice of $\psi(t_0)$ with $\psi^1(t_0)$ nonzero, then $S(t_0) = \hat{G}$.*

Theorem 2 provides a sufficient condition for complete recoverability. The proof of this theorem is established by a sequence of three lemmas.

LEMMA 1. *$S(t_0) = \hat{G}$ if and only if for any number $H > 0$ and any choice $\psi(t_0) = [\psi^1(t_0), \psi^2(t_0)]$, with $\psi^1(t_0)$ nonzero, of the initial value of the adjoint vector, there exists a time $T > t_0$ and a control $v \in \Lambda(T)$, such that*

$$\int_{t_0}^{T} \psi^2(s) v(s)\, ds > H. \tag{15}$$

The proof of Lemma 1 is rather lengthy and is therefore omitted here. It is presented in [3]. The proof is based on the convexity property of $S(t_0)$, and the properties of the adjoint vector $\psi(t)$.

The proof of Theorem 2 is carried out in the following manner. First it is shown that if any component of $\psi^2(t)$ belongs to Class Γ_1 for a given choice of $\psi(t_0)$, with $\psi^1(t_0)$ nonzero, then (15) is satisfied. In a similar manner, the proof is carried out for Class Γ_2. Then, the proof of Theorem 2 immediately follows from Lemma 1.

LEMMA 2. *If for a given choice of initial condition $\psi(t_0)$ of the adjoint vector, with $\psi^1(t_0)$ nonzero, there exists a component of $\psi^2(t)$ belonging to Class Γ_1, then*

for any number $H > 0$ there exists a time $T > t_0$ and a control $v \in \Lambda(T)$, such that (15) is satisfied.

Proof. Suppose, for some $j, j = 1, \cdots, m$, that $\psi_{n+j}(t)$ belongs to Class Γ_1. Let the time T be defined as

$$T = \tilde{T} + 1/h_j \tag{16}$$

and define the control $v(t)$ on $t \in [t_0, T]$ by:

$$v_i(t) \equiv 0, \qquad i \neq j, \quad i = 1, 2, \cdots, m,$$
$$v_j(t) = \begin{cases} 0 & \text{if } t_0 \leqq t < \tilde{T}, \\ h_j \operatorname{sgn} [\psi_{n+j}(t)] & \text{if } \tilde{T} \leqq t \leqq T, \end{cases} \tag{17}$$

where $h_j > 0$ is the amplitude bound on the jth component of the admissible controls. From the definition of the class of admissible controls, (6), it is seen that $v(t)$ on $t \in [t_0, T]$ is an admissible control, i.e., $v \in \Lambda(T)$. For this choice of control, the left-hand side of (15) becomes

$$\int_{t_0}^{T} \psi^2(s) v(s)\, ds = h_j \int_{\tilde{T}}^{T} |\psi_{n+j}(s)| ds, \qquad v \in \Lambda(T). \tag{18}$$

However, from (11) and (16), (18) implies that

$$\int_{t_0}^{T} \psi^2(s) v(s)\, ds > h_j H/h_j = H, \qquad v \in \Lambda(T). \tag{19}$$

Since the above procedure is valid for any choice of the number $H > 0$, then (19) is equivalent to (15) of Lemma 1.

LEMMA 3. *Lemma 2 is also valid if Class Γ_1 is replaced by Class Γ_2 in the statement of the lemma.*

Proof. With reference to (14) and since $\tilde{H} > 0$, then $N = N(T) \geqq 2$. Therefore, by (13), there must be at least two extrema of $\psi_{n+j}(t)$ on the time interval (t_0, T). For convenience, suppose that the first extrema is a local maximum. There is no loss of generality since the indexing parameter can always be redefined. Let the control $v(t)$ on $t \in [t_0, T]$ be defined as

$$v_l(t) = 0, \qquad l \neq j, \quad l = 1, 2, \cdots, m,$$
$$v_j(t) = \begin{cases} (-1)^{i+1} h_j & \text{if } t_i \leqq t < t_i + \delta_i, \\ 0 & \text{if } t_i + \delta_i \leqq t < t_{i+1} - \delta_i, \\ (-1)^i h_j & \text{if } t_{i+1} - \delta_i \leqq t < t_{i+1}, \quad i = 1, 2, \cdots, N-1, \\ 0 & \text{elsewhere,} \end{cases} \tag{20}$$

where δ_i, $i = 1, 2, \cdots, N-1$, is given by

$$\delta_i = \begin{cases} \Delta\psi_i/(3r) & \text{if } \Delta\psi_i < 3r/h_j, \\ \Delta t_i/(h_j \beta) & \text{if } \Delta\psi_i \geqq 3r/h_j, \end{cases} \tag{21}$$

and the numbers r and β were defined in (12) and (13).

By applying the definition of Class Γ_2, it is straightforward to show that

$$\delta_i \leqq \min [1/h_j, \Delta t_i/3], \qquad i = 1, 2, \cdots, N-1. \tag{22}$$

The admissibility of the control $v(t)$ on $t \in [t_0, T]$ given by (20) then follows as a direct consequence of (22).

For this choice of admissible control the left-hand side of (15) becomes

$$\int_{t_0}^{T} \psi^2(s)v(s)\,ds = h_j \sum_{i=1}^{N-1} \left\{ (-1)^{i+1} \int_{t_i}^{t_i+\delta_i} \psi_{n+j}(t)\,dt + (-1)^i \int_{t_{i+1}-\delta_i}^{t_{i+1}} \psi_{n+j}(t)\,dt \right\}. \tag{23}$$

Let k be an odd number, $1 \leqq k \leqq N-1$, and consider the term $i = k$ of the summation in (23). Since k is odd, then $\psi_{n+j}(t_k)$ corresponds to a local maximum of $\psi_{n+j}(t)$. Therefore,

$$\begin{aligned} &\int_{t_k}^{t_k+\delta_k} \psi_{n+j}(t)\,dt \geqq \delta_k \psi_{n+j}(t_k + \delta_k), \\ &-\int_{t_{k+1}-\delta_k}^{t_{k+1}} \psi_{n+j}(t)\,dt \geqq -\delta_k \psi_{n+j}(t_{k+1} - \delta_k). \end{aligned} \tag{24}$$

However, since the derivative of $\psi_{n+j}(t)$ is bounded, then

$$\begin{aligned} \psi_{n+j}(t_k + \delta_k) &\geqq \psi_{n+j}(t_k) - r\delta_k, \\ -\psi_{n+j}(t_{k+1} - \delta_k) &\geqq -\psi_{n+j}(t_{k+1}) - r\delta_k. \end{aligned} \tag{25}$$

Let I_k represent the kth term of the summation in (23). Then, by combining (24) and (25), the following inequality for I_k is obtained:

$$I_k \geqq [\Delta\psi_k - 2r\delta_k]h_j\delta_k, \tag{26}$$

where, in this case, $\Delta\psi_k = \psi_{n+j}(t_k) - \psi_{n+j}(t_{k+1}) > 0$. Now consider the two possible values of δ_k from (21).

(i) $\delta_k = \Delta\psi_k/(3r)$. Substituting this value of δ_k into (26) yields

$$I_k \geqq h_j(\Delta\psi_k)^2/(9r). \tag{27}$$

(ii) $\delta_k = \Delta t_k/(h_j\beta)$. This yields $\delta_k \leqq 1/h_j$. Since, in this case, $\Delta\psi_k > 3r/h_j$, then $\delta_k \leqq \Delta\psi_k/(3r)$. Substituting these results into (26) yields $I_k \geqq (\Delta\psi_k)\Delta t_k/(3\beta)$. However, from (12), $\Delta t_k \geqq \Delta\psi_k/r$. Thus,

$$I_k \geqq (\Delta\psi_k)^2/(3r\beta). \tag{28}$$

The number $\beta > 0$ is defined by (13) as $\Delta t_k \leqq \beta$. Therefore, the inequality $\beta \geqq 3/h_j$ can always be satisfied for any given problem. With this inequality satisfied, it is seen that (27) implies (28). Hence, (28) is valid for both choices of δ_k.

If k is an even number, (28) can be established by similar reasoning. Substituting (28) into the summation (23) yields

$$\int_{t_0}^{T} \psi^2(s)v(s)\,ds \geqq \sum_{i=1}^{N-1} (\Delta\psi_i)^2/(3r\beta). \tag{29}$$

However, since $\psi_{n+j}(t)$ belongs to Class Γ_2, then from (14),

$$\int_{t_0}^{T} \psi^2(s)v(s)\,ds > \tilde{H}/(3r\beta). \tag{30}$$

For any number $H > 0$, let $\tilde{H} = 3r\beta H$. Since the above procedure is valid for any number $\tilde{H} > 0$, then (30) is equivalent to (15) of Lemma 1. This completes the proof of the sufficient condition for complete recoverability.

Theorems 1 and 2 are the basic results on complete recoverability. In order to apply these theorems to a given problem, the adjoint solution must be constructed as a function of the initial condition $\psi(t_0)$, with $\psi^1(t_0)$ nonzero. Then the conditions of the theorems can be checked to see if the set of recoverable initial conditions, $S(t_0)$, equals the restraint set $\hat{G}$. This procedure is illustrated by two examples in § 6.

6. Examples. An example of a system that is not completely recoverable is given below. It will be shown that this system does not satisfy the necessary condition given by Theorem 1.

Consider the following unstable system for $t \geqq 0$:

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= x_1 + u, \end{aligned} \tag{31}$$

with the control $u(t)$ bounded in both amplitude and rate for all time t. By using the transformation discussed in § 2, (31) may be reformulated as a bounded state variable process and written as

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} v, \tag{32}$$

where $x = [x_1, x_2, x_3]' = [x_1, x_2, u]'$, $v = \dot{u}$, and $[\cdot]'$ = transpose of $[\cdot]$. For this example, $n = 2$, $m = 1$ and therefore $\psi^2(t) = \psi_3(t)$. The third component of the adjoint solution is given by

$$\psi_3(t) = \eta_3 - \eta_1 + \eta_1 \cosh t - \eta_2 \sinh t, \tag{33}$$

with $\psi(0) = [\eta_1, \eta_2, \eta_3]$. Consider the particular choice of $\psi(0) = [2, 2, 2]$. For this choice, $\psi^1(0)$ is nonzero and $\psi_3(t) = 2e^{-t}$. Substituting for $\psi^2(t) = \psi_3(t)$ in (10) yields

$$\lim_{T \to \infty} \int_0^T |\psi^2(t)|\,dt = 2 < \infty. \tag{34}$$

Thus the necessary condition of Theorem 1 is violated and therefore $S(t_0) \neq \hat{G}$ in this case.

As an illustration of the sufficiency condition of Theorem 2, consider the following oscillatory system with bounded control amplitudes and rates for $t \geqq 0$:

$$\begin{aligned} \dot{x}_1 &= x_2 + u_1, \\ \dot{x}_2 &= -x_1 + u_2. \end{aligned} \tag{35}$$

The bounded state variable reformulation of (35) is given by

$$\dot{x} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} v, \tag{36}$$

where $x = [x_1, x_2, x_3, x_4]' = [x_1, x_2, u_1, u_2]'$, $v = [v_1, v_2]' = [\dot{u}_1, \dot{u}_2]'$. For this example $n = 2$, $m = 2$, and $\psi^2(t) = [\psi_3(t), \psi_4(t)]$. The third and fourth components of the adjoint solution are given by

$$\begin{aligned} \psi_3(t) &= -\eta_1 \sin t - \eta_2(1 - \cos t) + \eta_3, \\ \psi_4(t) &= \eta_1(1 - \cos t) - \eta_2 \sin t + \eta_4, \end{aligned} \tag{37}$$

with $\psi(0) = [\eta_1, \eta_2, \eta_3, \eta_4]$. It can be shown that both $\psi_3(t)$ and $\psi_4(t)$ of (37) belong to Class Γ_2, as long as $\psi^1(0) = [\eta_1, \eta_2]$ is nonzero. Thus $\psi^2(t)$ always belongs to Class Γ and, consequently, Theorem 2 is satisfied. Hence $S(t_0) = \hat{G}$ and therefore every initial condition is recoverable.

In the following section the two theorems developed in this paper will be compared to the necessary and sufficient condition for the complete recoverability of (1), in the case when instantaneous change of control is allowed.

7. Comparison with complete recoverability in the bounded amplitude case. If the constraint on the rate of change of the admissible controls, (3), is removed, then [2] gives the necessary and sufficient condition for the complete recoverability of (1) as

$$\lim_{T \to \infty} \int_{t_0}^{T} \| \lambda(t) F(t) \| \, dt = \infty, \tag{38}$$

where the n-dimensional vector $\lambda(t)$ represents any nontrivial solution to the adjoint equation $\dot{\lambda}(t) = -\lambda(t)E(t)$. The complete recoverability problem considered in this paper has included the rate constraint on the admissible controls. The reformulated $(n + m)$-dimensional problem is described by (5). The results for this problem are Theorems 1 and 2 which provide a necessary and a sufficient condition for the complete recoverability of (5), and hence (1).

A basic relationship between the two problems can be established by noting the form of the matrix $A(t)$ in (5). Since the $(n + m)$-dimensional adjoint vector $\psi(t) = [\psi^1(t), \psi^2(t)]$ satisfies the equation $\dot{\psi}(t) = -\psi(t)A(t)$, then the following two relations are established, viz.:

$$\begin{aligned} &\psi^1(t) = \lambda(t), \qquad \text{with } \psi^1(t_0) = \lambda(t_0) \quad \text{nonzero}, \\ &\dot{\psi}^2(t) = -\psi^1(t)F(t) = -\lambda(t)F(t). \end{aligned} \tag{39}$$

Certainly, every system that is completely recoverable in the case when the rate of change of the admissible controls is constrained should also be recoverable when the rate constraint is removed. Thus, if Theorem 2 is satisfied, then (38) should

also be satisfied. To see this note that from (39), (38) can be written as

$$\lim_{T\to\infty}\int_{t_0}^{T}\|\dot{\psi}^2(t)\|\,dt = \infty. \tag{40}$$

Now, suppose that Theorem 2 is satisfied and that some $\psi_{n+j}(t)$, $j = 1, 2, \cdots, m$, belongs to Class Γ_1. Since for all $T \geqq t_0$,

$$\int_{t_0}^{T}\|\dot{\psi}^2(t)\|\,dt \geqq \int_{t_0}^{T}|\dot{\psi}_{n+j}(t)|\,dt \geqq |\psi_{n+j}(T) - \psi_{n+j}(t_0)|, \tag{41}$$

then by (11) (the definitions of Class Γ_1), (41) implies that (40) is satisfied. Now suppose that some $\psi_{n+j}(t), j = 1, 2, \cdots, m$, belongs to Class Γ_2. From (12) and (13), and since $\Delta\psi_i = |\psi_{n+j}(t_{i+1}) - \psi_{n+j}(t_i)|$,

$$\int_{t_0}^{T}\|\dot{\psi}^2(t)\|\,dt \geqq \int_{t_0}^{T}|\dot{\psi}_{n+j}(t)|\,dt \geqq \sum_{i=1}^{N}\Delta\psi_i, \tag{42}$$

$$N = N(T), \quad T \geqq t_0.$$

Suppose that as $T \to \infty$, the summation in (42) converges. There are two possible cases. If $N = N(T)$ remains finite as $T \to \infty$, then (14) is not satisfied which contradicts the assumption that $\psi_{n+j}(t)$ belongs to Class Γ_2. If $N = N(T) \to \infty$ as $T \to \infty$, then $\lim_{i\to\infty}\Delta\psi_i = 0$, and therefore there exists an integer $M > 0$ such that $\Delta\psi_i \leqq 1/2$ for all $i > M$. Therefore $(\Delta\psi_i)^2 < \Delta\psi_i$ for all $i > M$. Thus by the comparison test [4], $\sum_{i=1}^{\infty}(\Delta\psi_i)^2$ converges, which again contradicts the definition of Class Γ_2.

It has been shown that if Theorem 2 is satisfied, then (38) is indeed satisfied. On the other hand it is not necessarily true that if (38) is satisfied, then Theorem 1 will be satisfied. In other words, a system may be completely recoverable with admissible controls that are bounded only in amplitude. However, the property of complete recoverability may be lost when constraints on the rate of change of the admissible controls are added. As an example, consider the following scalar system:

$$\dot{y}(t) = -y(t) + f(t)u(t), \qquad t \geqq 0, \tag{43}$$

with

$$f(t) = -e^{-t}(-1)^{k-1}\sin k\pi(t - t_{k-1}) \tag{44}$$

for $t_{k-1} \leqq t < t_k$ and $t_0 = 0$, $t_k = t_{k-1} + 1/k, k = 1, 2, \cdots$. In Appendix B, it is shown that this system satisfies (38) and yet does not satisfy Theorem 1. This occurs because of the rapidly changing sign of the function $f(t)$. If the rate of change of control is bounded, then the control cannot follow this rapid change in $f(t)$. On the other hand this problem does not exist when the rate of control is unconstrained.

8. Conclusions. For linear time-varying systems with bounded control amplitudes and rates, the set of initial states that can be steered to the origin in finite time is directly related to the augmented components of the adjoint solution to the reformulated bounded state variable processes. The two theorems presented

in this paper provide a systematic method for investigating the recoverability of these systems by testing the augmented components of the adjoint solution.

Appendix A.

A.1. Properties of $K(t, t_0)$. Consider the set of recoverability, given by

$$K(t, t_0) = \left\{ x | x = -\int_{t_0}^{t} \phi(t_0, s) B v(s)\, ds,\ v \in \Lambda(t) \right\}, \tag{A.1}$$

where $\phi(t_0, s)$ is a fundamental solution matrix and $\Lambda(t)$ is the class of admissible controls. Let the set $L(t, t_0)$ be defined as

$$L(t, t_0) = \left\{ \xi | \xi = \int_{t_0}^{t} \phi(t, s) B v(s)\, ds,\ v \in \Lambda(t) \right\}. \tag{A.2}$$

Now consider the following transformations, viz.:

$$w(s) = v(t + t_0 - s), \qquad s \in [t_0, t],\quad v \in \Lambda(t), \tag{A.3}$$

$$\zeta_{n+j}(s) = -x_{n+j}(t + t_0 - s), \qquad j = 1, 2, \cdots, m,\quad s \in [t_0, t]. \tag{A.4}$$

From (A.3) and (A.4), $d\zeta(s)/ds = w(s)$, $s \in [t_0, t]$ and $\zeta_{n+j}(t_0) = x_{n+j}(t) = 0$, $j = 1, 2, \cdots, m$. Let the set $Z(t, t_0)$ be defined as

$$Z(t, t_0) = \left\{ z | z = \int_{t_0}^{t} \phi(t, s) B w(s)\, ds \right\}, \tag{A.5}$$

where $w(s)$ on $s \in [t_0, t]$ is given by (A.3). It is shown in [5] that the set $Z(t, t_0)$, defined in (A.5), is compact, convex and varies continuously with t. Since the proof of these properties in [5] does not depend on the transformation given by (A.3) and (A.4), then the set $L(t, t_0)$ is also compact, convex and varies continuously with t. From (A.1) and (A.2), the sets $K(t, t_0)$ and $L(t, t_0)$ are related by the transformation

$$x = -\phi(t_0, t)\xi. \tag{A.6}$$

Since the two sets are related by a linear continuous transformation, then the set $K(t, t_0)$ is also compact, convex and varies continuously with t. The nondecreasing property of $K(t, t_0)$ can be established in the following manner. Suppose that the state $\tilde{x} \in K(t_1, t_0)$, where $t_1 \geqq t_0$. Then, by (A.1), there exists a control $\tilde{v} \in \Lambda(t_1)$, such that

$$\tilde{x} = -\int_{t_0}^{t_1} \phi(t_0, s) B \tilde{v}(s)\, ds. \tag{A.7}$$

Now, let the time t_2 be such that $t_2 \geqq t_1$ and define the control $\hat{v}(t)$ on $t \in [t_0, t_2]$ by

$$\hat{v}(t) = \begin{cases} \tilde{v}(t), & t_0 \leqq t \leqq t_1, \\ 0, & t_1 < t \leqq t_2. \end{cases} \tag{A.8}$$

Since $\tilde{v} \in \Lambda(t_1)$, then by the definition of the class of admissible controls, $\hat{v} \in \Lambda(t_2)$.

The corresponding state $\hat{x} \in K(t_2, t_0)$ is given by (A.1) as

$$\hat{x} = -\int_{t_0}^{t_2} \phi(t_0, s)B\hat{v}(s)\,ds = -\int_{t_0}^{t_1} \phi(t_0, s)B\tilde{v}(s)\,ds + 0 = \tilde{x}. \tag{A.9}$$

Thus the state $\tilde{x} \in K(t_2, t_0)$. Since the choices of $\tilde{x}$, t_1 and t_2, $t_0 \leqq t_1 \leqq t_2$, were arbitrary, the above is true for all states $\tilde{x} \in K(t_1, t_0)$. Thus, $K(t_1, t_0) \subset K(t_2, t_0)$, which establishes the nondecreasing property of $K(t, t_0)$.

A.2. Properties of $S(t_0)$. The set of all recoverable initial conditions is defined by

$$S(t_0) = \bigcup_{t \geqq t_0} K(t, t_0). \tag{A.10}$$

In order to show that $S(t_0)$ is convex, consider any two states $\tilde{x}$, $\hat{x} \in S(t_0)$. Then $\tilde{x} \in K(t_a, t_0)$ and $\hat{x} \in K(t_b, t_0)$ for some time instants $t_a, t_b \geqq t_0$. Let the time t_c be defined by

$$t_c = \max [t_a, t_b]. \tag{A.11}$$

Then $\tilde{x}$, $\hat{x} \in K(t_c, t_0)$ and $x \in K(t_c, t_0) \subset S(t_0)$ for all states x given by

$$x = \alpha\tilde{x} + (1 - \alpha)\hat{x} \qquad \text{for all } 0 \leqq \alpha \leqq 1. \tag{A.12}$$

Since the states $\tilde{x}$ and $\hat{x}$ and the times t_a and t_b were arbitrary, the set $S(t_0)$ is convex.

Appendix B.

B.1. Example. The scalar system under consideration is given by

$$\dot{y}(t) = -y(t) + f(t)u(t), \qquad t \geqq 0, \tag{B.1}$$

with

$$f(t) = -e^{-t}(-1)^{k-1} \sin k\pi(t - t_{k-1}) \tag{B.2}$$

for $t_{k-1} \leqq t < t_k$, and $t_0 = 0, t_k = t_{k-1} + 1/k, k = 1, 2, \cdots$.

First, suppose that the control $u(t)$ is bounded only in amplitude. Since $n = 1$, the adjoint solution $\lambda(t)$ is scalar and is given by $\lambda(t) = \lambda(0)e^t$. By direct substitution,

$$\lim_{T \to \infty} \int_0^T |\lambda(t)f(t)|\,dt = \sum_{k=1}^{\infty} 2|\lambda(0)|/(k\pi) = \infty \tag{B.3}$$

for all choices of $\lambda(0) \neq 0$. Thus, (38) is satisfied, and therefore (B.1) is completely recoverable for the case that the control u is only bounded in amplitude.

Now, suppose that a rate constraint on the control $u(t)$ is added, i.e., $|\dot{u}(t)| \leqq h$, $h > 0$, for all $t \geqq 0$. System (B.1) can be reformulated as a bounded state variable problem, i.e.,

$$\dot{x}(t) = \begin{bmatrix} -1 & f(t) \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} v(t), \tag{B.4}$$

where $x = [x_1, x_2]' = [y, u]'$ and $v = \dot{u}$. Since $m = 1$, then $\psi^2(t)$ is scalar. The

general expression for $\psi^2(t)$ is given by (see (39))

$$\psi^2(t) = \psi^2(0) + \int_0^t [-\lambda(s)f(s)]\,ds. \tag{B.5}$$

Substituting for $\lambda(s)f(s)$ in (B.5) and applying (B.2) yields the following expression for $\psi^2(t)$:

$$[\psi^2(t) - \psi^2(0)]/\psi^1(0) = \begin{cases} (1 - \cos \pi t)/\pi & \text{if } 0 \leqq t < t_1, \\ \sum_{k=1}^{N-1} (-1)^{k-1}2/(k\pi) + (-1)^{N-1}(1 - \cos N\pi(t - t_{N-1}))/(N\pi) & \\ & \text{if } t_{N-1} \leqq t < t_N, \quad N > 1. \end{cases} \tag{B.6}$$

With the above expression for $\psi^2(t)$, the left-hand side of (10) (Theorem 1) becomes

$$\lim_{T\to\infty} \int_0^T \|\psi^2(t)\|\,dt = \lim_{T\to\infty} \int_0^T |\psi^2(t)|dt = I_1 + I_2, \tag{B.7}$$

where

$$I_1 = \left(\frac{1}{\pi}\right) \int_0^{t_1} |\psi^2(0) + \psi^1(0)(1 - \cos \pi t)|\,dt \tag{B.8}$$

and

$$I_2 = \sum_{N=2}^{\infty} \int_{t_{N-1}}^{t_N} |\psi^2(0) + 2\psi^1(0) \sum_{k=1}^{N-1} (-1)^{k-1}/(k\pi) + [(-1)^{N-1}\psi^1(0)(1 - \cos N\pi(t - t_{N-1}))]/(N\pi)|\,dt. \tag{B.9}$$

By (B.2), $t_1 = 1$. Thus $I_1 \leqq H_1 < \infty, H_1 > 0$; i.e., the integral I_1 of (B.8) is bounded for all choices of $\psi(0)$ with $\psi^1(0)$ nonzero. Now consider the integral I_2 of (B.9). For a particular choice of $\psi(0)$, let $\psi^1(0)$ be arbitrary ($\psi^1(0) \neq 0$) and let

$$\psi^2(0) = -[2\psi^1(0) \log 2]/\pi. \tag{B.10}$$

Since $\sum_{k=1}^{\infty} (-1)^{k-1}/k = \log 2$, then with the aid of (B.10), (B.9) can be written as

$$I_2 = \sum_{N=2}^{\infty} \int_{t_{N-1}}^{t_N} |-2\psi^1(0) \sum_{k=N}^{\infty} (-1)^{k-1}/(k\pi) + (-1)^{N-1}\psi^1(0)(1 - \cos N\pi(t - t_{N-1}))/(N\pi)|\,dt. \tag{B.11}$$

Applying Leibniz's alternating series test [4] to (B.11), and noting that $t_N - t_{N-1} = 1/N$, yields

$$I_2 \leqq \sum_{N=2}^{\infty} \{|2\psi^1(0)/(N^2\pi)| + |(-1)^{N-1}\psi^1(0)/(N^2\pi)|\} \leqq H_2 < \infty, \; H_2 > 0. \tag{B.12}$$

Substituting the result from (B.12) along with the result that $I_1 \leqq H_1 < \infty$ into

(B.7) yields

$$\lim_{T\to\infty}\int_0^T \|\psi^2(t)\|\,dt = I_1 + I_2 \leqq H_1 + H_2 = H < \infty, \qquad H > 0, \tag{B.13}$$

for these particular choices of $\psi(0) = [\psi^1(0), \psi^2(0)]$, with $\psi^1(0) \neq 0$. Thus, (10) of Theorem 1 does not hold and therefore the theorem is violated.

System (B.1) is then not completely recoverable in the case when the admissible controls are constrained both in amplitude and rate, although it is completely recoverable when the rate constraint is removed. Notice that the sufficiency condition (Theorem 2) is also not satisfied for (B.1), since $\psi^2(t)$ does not belong to Class Γ.

REFERENCES

[1] J. P. LaSalle, *The optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. V, Annal. Math. Studies No. 45, Princeton University Press, Princeton, 1960, pp. 1–24.

[2] J. L. LeMay, *Recoverable and reachable zones for control systems with linear plants and bounded controller outputs*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 346–354.

[3] J. Y. S. Luh and J. S. Shafran, *Minimal time control of linear systems with control amplitude and rate saturations*, Rep. TREE 68-33, School of Electrical Engineering, Purdue University, Lafayette, Indiana, 1968.

[4] R. G. Bartle, *The Elements of Real Analysis*, John Wiley, New York, 1966.

[5] W. W. Schmaedeke and D. L. Russell, *Time optimal control with amplitude and rate limited controls*, this Journal, 2 (1964), pp. 373–395.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

A FREQUENCY CRITERION FOR OSCILLATORY SOLUTIONS*

GERALD S. LELLOUCHE†

1. We examine the solutions of a certain integral equation (to be given below). We assume that the solutions exist locally and show that if a certain frequency criterion is satisfied that the solutions exist globally and either have an infinite number of zeros on the line $[0, \infty)$ or else are asymptotically zero.

Consider the solutions of the integral equation

$$\int_0^t \sigma(x)\,dx = f(t) - f(0) + \int_0^t m(t-x)\varphi[\sigma(x)]\,dx, \tag{1}$$

where the real functions $f(t)$, $m(t)$ and $\varphi(\sigma)$ satisfy the following conditions[1]:

(i) $f(t)$ is continuous and defined for $t \geqq 0$, and $f'(t)$, $f''(t)$ and $f'''(t)$ are in $L_1[0, \infty)$. It follows that

$$f^N(t) = -\int_t^\infty f^{N+1}(x)\,dx. \tag{2a}$$

(ii) The kernel $m(t)$ continuous and defined for $t \geqq 0$ is given by

$$m(t) \equiv j(t) - j(0) - \rho t, \tag{2b}$$

$$\rho > 0, \tag{2c}$$

where $j(t)$, $j'(t)$ and $j''(t)$ are defined by the relation

$$j^N(t) = -\int_t^\infty j^{N+1}(x)\,dx \tag{2d}$$

for $t \geqq 0$, and they are in $L_1[0, \infty) \cap L_2[0, \infty)$.

(iii) $\varphi(\sigma)$ is a bounded and locally Lipschitzian function of σ and satisfies

$$\sigma\varphi(\sigma) > 0, \qquad \sigma \neq 0, \tag{2e}$$

$$\varphi(0) = 0. \tag{2f}$$

We suppose now that:

(iv) If

$$J_1(i\omega) \equiv \int_0^\infty j'(\tau)e^{-i\omega\tau}\,d\tau, \tag{2g}$$

there exist two real numbers $q_1, q_2 \geqq 0$ such that

$$\operatorname{Re}\{Z(i\omega)\} \equiv \operatorname{Re}\left\{(1 + i\omega q_1)J_1(i\omega) - \rho q_1 + \frac{q_2}{i\omega}(J_1(i\omega) - J_1(0))\right\} \leqq 0 \tag{3}$$

for all real $\omega \neq 0$.

* Received by the editors January 23, 1969, and in revised form October 7, 1969.

† Nuclear Engineering Department, Brookhaven National Laboratory, Upton, Long Island, New York 11973. This work was supported by the United States Atomic Energy Commission.

[1] While we avoid the question of demonstrating local existence, conditions (i)–(iii) are sufficiently stringent to assure that solutions do exist. The method of successive approximations may be used to demonstrate that such solutions exist.

(v) The number $q_2 \geqq 0$ also satisfies

$$\rho - q_2 J_1(0) > 0. \tag{4}$$

Then we have the following theorem.

THEOREM 1. *If conditions* (i)–(v) *are satisfied, all solutions* $(\sigma(t))$ *of* (1) *are defined for all* $t \geqq 0$ *and with the following properties*: *either*

(a) $\lim_{t\to\infty} \sigma(t) \to 0$ *or*

(b) $\sigma(t)$ *has an infinite number of zeros on the line* $0 \leqq t < \infty$.

Digression. If we differentiate (1) we find

$$\sigma(t) = f'(t) + \int_0^t m'(t-x)\varphi[\sigma(x)]\,dx; \tag{1'}$$

this equation is the one studied by Corduneanu [1], and in particular if $q_2 = 0$, then Corduneanu has shown that (3) is a sufficient condition for asymptotic stability in the large. We do not consider the question of whether conditions (i)–(v), with $q_2 > 0$, are sufficient to prove asymptotic stability in the large.

Proof of Theorem 1. The method of proof is that due to Popov [2]. Set

$$\varphi_t(\tau) \equiv \varphi[\sigma(\tau)],^2 \qquad 0 \leqq \tau \leqq t, \tag{5a}$$

$$\varphi_t(\tau) = 0, \qquad \tau > t. \tag{5b}$$

Consider the function defined for $\tau \geqq 0$

$$\lambda_t(\tau) = \int_0^\tau [j'(\tau - x) + q_2 j''(\tau - x) + q_2 j(\tau - x)]\varphi_t(x)\,dx + q_1 m'(0)\varphi_t(\tau). \tag{6a}$$

Since for $\tau > t$

$$\lambda_t(\tau) = \int_0^t [j'(\tau - x) + q_1 j''(\tau - x) + q_2 j(\tau - x)]\varphi_t(x)\,dx, \tag{6b}$$

by condition (ii) we see that $\lambda_t(\tau)$ is in $L_1[0, \infty) \cap L_2[0, \infty)$, and we have

$$\Lambda_t(i\omega) = \int_0^\infty \lambda_t(\tau)e^{-i\omega\tau}\,d\tau = Z(i\omega)\tilde{\varphi}_t(i\omega), \tag{7a}$$

where

$$Z(i\omega) = (1 + i\omega q_1)J_1(i\omega) - \rho q_1 + \frac{q_2}{i\omega}[J_1(i\omega) - J_1(0)], \tag{7b}$$

$$\tilde{\varphi}_t(i\omega) = \int_0^\infty \varphi_t(\tau)e^{-i\omega\tau}\,d\tau = \int_0^t \varphi[\sigma(\tau)]e^{-i\omega\tau}\,d\tau. \tag{7c}$$

Now consider the real function

$$\eta(t) = \int_0^t \lambda_t(\tau)\varphi[\sigma(\tau)]\,d\tau = \int_0^\infty \lambda_t(\tau)\varphi_t(\tau)\,d\tau$$

---

[2] Condition (iii) guarantees that $\varphi[\sigma(t)]$ is locally in L_2, hence that $\varphi_t(\tau)$ is in L_2 for all $t \geqq 0$.

whence, by Parsevals relation,

$$\eta(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} Z(i\omega)|\tilde{\varphi}_t(i\omega)|^2 \, d\omega. \tag{8a}$$

Since $\eta(t)$ is real, Im $[Z(i\omega)]$ is antisymmetric in ω and

$$\eta(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \text{Re}\,[Z(i\omega)]|\tilde{\varphi}_t(i\omega)|^2 \, d\omega \tag{8b}$$

and, by condition (iv),

$$\eta(t) \leqq 0. \tag{8c}$$

Starting from the real form for $\lambda_t(\tau)$ we find (using (1, 1′) and the derivative of (1′) as well):

$$\begin{aligned} &\int_0^t \varphi[\sigma(x)]\sigma(x)\,dx + q_1 \int_0^\sigma \varphi(\sigma')\,d\sigma' + q_2 \int_0^t \varphi[\sigma(x)] \int_0^x \sigma(y)\,dy \\ &+ q_2\rho \int_0^t \varphi[\sigma(x)]\,dx \int_0^x (x-y)\varphi[\sigma(y)]\,dy + \frac{(\rho + q_2 j(0))}{2}\left[\int_0^t \varphi[\sigma(x)]\,dx\right]^2 \\ &- \int_0^t g(x)\varphi[\sigma(x)]\,dx - q_1 \int_0^{\sigma(0)} \varphi(\sigma')\,d\sigma' \leqq 0, \end{aligned} \tag{9}$$

where

$$g(x) \equiv f'(x) + q_1 f''(x) + q_2(f(x) - f(0)).$$

From condition (iii) each of the first two integrals is nonnegative. If $\sigma(t)$ (hence $\varphi[\sigma]$) has only a fixed sign then the third and fourth integrals are always positive. However if $\sigma(t)$ changes sign then the third and fourth integrals are not necessarily positive. We return to this in a moment. Consider the last three integrals: let $\Phi(t) = \int_0^t \varphi[\sigma(x)]\,dx$; then $\left|\int_0^t g(x)\varphi[\sigma(x)]\,dx\right| \leqq K \sup|\Phi(\chi)|$, $0 \leqq \chi \leqq t$, where K depends only on $f(t)$ and its derivatives; hence the last three integrals sum to a value greater than

$$\frac{1}{2}[\rho + q_2 j(0)]\Phi^2(t) - K \sup|\Phi(\chi)| - q_1 \int_0^{\sigma(0)} \varphi(\sigma')\,d\sigma', \quad 0 \leqq \chi \leqq t.$$

Since $\Phi^2(t)$ attains $(\sup|\Phi|)^2$ somewhere on the line $[0, t]$ (if $\varphi[\sigma]$ has a fixed sign, $\Phi(t)$ is maximum at the endpoint of the line) we see that by conditions (i), (ii) and (v) the last three terms sum positive for $t >$ some t^* unless $\varphi[\sigma(t)] \to 0$. We find then that the first four integrals and the sum of the last three are positive, contradicting the inequality; hence $\sigma(t)$ cannot have a fixed sign unless $\sigma(t)$ goes asymptotically to zero.

Suppose $\sigma(t)$ has only a fixed number of sign changes, and for $t > t_1$ achieves a fixed sign. This cannot occur either, since $\int_{t_1}^t \varphi(\sigma(x))\,dx$ becomes unbounded (unless $\lim_{t\to\infty} \sigma(t) \to 0$); hence the last three terms will again sum to a positive

value for $t > t^* > t_1$. Furthermore the third integral can be written in the form

$$I_3 \equiv \int_{t_1+\varepsilon}^{t} \varphi[\sigma(x)]\,dx \int_{t_1+\varepsilon}^{x} \sigma(y)\,dy + \int_0^{t_1+\varepsilon} \sigma(y)\,dy \int_{t_1+\varepsilon}^{t} \varphi[\sigma(x)]\,dx + \int_0^{t_1+\varepsilon} \varphi[\sigma(x)]\,dx \int_0^{x} \sigma(y)\,dy, \tag{10}$$

where $\varepsilon > 0$ is defined such that $|\sigma(t)| > \eta$ for $t > t_1 + \varepsilon$ and similarly $|\varphi[\sigma(t)]| > \delta$ for $t > t_1 + \varepsilon$. Assume $\sigma > 0$ for $t > t_1$; then for $t > t_1 + \varepsilon$ we have the following lower estimate for I_3:

$$I_3 > \frac{(t - t_1 - \varepsilon)^2}{2}\eta\delta + \delta(t - t_1 - \varepsilon)\int_0^{t_1+\varepsilon} \sigma(y)\,dy + \int_0^{t_1+\varepsilon} \varphi[\sigma(x)]\,dx \int_0^{x} \sigma(y)\,dy;$$

for t sufficiently large this term is and remains thereafter positive (similarly if $\sigma < 0$ for $t > t_1$). An equivalent analysis can be done for the fourth integral. Since for t sufficiently large we violate the inequality (9), $\sigma(t)$ cannot have a finite number of zeros on $[0, \infty)$. Suppose however $\sigma(t)$ diverges at a finite time t_f after a finite number of zeros. Equation (1′) shows that if $\sigma(t)$ diverges at $t_f < \infty$ then (since $m'(t)$ is bounded by condition (ii)) both $\varphi(\sigma)$ and $\int_0^t m'(t - x)\varphi(\sigma(x))\,dx$ diverge as $t \to t_f$; from conditions (i), (ii) and (1′) we also find

$$|\sigma(t_f)| < K_1 + K_2 \int_0^{t_f} |\varphi[\sigma(x)]|\,dx,$$

and if the last zero of $\sigma(t)$ occurs at $t = t_1 < t_f$ then (taking σ to diverge positively) $\int_{t_1}^{t_f} \varphi[\sigma(x)]\,dx$ diverges as well (similarly if $t_1 + \varepsilon < t_f$, $\int_{t_1+\varepsilon}^{t_f} \varphi[\sigma]\,dx \to \infty$). We now consider the sign of the third and fourth integrals.

$$\begin{aligned} I_3 &= \int_0^{t_f} \varphi[\sigma(x)]\,dx \int_0^{x} \sigma(y)\,dy \\ &= \int_{t_1}^{t_f} \varphi\,dx \int_{t_1}^{x} \sigma\,dy + \int_0^{t_1} \sigma(y)\,dy \int_{t_1}^{t_f} \varphi\,dx + \int_0^{t_1} \varphi\,dx \int_0^{x} \sigma\,dy; \end{aligned}$$

if we take the sign of σ (hence of $\varphi[\sigma]$) to be >0 for $t_f > t > t_1$ then the first term on the right is >0. Now we add the second integral in (9) to the third and find

$$\begin{aligned} I &\equiv \int_0^{t_f} \varphi[\sigma]\sigma\,dx + q_2 \int_0^{t_f} \varphi[\sigma]\,dx \int_0^{x} \sigma\,dy \\ &> \int_{t_1}^{t_f} \varphi[\sigma]\,dx\left[\sigma(x) + q_2 \int_0^{t_f} \sigma(y)\,dy\right] + q_2 \int_0^{t_1} \varphi\,dx \int_0^{x} \sigma\,dy. \end{aligned}$$

The integral $\int_0^{t_1} \sigma(y)\,dy$ is a constant while $\sigma(x)$ is of fixed sign for $t_1 < t < t_f$ and diverges for $t \to t_f$; hence $\sigma(x) + q_2 \int_0^{t_1} \sigma(y)\,dy$ attains the same fixed sign that

$\varphi[\sigma]$ has for some $t_f > t = t_2 \geqq t_1$ and diverges as $t \to t_f$. Now for $t > t_2 + \varepsilon$ we find that if $\sigma(x) + q_2 \int_0^{t_1} \sigma(y)\,dy > \eta$ then $I > \eta \int_{t_1+\varepsilon}^{t_f} \varphi[\sigma]\,dx + q_2 \int_0^{t_1} \varphi\,dx \int_0^x \sigma\,dy$ and the sum of the second and third integrals diverges. The fourth integral can be handled as follows:

$$I_4 \equiv \int_0^{t_f} \varphi\,dx \int_0^x (x - y)\varphi\,dy$$

$$= \int_0^{t_1} \varphi\,dx \int_0^x (x - y)\varphi\,dy + \int_{t_1}^{t_f} \varphi\,dx \int_0^{t_1} (x - y)\varphi\,dy + \int_{t_1}^{t_f} \varphi\,dx \int_{t_1}^x (x - y)\varphi\,dy.$$

The third term is positive since $\varphi[\sigma]$ has a fixed sign between t_1 and t_f; the first term is constant. We add the second term to I and find a new term in the bracket of the form

$$\int_{t_1}^{t_f} \varphi(x)\,dx \left[\sigma(x) + q_2 \int_0^{t_1} \sigma(y)\,dy + q_2\rho \int_0^{t_1} (x - y)\varphi\sigma(y)\,dy \right].$$

Again, since $\sigma(x)$ diverges and the other two terms are bounded, there is a $t_f > t_3 \geqq t_1$ for which the sign of the bracket becomes fixed and as before we find that the sum of the integrals becomes positive, in violation of (9); hence a finite escape time cannot exist. This proves the theorem since $\sigma(t)$ must either be a continuously oscillating function, finite for all t finite, or else go to zero asymptotically.

2. An application to linear systems. We assume $\varphi(\sigma) = a\sigma$; in this case the first three integrals of (9) are each $\geqq 0$. Theorem 1 states that any nonasymptotically zero solution oscillates. If the fourth integral in (9) can be shown to be greater than $-A < 0$ $(A < \infty)$ on some sequence $\{t_n\}$, then on that sequence $\{t_n\}$ the first three integrals diverge positively, and for $n > N$ a violation of Theorem 1 is reached. By direct substitution in the fourth integral, solutions of the form $\sin(\omega t)e^{\lambda t}$ show such a violation; hence no such solution can exist (independently of $\omega > 0$ or $\lambda \geqq 0$). Solutions of the form $e^{\lambda t} \sum_n a_n \sin \omega_n t$ can however occur and need not violate the criterion if $\sum_n a_n/\omega_n = 0$ (even for $\lambda = 0$). Hence Theorem 1 does not guarantee boundedness.

REFERENCES

[1] C. CORDUNEANU, *Concerning an integral equation occurring in automatic control theory*, C. R. Acad. Sci., Paris, 256 (1963), pp. 3564–3567.

[2] V. M. POPOV, *Absolute stability of nonlinear systems of automatic control*, Automat. Control, 22 (1962), pp. 857–875.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

A NONGRADIENT AND PARALLEL ALGORITHM FOR UNCONSTRAINED MINIMIZATION*

D. CHAZAN AND W. L. MIRANKER†

Abstract. The purpose of this paper is to describe an algorithm for unconstrained optimization which is suitable for execution on a parallel computer. A nongradient method similar in nature to Powell's method is used and it is shown that the algorithm terminates at the minimum for quadratics and converges for strictly convex twice continuously differentiable functions.

1. Introduction. In this paper we consider an algorithm suitable for unconstrained minimization of strictly convex functions. The method is a nongradient method since it requires no information about derivatives. In addition, it has been designed as a parallel method and may be executed simultaneously on a set of arithmetic processors. Parallelism has recently attracted some attention in various fields of computation and computer science, e.g., [4], [5], [6]. This paper gives a procedure for utilizing parallelism in the area of unconstrained minimization.

Algorithms for unconstrained minimization usually proceed by a sequence of univariate minimizations. The directions in which to make a univariate search typically depend on the gradient at the current point. Steepest descent [1], conjugate gradient [1], and variable metric [2] are examples of such methods. Since gradient computation is costly in certain computational situations, methods which do not compute gradients have been devised. Apart from those methods which estimate gradients by finite differences, such nongradient methods include pattern search [1] and a conjugate method devised by Powell [3]. This latter method has been the motivation for the study presented here.

We will now give a brief description of Powell's minimization method which requires no derivative computation. In m-space we are given a point p and m directions $v^1, \cdots, v^m$. Starting at p we make m univariate minimizations in sequence in the directions $v^1, \cdots, v^m$, respectively. This procedure produces a polygonal trajectory which terminates at a point, q, say. We now make one more univariate minimization starting at q and in the direction $v^{m+1} = q - p$. Let r be the point which this last minimization results in. This is one cycle in the algorithm. To execute the next cycle, we update p and $v^1, \cdots, v^m$ as follows: $r \to p$ and $v^{i+1} \to v^i$, $i = 1, \cdots, m$.

Figure 1.1 illustrates two cycles of the algorithm in 3-space. Arrows denote univariate minimizations and are labeled by their directions.

Powell tried out his algorithm and obtained convergence in several examples. However, he did not produce a proof of convergence for his algorithm. He does produce a discussion of the method when it is applied to a function $f(x)$ of the form

$$f(x) = xAx + bx + c, \tag{1.1}$$

where c is a scalar, b is an m-vector, and A an $m \times m$ positive, definite, symmetric

* Received by the editors September 24, 1968, and in revised form October 10, 1969.

† Thomas J. Watson Research Center, International Business Machines Corporation, Yorktown Heights, New York 10598.

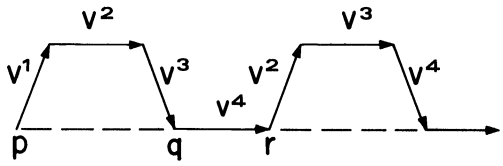


FIG. 1.1

matrix. In this discussion Powell claimed that his algorithm terminates in at most m cycles at the minimum of $f(x)$.

Except for a small gap, this discussion is a proof of convergence in this case. Subsequently, W. Zangwill [6] modified Powell's algorithm and filled the gap in Powell's argument for quadratic functions. Moreover, Zangwill produced a convergence proof for this modification in the general case when $f(x)$ is a strictly convex function. Zangwill's algorithm is identical to Powell's, except that each cycle of minimizations is augmented by an additional minimization in a coordinate direction. The coordinate directions are chosen in cyclical order.

The methods of Powell and of Zangwill, as well as most other minimization methods, proceed by means of a sequence of univariate minimizations. Thus they are, of necessity, sequential computations. If we have at our disposal a computer with a set of arithmetic processors capable of simultaneous operation, and if we seek to exploit this computer to improve the speed of execution of a typical sequential minimization algorithm, we will, in general, be unable to do so. The method discussed in this paper proceeds by simultaneous univariate minimization, with simultaneity of degree as high as the dimension of the problem, and so is appropriate for exploitation of the type of parallel computer in question.

We will now sketch our parallel algorithm for unconstrained minimizations.

First we choose a set U of m linearly independent unit vectors $u_1, \cdots, u_m$. At each cycle of the algorithm we will select a vector from this set in cyclical order. We are given a point p and $m-1$ vectors $v^1, \cdots, v^{m-1}$. We select a vector from U and call it v^m. Starting at p, construct a polygonal path by stringing together the m vectors $v^1, \cdots, v^m$ (no minimizations). Then from each vertex of this polygonal path (excluding its initial point p), perform m simultaneous (in parallel) univariate minimizations in the common direction v^1. These minimizations produce a displacement $\alpha_i v^1$, $i = 1, \cdots, m$, from the vertex $p + \sum_{j \leq i} v^j$, $i = 1, \cdots, m$, of the polygonal path. Now update p and $v^1, \cdots, v^{m-1}$ as follows. Let $p + v^1 + \alpha_1 v^1$ be the updated p. Let $v^{i+1} + (\alpha_{i+1} - \alpha_i)v^1$ be the updated v^i, $i = 1, \cdots, m-1$. This completes one cycle of the algorithm.

Figure 1.2 illustrates one cycle in the algorithm. Line segments with arrowheads denote minimizations. The updated p and the updated v^i are given an asterisk. In the schematic, v^3 is chosen as u_j so that $(v^3)^*$ must be u_{j+1}. As above, parallel and nonparallel lines may appear nonparallel or parallel, respectively, in the schematic. For example, the broken line from p to p^* given by $p + v^1 + \alpha_1 v^1$ is, in fact, a straight-line segment.

It is seen that our algorithm uses features found in Powell's method as well as a feature resembling the modification introduced by Zangwill. However, it is quite different from either of these methods and has as its principal objective to increase the speed of a calculation by executing simultaneous or parallel

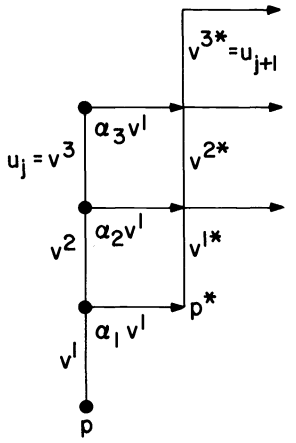


FIG. 1.2

minimizations. In the case that $f(x)$ is a quadratic in m-space our algorithm is finitely convergent and employs m^2 minimizations to locate the minimum. In the cases that they are finitely convergent, Powell's method requires m^2 minimizations also while Zangwill's requires $(m+1)^2$. (It is a simple matter to modify all of these methods to cut these numbers in half.) However, the latter two methods are purely sequential. If we assume that a univariate functional minimization takes one unit of time, the latter two methods (suitably modified) require approximately m^2 time steps. The method in this paper is a parallel method and requires only m time steps to find the minimum of a quadratic. Thus there is a gain in speed of computation which is linear in the number of processors used for minimizing a quadratic function.

We have not produced an estimate of the increase in speed of calculation for our algorithm for the nonquadratic case. This remains an open question.

In § 2 we define the parallel minimization algorithm in m-space. We then consider the case of quadratic functions and show that the algorithm converges in finitely many steps. In § 3 we give an outline of the proof of convergence of the algorithm for the class of locally convex functions. In §§ 4, 5 and 6 the technical details of the proof are furnished.

Remark on parallel operation. It is possible that in one cycle of m simultaneous univariate minimizations one minimization is much more difficult than the $m-1$ others. This is a cost parallelism which in practice may be reduced in various ways.

2. Description of algorithm and the case of quadratic functions. Let $f(x)$ denote the function to be minimized. Then the parallel nongradient minimization algorithm is given formally as follows.

Let $U = \{u_i, i = 1, \cdots, m\}$, where the u_i are linearly independent m-vectors.

Let $\beta_r, r = 1, \cdots$, be a sequence of positive scalars tending to zero. Let $w_n^m = u_i$ if $n \equiv i \bmod m, i = 1, \cdots, m$.

Let $p_n^1, n = 1, \cdots$, be a sequence of points in m-space, and let v_{n+j}^j, $j = 1, \cdots, m; n = 1, \cdots$, be a sequence of m-vectors. For each $n, n = 1, \cdots$, a step in the algorithm is given by a mapping

$$(2.1) \qquad (p_n^1, v_{n+1}^1, v_{n+2}^2, \cdots, v_{n+m}^m) \to (p_{n+1}^1, v_{n+2}^1, v_{n+m}^2, \cdots, v_{n+m+1}^m)$$

defined as follows.

Determine the scalars α_{n+1}^j, $j = 1, \cdots, m$, by performing (simultaneously) the m univariate minimizations

$$\min_{\alpha_{n+1}^j} f\left(p_n^1 + \sum_{i \leqq j} v_{n+i}^i + \alpha_{n+1}^j v_{n+1}^1\right), \qquad j = 1, \cdots, m. \tag{2.2}$$

Then the mapping (2.1) is defined by

$$\begin{aligned} p_{n+1}^1 &= p_n^1 + (1 + \alpha_{n+1}^1)v_{n+1}^1, \\ v_{n+j}^j &= (\alpha_n^{j+1} - \alpha_n^j)v_n^1 + v_{n+j}^{j+1}, \qquad j = 1, \cdots, m-1, \\ v_{n+m}^m &= \beta_{n+m} w_n^m. \end{aligned} \tag{2.3}$$

The algorithm is schematized in Fig. 2.1.

We may note that if v_n^i turns out to be zero v_n^1 would be zero and our algorithm (as defined) would simply imply that $p_{n+1}^1 = p_n^1$ and $v_{n+1}^j = v_n^j$.

If $f(x)$ is a quadratic given by (1.1) with A positive definite, it is easy to see that the algorithm converges in at most m steps. Consider to this end the following two lemmas found in [3].

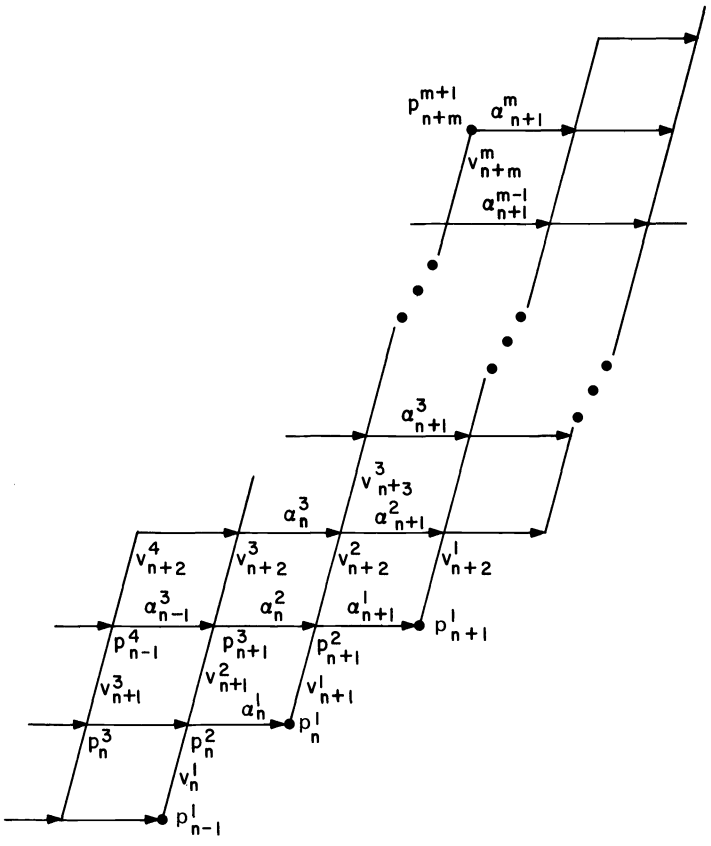


FIG. 2.1

LEMMA 2.1. *If $q^1, \cdots, q^k$, $k \leqq m$, are mutually conjugate directions, then the minimum of the quadratic function $f(x)$, where x is a general point in the k-dimensional space containing the directions $q^1, \cdots, q^k$, may be found by searching along each of the directions once only and, moreover, in any order.*

LEMMA 2.2. *If y and z are the locations of minima of $f(x)$ in a space containing the direction q, then the vector $y - z$ is the conjugate to q.*

Referring to Fig. 2.1, we see that the vector v^1_{n+1} is conjugate to v^1_{n-j}, $j = 0, \cdots, m-2$. This follows from Lemma 2.2 since v^1_{n+1} is determined by two points which are locations of minima of $f(x)$ in a space containing v^1_{n-j}, $j = 0, \cdots, m-2$. Thus any sequence of m vectors v^1_{r+j}, $j = 1, \cdots, m$, $r \geqq 0$, are mutually conjugate. Lemma 2.1 then implies that every point p^1_n, $n \geqq m$, lies at the minimum of $f(x)$.

Notice that the proof is identical to Powell's; but, whereas it is insufficient for his own algorithm, it is complete for ours.

3. Outline of the proof of convergence. We have seen that the parallel minimization scheme defined in the previous section locates the minimum of $f(x)$ in a finite number of steps if $f(x)$ is quadratic. The purpose of this and the following sections will be to demonstrate convergence of the sequence of iterates to the minimum of f when f is twice continuously differentiable, strictly convex, and tends to ∞ as $\|x\| \to \infty$. It is not at all clear that the strict convexity is essential for convergence. Indeed, a close relative of this method, the method of coordinate search, does not require such an assumption. This difficulty stems from the fact that at some point in the procedure we have to guarantee that the search directions keep spanning the whole space. This is always true for the method of coordinate search. To prove this fact for our scheme, we shall use the strict convexity of $f(x)$ to "localize" a finite sequence of steps around some point and use linearizations of the function near this point to obtain the spanning property.

Since the proof, while elementary, is composed of a large number of disjoint parts, we shall present here a short summary to serve as a guideline.

Let us start by noting that the sequence of points p^1_n is a descending sequence; i.e., $f(p_{n+1}) \leqq f(p_n)$. Then proceeding by contradiction, we shall assume that $f(p^1_n) > a_1 > \min_p f(p)$. Thus there exists some value a_2 so that $f(p^1_n) \downarrow a_2$ and the sequence p^1_n is bounded between the two contour surfaces $f(p) = a_2$ and $f(p) = a_2 + \varepsilon_n$ with $\varepsilon_n \to 0$. Clearly, the accumulation points of p^1_n lie on $\{p : f(p) = a_2\}$. Let p_0 be one such point, and let g_0 be the gradient at p_0. Also, let g_n be the gradient at p^1_n.

The main part of the proof will consist of showing that there exist a positive integer j_0 and a positive scalar α so that for each m-vector v,

$$\max_{1 \leqq j \leqq j_0} \frac{\langle v, v^1_{n+j} \rangle}{\|v\| \, \|v^1_{n+j}\|} > \alpha.$$

From this it will follow that any sequence of $j_0 \geqq m$ successive v^1_n (search directions) spans the space. We combine this with the fact that

$$\frac{\langle v^1_n, g_n \rangle}{\|v^1_n\| \, \|g_n\|} = 0$$

and that g_{n+j}, $j = 1, \cdots, j_0$, can be made arbitrarily close to g_0 by continuity of the gradient to conclude that $\langle v^1_{n+j}, g_0 \rangle / (\|v^1_n\| \, \|g_0\|)$, $j = 1, \cdots, j_0$, has zero as a

limit point. This contradicts the fact just cited with g_0 taking the place of v, that the normalized projection of any vector onto v_n^1 is greater than $\alpha > 0$.

As noted above, the critical part of the argument presented is the spanning property of the vectors v_{n+j}^1, $1 \leqq j \leqq j_0$, for some j_0. If $f(\cdot)$ were a positive definite quadratic form, the vectors v_{n+j}^1, $1 \leqq j \leqq m$, would be pairwise conjugate and would, therefore, certainly have the spanning property. In general, this is not the case. However, suppose we could localize a block of Fig. 2.1, i.e., a collection of points $\{p_{n+j}^i : 1 \leqq j \leqq j_0, 1 \leqq i \leqq m\}$, inside a neighborhood N of the accumulation point p_0. We could then approximate $f(\cdot)$ by a quadratic consisting of the first three terms in its Taylor series about p_0 and obtain approximate conjugacy statements.

Conjugacy here of $v_{n+j_1}^1$ and $v_{n+j_2}^1$ is taken to mean $\langle v_{n+j_1}^1, f_{xx}(p_0) v_{n+j_2}^1 \rangle = 0$, where $f_{xx}(p_0)$ is the Hessian matrix of f at p_0. Thus, to conclude the proof, two facts have to be verified:

(a) A "block" of the p_j^i can be put into an arbitrarily small neighborhood of p_0;

(b) If the p_j^i are sufficiently close to p_0, the search directions v_{n+j}^1, $1 \leqq j \leqq m$, are approximately conjugate.

Statement (b) as it stands turns out not to be true. There may be freak circumstances where $v_{n+j_1}^1$ and $v_{n+j_1+j_2}^1$, $1 \leqq j_2 < m$, are not even approximately conjugate. It is, however, always true that v_n^1 is (approximately) conjugate to v_{n+1}^j, $1 \leqq j \leqq m$. For $j = 1$ this last property asserts that successive search directions are (approximately) conjugate. This, together with the fact that the v_n^m have the spanning property, is enough to induce the spanning property for v_{n+j}^1, $1 \leqq j \leqq 2m - 1$.

The proof of statement (a) requires the following assertions:

(a1) The sequence $\|v_n^i\| \to 0$ as $n \to \infty$, $1 \leqq i \leqq m$;

(a2) $\|p_{n+1}^1 - p_n^1\| \to 0$ as $n \to \infty$.

From (a2) and the fact that p_0 is a limit point of p_n^1, we are able to force the whole sequence p_{n+j}^1, $1 \leqq j \leqq j_0$, for any fixed j_0, into an arbitrarily small neighborhood of p_0 infinitely often. Combining this with (a1), we are assured that the p_{n+j}^i, $1 \leqq i \leqq m$, $1 \leqq j \leqq j_0$, lies in N also.

The next section will be devoted to the statement of an elementary theorem which will allow us to obtain (a1). This uses the fact that the $v_n^m \to 0$, which in turn follows from the fact that the $\beta_n \to 0$. This theorem will also be used in § 6 to obtain the approximate conjugacy of v_n^1 to the v_{n+1}^j, $1 \leqq j \leqq m$.

In § 5 we shall prove (a2) using the strict convexity of f and the fact that p_n is in the strip $\{p : a \leqq f(p) \leqq a + \varepsilon_n\}$.

In § 6 we shall combine all these facts together to obtain the convergence statement.

4. Proof of $\|v_n^i\| \to 0$ as $n \to \infty$. As noted above, the argument demonstrating convergence depends on the approximate conjugacy of successive minimization directions. We shall now state a lemma which will formalize the notion of approximate conjugacy, and apply it to deduce that $v_n^i \to 0$ (assertion (a1)).

We recall that the function $f(x)$ is twice continuously differentiable, strictly convex and tends to infinity as $\|x\| \to \infty$.

LEMMA 4.1. *Let $f(x)$ be a twice differentiable and strictly convex function of the m-vector x and let $f(x) \to \infty$ as $\|x\| \to \infty$. Let v be a given m-vector and let S be the subspace orthogonal to v. Let $g: S \to R^1$ be defined as follows:*

$$f(w + g(w)v) \leqq f(w + kv)$$

for all real k. Then

(a) *$g(w)$ is a well-defined differentiable function;*

(b) $\langle v, f_{ww}(w + g(w)v\Delta w\rangle + \langle v, f_{ww}(w + g(w)v)v\Delta g\rangle + h(\Delta w) = 0$, *where $|h(\Delta w)| = o(\Delta w)$.*

Furthermore, if S' is the surface, $S' = \{vg(w) + w : w \in R^m\}$ and q_1, q_2 are two points on S' which have the form $w_1 + g(w_1)v$ and $w_1 + g(w_2)v$, respectively, then

$$\text{(c)} \qquad \left| \frac{\langle v, f_{ww}(q_1)(q_1 - q_2)\rangle}{\|v\| \, \|w_1 - w_2\|} \right| \leqq O(\|w_1 - w_2\|).$$

We may note that this also implies that

$$\langle v, f_w(q_1)(q_1 - q_2)\rangle \leqq O(q_1 - q_2)\|q_1 - q_2\| \, \|v\|$$

since

$$\|q_1 - q_2\|^2 = \|w_1 - w_2\|^2 + \|v\|^2 g^2(w) = \|w_1 - w_2\|^2.$$

Proof. $f(w + kv)$ is strictly convex in k and tends to ∞ as $k \to \infty$. Thus it has a unique, well-defined minimum at $g(w)$. Let $G(w) = f_w(w)$. Then $g(w)$ satisfies the equation $\langle v, G(w + g(w)v)\rangle = 0$. Since $G_w(w + kv)$ is positive definite, it follows from the implicit function theorem that $g(w)$ is differentiable. This demonstrates (a). Now

$$0 = \delta\langle v, G(w + g(w)v)\rangle = \langle v, H(w + g(w)v) \cdot (\delta w + g_w(w)\delta w v)\rangle + h(\Delta w),$$

where $H(w) = G_w$ is the Hessian matrix of f and $\langle v, G(v + g(w)v)\rangle$ was expanded in a Taylor series. This demonstrates (b). Dividing through by $\|v\| \, \|w_1 - w_2\|$ demonstrates (c).

COROLLARY. *Let $\{q_n^1, q_n^2\}$ be two sequences of points in R^m, which are contained in a compact set. Let $q_n^1 - q_n^2 \to 0$ as $n \to \infty$. Then $q_n^1 + vg(q_n^1) - (q_n^2 + vg(q_n^2)) \to 0$.*

Proof. The statement follows immediately from the uniform continuity of g on a compact set.

We will now use this corollary to show that the network of Fig. 2.1 collapses in its vertical direction as it evolves, so that the whole forms a narrowing tube.

Let $q_k^1 = p_k^m$ and $q_k^2 = p_k^m + v_{k+1}^m$ (see Fig. 4.1). Since $\sum_{j=1}^{\infty} \beta_j < \infty$, the p_k^m do indeed lie in a compact set. Then the corollary allows us to conclude immediately that $v_{k+1}^{m-1} \to 0$ as $k \to \infty$. By induction, it follows similarly that $v_{k+1}^j \to 0$, $j = 1, \cdots, m$, as $k \to \infty$. In this way we obtain m sequences, all of which converge to zero. Thus for every ε, there exist $k_1, \cdots, k_m$ so that $\|v_k^i\| < \varepsilon$ for $k \geqq k_i$. Hence $\|v_k^i\| \leqq \varepsilon$ for all $k \geqq \max_{1 \leqq i \leqq m} k_i$.

5. Proof of $\|p_{n+1}^1 - p_n^1\| \to 0$ as $n \to \infty$. Let us note now that the sequence of points p_j^1 has the property:

$$f(p_{j+1}^1) \leqq f(p_j^1), \tag{5.1}$$

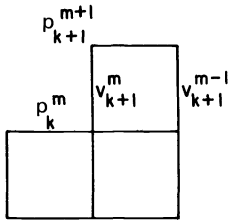


FIG. 4.1

which follows from the fact that p_{j+1}^1 is the point where the one-dimensional minimum of f in the direction v_{j+1}^1 starting from p_{j+1}^2 is obtained, and that the line $p_{j+1}^2 + \alpha v_{j+1}^1$ contains p_j^1. In particular, $f(p_j^1) \leqq f(p_1^1)$.

Let us also remark that $\{p : f(p) \leqq a\}$ is compact since $f(p) \to \infty$ as $\|p\| \to \infty$, and that the sequence p_j^1 must, therefore, lie in a compact set. We wish to show that any accumulation point p of the sequence p_j^1 has the property $f(p) \leqq f(q)$ for all points $q \in R^m$.

In this section we shall show that $\|p_{j+1}^1 - p_j^1\| \to 0$. Thus if p is an accumulation point of p_j^1 then $p_j^1, j_1 \leqq j \leqq j_1 + j_0$, can be made to stay within ε of p for any fixed j_0 and for infinitely many values of j_1. Consider now the following lemma.

LEMMA 5.1. *Let q_n, w_n be a sequence of points satisfying the following relation*:

$$a \leqq f(q_n + \alpha w_n) \leqq a + \varepsilon_n$$

with $\varepsilon_n \to 0, 0 \leqq \alpha \leqq 1$. If f is continuous, strictly convex, and $f(p) \to \infty$ as $\|p\| \to \infty$, then $\|w_n\| \to 0$.

Proof. Suppose otherwise. Then there exists a subsequence of $\|w_n\|$ with $\|w_{n_i}\| \geqq c > 0$. Since the set $S = \{q : f(q) \leqq a + \varepsilon_1\}$ is closed and bounded, it is compact. Similarly, the set $\{v : v = q_1 - q_2; q_1, q_2 \in S\}$ is also compact. Hence q_n, w_n lie in compact sets and there exists a subsequence m_i of n_i with $(q_{n_i}, w_{n_i}) \to (q, w)$. By continuity, $f(q_{n_i} + \alpha w_{m_i}) \to f(q + \alpha w)$. Clearly

$$a \leqq f(q + \alpha w) \leqq a + \varepsilon_n$$

for all n and

$$\|w\| \geqq c.$$

But then f is constant along w which contradicts the strict convexity of f.

Using this lemma, we obtain the stated conclusion, $\|p_{j+1}^1 - p_j^1\| \to 0$, by letting $q_j = p_j^1$ and $w_j = p_{j+1}^1 - p_j^1$.

6. Proof of convergence. From the result of the last two sections, we may conclude that for every $\varepsilon > 0$ and $j_0 > 0$, there exists a $j \geqq 0$ so that $\|p_{j+k}^i - p\| < \varepsilon$, $0 \leqq k \leqq j_0, 1 \leqq i \leqq m$. This statement may be viewed geometrically as a collapsing of the array of Fig. 2.1 into an ε-neighborhood of f for durations of j_0 steps at a time. Of course, p is an accumulation point of the sequence $\{p_j^n\}$. We now wish to use these facts to show that for ε sufficiently small, it follows that for infinitely many j,

$$\max_{1 \leqq k \leqq 2m-1} \left| \left\langle f_{pp}(p)w, \frac{v_{j+k}^1}{\|v_{j+k}^1\|} \right\rangle \right| > \gamma$$

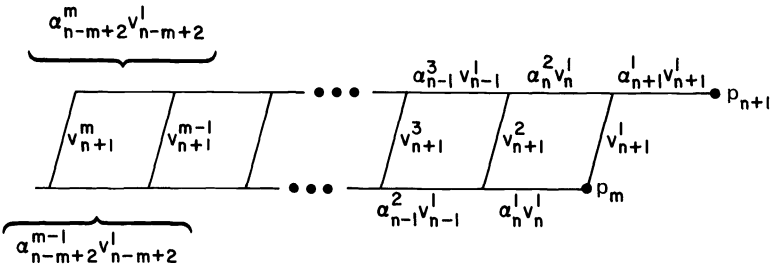


FIG. 6.1

for some $\gamma > 0$ and any w with $\|w\| = 1$. We shall accomplish this by noting that if $2m - 1$ successive search directions stay close to some subspace of R^m, then m successive v_j^m will also stay close to that subspace. This contradicts the fact that any m successive v_j^m are linearly independent and cannot collapse into a subspace.

Let $H = f_{pp}(p)$, where p is a limit point of the sequence $\{p_j^1\}$, and let $\|v\|_H^2 = \langle Hv, v\rangle$ for any symmetric H.

LEMMA 6.1. *There exists a positive γ so that for an ε sufficiently small, whenever $\|p^i_{j+k} - p\| < \varepsilon$, $1 \leqq k \leqq 2m - 1$, $1 \leqq i \leqq m$, we have*

$$\max_{1 \leqq k \leqq 2m-1} \left| \left\langle Hw, \frac{v^1_{j+k}}{\|v^1_{j+k}\|} \right\rangle \right| > \gamma$$

for any w.

Thus any $2m + 1$ successive search directions have the spanning property.

Proof. Suppose otherwise. Then for any ε, γ there exist j, w so that $\|p^i_{j+k} - p\| \leqq \varepsilon$, $1 \leqq k \leqq 2m - 1$, $1 \leqq i \leqq m$, but

$$\left| \left\langle Hw, \frac{v^1_{j+k}}{\|v^1_{j+k}\|} \right\rangle \right| \leqq \gamma, \qquad 1 \leqq k \leqq 2m - 1.$$

Let $z_i = (\alpha^{m-i}_{j+k+i} - \alpha^{m-i+1}_{j+k+i})v^1_{j+k+i}$, $i = 1, \cdots, m-1$, $z_m = v^1_{j+k+m}$. Then, by viewing Fig. 6.1, it is easily seen that

$$v^m_{j+k+m} = \sum_{i=1}^{m} z_i.$$

Since the z_i are all of the form $c \cdot v^1_{j+k}$ for some real c and some integer k, we know that

$$\left| \left\langle Hw, \frac{z_i}{\|z_i\|} \right\rangle \right| \leqq \gamma.$$

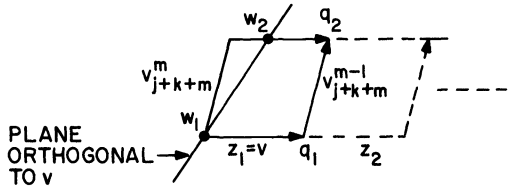


FIG. 6.2

We would like to conclude from this that

$$\left|\left\langle Hw, \frac{v^m_{j+k+m}}{\|v^m_{j+k+m}\|}\right\rangle\right|$$

is small for m successive values of k to contradict the independence of m successive v^m_{j+k+m}. To do so it is enough to show that $\|v^m_{j+k+m}\| \geqq c\|z_i\|$ for some constant c independent of j, ε. In this case,

$$\begin{aligned}\left|\left\langle Hw, \frac{v^m_{j+k+m}}{\|v^m_{j+k+m}\|}\right\rangle\right| &\leqq \left|\sum \left\langle Hw, \frac{z_i}{\|v^m_{j+k+m}\|}\right\rangle\right| \\ &\leqq \left|\sum \left\langle Hw, \frac{z_i}{c\|z_i\|}\right\rangle\right| \\ &\leqq \sum \frac{\gamma}{c} = \frac{m\gamma}{c}\end{aligned}$$

and the desired contradiction would follow.

We shall now show that v^m_{j+k+m} satisfies

$$\|v^m_{j+k+m}\|^2 \geqq c\|z_i\|^2(1 - O(\varepsilon))/\|H^{-1}\|\,\|H\|.$$

Indeed since

$$\|v^m_{j+k+m}\|^2 \geqq \|v^m_{j+k+m}\|^2_H/\|H\|$$

it suffices to show

$$\|v^m_{j+k+m}\|^2_H = \Big\|z_1 + \sum_{i=2}^{m} z_i\Big\|^2_H \geqq (1 - O(\varepsilon))\|z_i\|^2_H,$$

since $\|z_i\|^2_H \geqq \|z_i\|^2/\|H^{-1}\|$. Referring to Figs. 6.1 and 6.2, we have $\sum_{i=2}^m z_i = v^{m-1}_{j+k+m}$. Let $v = z_1$, let P be the orthogonal projection onto the plane orthogonal to v, let $w_1 = p^m_{j+k+m}$, $w_2 = w_1 + P(v^m_{j+k+m})$ and $q_1 = p^{m-2+i}_{j+k+m-2+i}$, $i = 1, 2$. Then $q_1 - q_2 = v^{m-1}_{j+k+m}$. Since v^{m-1}_{j+k+m} and v^m_{j+k+m} differ by a multiple of v, $p(v^{m-1}_{j+k+m}) = p(v^m_{j+k+m})$. Applying Lemma 4.1:

$$\begin{aligned}\langle z_1, f_{pp}(q_1)v^{m-1}_{j+k+m}\rangle &= O(w_1 - w_2)\|z_1\|\,\|w_1 - w_2\| \\ &= O(P(v^{m-1}_{j+k+m}))\|z_1\|_H\|Pv^{m-1}_{j+k+m}\|\,\|H^{-1}\| \\ &\leqq O(v^{m-1}_{j+k+m})\|z_1\|_H\|v^{m-1}_{j+k+m}\|_H.\end{aligned}$$

It follows that

$$\begin{aligned}\|v^m_{j+k+m}\|^2_H &= \left\|z_1 + \sum_{i=2}^{m} z_i\right\|^2_H = \|z_1 + v^{m-1}_{j+k+m}\|^2_H \\ &= \|z_1\|^2_H + \left\|\sum_{i=2}^{m} z_i\right\|^2_H + 2\langle z_1, Hv^{m-1}_{j+k+m}\rangle \\ &\geqq \max\left(\|z_1\|^2_H, \left\|\sum_{i=2}^{m} z_i\right\|^2_H\right) \\ &\quad -2\langle z_1, (f_{pp}(p^{m-1}_{j+k+m-1}) + O(\varepsilon))v^{m-1}_{j+k+m}\rangle\end{aligned}$$

$$\geqq \max\left(\|z_1\|_H^2, \left\|\sum_{i=2}^{m} z_i\right\|_H^2\right) - O(\varepsilon)\|z_1\|_H\left\|\sum_{i=2}^{m} z_i\right\|_H$$

$$\geqq \max\left(\|z_1\|_H^2, \left\|\sum_{i=2}^{m} z_i\right\|_H^2\right)(1 - O(\varepsilon)).$$

Continuing by induction by applying the same argument to $\sum_{i=k}^{m} z_i$, we obtain the desired result:

$$\|v_{j+k+m}^{m}\|_H^2 \geqq \max \|z_i\|_H^2(1 - O(\varepsilon)).$$

With the help of this lemma and the outline of the proof given in § 3, we obtain our final result.

THEOREM 6.1. *The sequence p_j^1 defined in § 2 converges if f is twice continuously differentiable and strictly convex; i.e., $f_{pp} \geqq 0$ and $f(x) \to \infty$ as $x \to \infty$.*

REFERENCES

[1] D. J. WILDE, *Optimum Seeking Methods*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

[2] R. FLETCHER AND M. J. D. POWELL, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163–168.

[3] M. J. D. POWELL, *An efficient method for finding the minimum of a function of several variables without calculating derivatives*, Ibid., 7 (1964), pp. 155–162.

[4] D. CHAZAN AND W. L. MIRANKER, *Chaotic relaxation*, Linear Algebra and Its Applications, 2 (1969), pp. 199–222.

[5] R. M. KARP AND R. E. MILLER, *Parallel program schemata*, J. Comput. Systems Sci., 3 (1969), pp. 147–195.

[6] W. I. ZANGWILL, *Minimizing a function without calculating derivatives*, Comput. J., 10 (1967), pp. 293–296.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

A GENERALIZATION OF THE METHOD OF BALAKRISHNAN: INEQUALITY CONSTRAINTS AND INITIAL CONDITIONS*

ARNOLD P. JONES AND GARTH P. McCORMICK†

1. Introduction. The application of penalty methods to optimal control problems has received a limited amount of attention since R. Courant's original efforts [2]. The method has been most fully developed in mathematical programming both from a theoretical and computational point of view. This area of study has been developed and exploited primarily by A. V. Fiacco and G. P. McCormick [3]. Attempts to extend these methods to optimal control and trajectory optimization problems have, in the past, been primarily aimed at removing troublesome intermediate state and/or terminal state constraints via a penalty function and then utilizing the associated necessary conditions for optimality to solve the transformed problem, still an optimal control problem. Investigations in this direction may be found in [4], [5], [7], [10] and [11].

Recently, A. V. Balakrishnan [1] has succeeded in incorporating the system differential equations into a penalty function, thereby avoiding the previously essential step of having to solve these differential equations. In this paper we consider a fixed endpoint optimal control problem with specified intermediate state constraints and control constraints. We assume these constraints are given by finite systems of inequalities. In contrast to previous works, we view the initial conditions on the state vector as well as the system differential equations as equality constraints. We construct a penalty function with appropriate terms for all of these constraints and show that, under reasonable assumptions, we generate a sequence of minimizing points of this penalty function which converge to the solution of the original control problem. The penalty function used is a realization from the class of so-called mixed interior-exterior penalty functions developed in [3] and generalized in [6]. As such, this approach is a generalization of previous works on penalty methods in optimal control theory. No computational exploitation of the method is presented.

2. Statement of problem. This paper is concerned primarily with the following problem which we will refer to as problem (A). Find admissible functions $x(\cdot)$ and $u(\cdot)$, where $x:[0, T] \to E^n$ and $u:[0, T] \to E^m$, such that $\int_0^T f^0(x(t), u(t), t)\, dt$ is minimized,

$$\frac{dx}{dt} \equiv \dot{x}(t) = f(x(t), u(t), t) \qquad \text{a.e. } [0, T],$$

$$x(0) = x_0$$

and

$$h_j(x(t), t) \geqq 0, \quad j = 1, \cdots, q, \quad t \in [0, T],$$

$$g_i(u(t), t) \geqq 0, \quad i = 1, \cdots, p, \quad \text{a.e. } [0, T].$$

* Received by the editors May 29, 1969, and in revised form November 10, 1969.

† Advanced Research Department, Research Analysis Corporation, McLean, Virginia 22101.

The vector x_0 is specified along with the functions f^0, f, g_i and h_j. The notation a.e.$[0, T]$ means almost everywhere on the interval $[0, T]$.

3. Admissible functions. We first consider the following definitions: For each $t \in [0, T]$, let

$$\Omega_x(t) = \{x(t) | h_j(x(t), t) \geqq 0, j = 1, \cdots, q\} \subset E^n,$$
$$\mathring{\Omega}_x(t) = \{x(t) | h_j(x(t), t) > 0, j = 1, \cdots, q\},$$
$$\Omega_u(t) = \{u(t) | g_i(u(t), t) \geqq 0, i = 1, \cdots, p\} \subset E^m,$$
$$\mathring{\Omega}_u(t) = \{u(t) | g_i(u(t), t) > 0, i = 1, \cdots, p\}.$$

Now let

$$X = \{x(\cdot) | x(t) \in \Omega_x(t), \text{ all } t \in [0, T]\},$$
$$\mathring{X} = \{x(\cdot) | x(t) \in \mathring{\Omega}_x(t) \text{ a.e. } [0, T], \text{ and } \int_0^T \frac{dt}{h_j(x(t), t)} < +\infty, j = 1, \cdots, q\},$$
$$U = \{u(\cdot) | u(t) \in \Omega_u(t), \text{ a.e. } [0, T]\},$$
$$\mathring{U} = \{u(\cdot) | u(t) \in \mathring{\Omega}_u(t) \text{ a.e. } [0, T], \text{ and } \int_0^T \frac{dt}{g_i(u(t), t)} < +\infty, i = 1, \cdots, p\},$$
$$Q = \{(x(\cdot), u(\cdot)) | x(t) - x_0 - \int_0^t f(x(s), u(s), s)\, ds = 0, \text{ all } t \in [0, T]\}.$$

We shall consider as admissible state functions all functions $x(\cdot)$ which are absolutely continuous with derivatives, $\dot{x}(\cdot)$, which belong to $L_2[0, T]$ and such that the functions $x(\cdot) \in X$. Bounded, measurable (Lebesgue) functions $u(\cdot)$ such that $u(\cdot) \in U$ are said to be admissible controls. We assume $\mathring{X}$, $\mathring{U}$ and Q are not empty.

4. The penalty function. Consider a decreasing null sequence of positive real numbers $\{r_k\}$. For each $r_k > 0$ consider the following definition.

DEFINITION.

$$P(r_k, x(\cdot), u(\cdot))$$
$$\equiv \frac{1}{r_k}\|x(0) - x_0\|^2 + \int_0^T \Big\{ f^0(x(t), u(t), t) + \frac{1}{r_k}\|\dot{x}(t) - f(x(t), u(t), t)\|^2 + r_k \sum_{i=1}^{p} \frac{1}{g_i(u(t), t)} + r_k \sum_{j=1}^{q} \frac{1}{h_j(x(t), t)} \Big\} dt. \tag{4.1}$$

Here $\|\cdot\|$ denotes the Euclidian norm in E^n. It is to be noted that the penalty function used in [1] is a special case of (4.1).

Essentially, the problem before us is to minimize, for each fixed k, i.e., $r_k > 0$, the functional $P(r_k, x(\cdot), u(\cdot))$ over $\mathring{X} \times \mathring{U}$. If, for each k, a minimum $(x_0^k(\cdot), u_0^k(\cdot))$ exists then we show that this sequence of minima converges, as $k \to +\infty$, i.e..

$r_k \downarrow 0$, to a pair $(x^*(\cdot), u^*(\cdot)) \in (X \times U) \cap Q$ and that

$$\lim_{k\to\infty} P(r_k, x_0^k(\cdot), u_0^k(\cdot)) = \int_0^T f^0(x^*(t)u^*(t), t)\,dt$$

$$= \min_{X\times U\cap Q} \int_0^T f^0(x(t), u(t), t)\,dt.$$

5. Assumptions. Let us make the following assumptions on the problem functions f, f^0, h_j and g_i. The function f is assumed to be measurable in t for fixed x and u, continuous in x for fixed t, u, and continuous in u for fixed t, x. In addition, we assume that there exists an integrable function M on $[0, T]$ such that

$$\|f(x, u, t)\| \leqq M(t) \qquad \text{for } 0 \leqq t \leqq T$$

for all admissible functions $x(\cdot)$, $u(\cdot)$. The function f^0 is assumed to be integrable in t for fixed x and u. The functions $h_j, j = 1, \cdots, q$, are assumed continuous in t for each fixed x and continuous in x for each fixed t. The functions $g_i, i = 1, \cdots, p$, are assumed to be continuous in t for each fixed u and continuous in u for each fixed t.

DEFINITION.

$$\Omega_x = \bigcup_{t\in[0,T]} \Omega_x(t)$$

and

$$\Omega_u = \bigcup_{t\in[0,T]} \Omega_u(t).$$

We now assume that Ω_x is bounded, as a subset of E^n, and Ω_u is compact and convex, as a subset of E^m. Further, we assume:

(A.1) $\inf_{X\times U} \int_0^T f^0(x(t), u(t), t)\,dt = \alpha > -\infty$,

(A.2) for each $\varepsilon > 0$ there exist functions $(x^\varepsilon(\cdot), u^\varepsilon(\cdot)) \in \mathring{X} \times \mathring{U} \cap Q$ such that

$$\int_0^T f^0(x^\varepsilon(t), u^\varepsilon(t), t)\,dt < \inf_{X\times U\cap Q} \int_0^T f^0(x(t), u(t), t)\,dt + \varepsilon.$$

Remark. Assumption (A.2) is an "interior" approximability assumption which is necessary for penalty function approaches to control problems. For a somewhat similar hypothesis as well as an example of where such an assumption is not valid see [7].

(A.3) If $u_n(\cdot) \in \mathring{U}$ converges weakly to $u_0(\cdot) \in U$, and $x_n(\cdot) \in X$ converges uniformly to $x_0(\cdot) \in X$, then:

$$\int_0^T \|\dot{x}_0(t) - f(x_0(t), u_0(t), t)\|^2\,dt \leqq \liminf_n \int_0^T \|\dot{x}_n(t) - f(x_n(t), u_n(t), t)\|^2\,dt \tag{5.1}$$

and

$$\int_0^T f^0(x_0(t), u_0(t), t)\,dt \leqq \liminf_n \int_0^T f^0(x_n(t), u_n(t), t)\,dt. \tag{5.2}$$

Remark. Expressions (5.1) and (5.2) with $x_n(\cdot) \equiv x_0(\cdot)$ are precisely the lower semicontinuity assumptions made in [1]. As remarked in [1], (5.1) and (5.2) are valid if the problem is linear.

6. Fundamental results. We need the following lemma for subsequent developments.

LEMMA. *For each* $r_k > 0$, $P(r_k, x(\cdot), u(\cdot))$ *is bounded below on* $\mathring{X} \times \mathring{U}$.

Proof. By the definition of $P(r_k, x(\cdot), u(\cdot))$ we have, on $\mathring{X} \times \mathring{U}$, that

$$P(r_k, x(\cdot), u(\cdot)) \geqq \inf_{\mathring{X} \times \mathring{U}} \int_0^T f^0(x(t), u(t), t)\, dt$$

$$\geqq \inf_{X \times U} \int_0^T f^0(x(t), u(t), t)\, dt = \alpha > -\infty$$

by (A.1) and the fact that $\mathring{X} \times \mathring{U} \subset X \times U$.

We now state and prove our main result.

THEOREM. *If* Ω_u *is convex and compact,* Ω_x *is bounded, and the assumptions in* § 5 *hold then*:

(a) *for each* $r_k > 0$, $P(r_k, x(\cdot), u(\cdot))$ *is minimized in* $\mathring{X} \times \mathring{U}$ *at a point* $(x_0^k(\cdot), u_0^k(\cdot))$,

(b) $\lim_{k \to \infty} P(r_k, x_0^k(\cdot), u_0^k(\cdot)) = \min_{X \times U \cap Q} \int_0^T f^0(x(t), u(t), t)\, dt$,

(c) *limit points of* $x_0^k(\cdot), u_0^k(\cdot)$ *solve* (A).

Proof. (It is to be noted that the proof of (a) is based, in part, on that given in [1].)

(a) Let $r_k > 0$ be fixed. Let B_k denote the infimum, over $\mathring{X} \times \mathring{U}$, of $P(r_k, x(\cdot), u(\cdot))$. B_k is finite, by the previous lemma. Let $\{x_n^k(\cdot), u_n^k(\cdot)\}$ be a sequence of admissible functions such that

$$B_k = \lim_{n \to \infty} \left[\frac{1}{r_k} \| x_n^k(0) - x_0 \|^2 \right.$$

$$+ \int_0^T \left\{ f^0(x_n^k(t), u_n^k(t), t)\, dt + \frac{1}{r_k} \| \dot{x}_n^k(t) - f(x_n^k(t), u_n^k(t), t) \|^2 \right. \tag{6.1}$$

$$\left. \left. + r_k \sum_{i=1}^{p} \frac{1}{g_i(u_n^k(t), t)} + r_k \sum_{j=1}^{q} \frac{1}{h_j(x_n^k(t), t)} \right\} dt \right].$$

We now note that the range of $x_n^k(\cdot)$ is bounded, i.e., Ω_x is bounded as a subset of E^n. From this it follows that $x_n^k(\cdot)$ is a set of uniformly bounded continuous functions. We shall now show that they are also equicontinuous. To this end we must first show that $\left\{ \int_0^T \| \dot{x}_n^k(t) \|^2 \right\} dt$ is uniformly bounded. Consider $\int_0^T \| \dot{x}_n^k(t) - f(x_n^k(t), u_n^k(t), t) \|^2\, dt \geqq 0$. Since $B_k \geqq \alpha > -\infty$ we have that if $\int_0^T \| \dot{x}_n^k(t) - f(x_n^k(t), u_n^k(t), t) \|^2\, dt \to +\infty$ then necessarily $\int_0^T f^0(x_n^k(t), u_n^k(t), t)\, dt \to -\infty$. This is a contradiction of (A.1). Hence $\int_0^T \| \dot{x}_n^k(t) - f(x_n^k(t), u_n^k(t), t) \|^2\, dt = A$, where $A < +\infty$, and we have

the following:

$$0 \leqq \int_0^T \|\dot{x}_n^k(t) - f(x_n^k(t), u_n^k(t), t)\|^2 \, dt$$

$$= \int_0^T \|\dot{x}_n^k(t)\|^2 + \int_0^T \|f(x_n^k(t), u_n^k(t), t)\|^2 \, dt - 2\int_0^T \langle \dot{x}_n^k(t), f(x_n^k(t), u_n^k(t), t)\rangle \, dt = A,$$

where $\langle \cdot, \cdot \rangle$ is the inner product of vectors in E^n. Remembering that $\|f(x, u, t)\| \leqq M(t)$ we see

$$\int_0^T \|\dot{x}_n^k(t)\|^2 \, dt \leqq 2\int_0^T \|\dot{x}_n^k(t)\| M(t) \, dt + \int_0^T M^2(t) \, dt + A.$$

By the Cauchy-Schwarz inequality we now have

$$\int_0^T \|\dot{x}_n^k(t)\|^2 \, dt - 2\left(\int_0^T \|\dot{x}_n^k(t)\|^2 \, dt\right)^{\frac{1}{2}}\left(\int_0^T M^2(t) \, dt\right)^{\frac{1}{2}} - \int_0^T M^2(t) \, dt \leqq A$$

or

$$\left[\left(\int_0^T \|\dot{x}_n^k(t)\|^2 \, dt\right)^{\frac{1}{2}} - \left(\int_0^T M^2(t) \, dt\right)^{\frac{1}{2}}\right]^2 \leqq A + 2\int_0^T M^2(t) \, dt$$

which implies

$$\int_0^T \|\dot{x}_n^k(t)\|^2 \, dt \leqq \left[\left(A + 2\int_0^T M^2(t) \, dt\right)^{\frac{1}{2}} + \left(\int_0^T M^2(t) \, dt\right)^{\frac{1}{2}}\right]^2.$$

Hence we see that $\left\{\int_0^T \|\dot{x}_n^k(t)\|^2 \, dt\right\}$ is uniformly bounded. So that again by the Cauchy-Schwarz inequality and the monotonicity of the integral, as a set function, we have

$$\|x_n^k(t_1) - x_n^k(t_2)\|^2 = \left\|\int_{t_1}^{t_2} \dot{x}_n^k(t) \, dt\right\|^2 \leqq |t_2 - t_1| \int_0^T \|\dot{x}_n^k(t)\|^2 \, dt$$

and we see that the $x_n^k(\cdot)$'s are equicontinuous. By Arzela's theorem there exists a subsequence, again denoted by $\{x_n^k(\cdot)\}$, which converges uniformly on $[0, T]$ to a continuous function $x_0^k(\cdot)$. Again, from the uniform boundedness of $\int_0^T \|\dot{x}_n^k(t)\|^2 \, dt$ we can extract a further subsequence from $\{\dot{x}_n^k(\cdot)\}$ such that $\dot{x}_n^k(\cdot) \to y(\cdot)$, where the convergence is weak sequential convergence in $L_2[0, T]$. Hence $y(\cdot) \in L_2[0, T]$ and also $x_0^k(t) - x_0^k(0) = \int_0^t \dot{x}_0^k(s) \, ds = \lim_n \int_0^t x_n^k(s) \, ds = \int_0^t y(s) \, ds$. Therefore $\dot{x}_0^k(\cdot) = y(\cdot)$ for almost all $t \in [0, T]$. Hence, $\dot{x}_0^k(\cdot) \in L_2[0, T]$. Since Ω_u is compact and convex in E^m and since $[0, T]$ is compact in E^1 there exists a subsequence, relabeled $\{u_n^k(\cdot)\}$, which has a weak sequential limit function $u_0^k(\cdot)$, and which is bounded measurable (see, e.g., [8]). Since $\{x_n^k(\cdot), u_n^k(\cdot)\}$ is a minimizing sequence for $P(r_k, x(\cdot), u(\cdot))$ we have as a consequence that the sequence (in n), for $i = 1, \cdots, p$ and $j = 1, \cdots, q$,

$$\int_0^T \frac{1}{h_j(x_n^k(t), t)} dt \quad \text{and} \quad \int_0^T \frac{1}{g_i(u_n^k(t), t)} dt,$$

are bounded and

$$\frac{1}{h_j(x_n^k(t), t)} \to \frac{1}{h_j(x_0^k(t), t)}$$

and

$$\frac{1}{g_i(u_n^k(t), t)} \to \frac{1}{g_i(u_0^k(t), t)} \qquad \text{a.e. } [0, T].$$

Then by Fatou's lemma (see, e.g., [9]) we have

$$\int_0^T \frac{1}{g_i(u_0^k(t), t)}\, dt \leqq \liminf \int_0^T \frac{1}{g_i(u_n^k(t), t)}\, dt \tag{6.2}$$

and

$$\int_0^T \frac{1}{h_i(x_0^k(t), t)}\, dt \leqq \liminf \int_0^T \frac{1}{h_j(x_n^k(t), t)}\, dt. \tag{6.3}$$

We note also that, since $x_n^k(\cdot) \to x_0^k(\cdot)$ uniformly on $[0, T]$, we have

$$\lim_n \|x_n^k(0) - x_0\|^2 = \|x_0^k(0) - x_0\|^2. \tag{6.4}$$

As a consequence of (6.2) and (6.3) we see that $(x_0^k(\cdot), u_0^k(\cdot)) \in \mathring{X} \times \mathring{U}$. Now by recalling (5.1) and (5.2) and (6.2), (6.3) and (6.4) we have

$$\begin{aligned} \frac{1}{r_k}\|x_0^k(0) - x_0\|^2 + \int_0^T \Bigg\{ & f^0(x_0^k(t), u_0^k(t), t) + \frac{1}{r_k}\|\dot{x}_0^k(t) - f(x_0^k(t), u_0^k(t), t)\|^2 \\ & + r_k \sum_{i=1}^{p} \frac{1}{g_i(u_0^k(t), t)} + r_k \sum_{j=1}^{q} \frac{1}{h_j(x_0^k(t), t)} \Bigg\} dt. \end{aligned} \tag{6.5}$$

Therefore

$$P(r_k, x_0^k(\cdot), u_0^k(\cdot)) = B_k,$$

i.e., the infimum of $P(r_k, x(\cdot), u(\cdot))$ is actually achieved on $\mathring{X} \times \mathring{U}$ and (a) is proved.

(b) and (c). We know from (4.1) that

$$\int_0^T f^0(x_0^k(t), u_0^k(t), t)\, dt \leqq P(r_k, x_0^k(\cdot), u_0^k(\cdot)) \leqq P(r_k, x(\cdot), u(\cdot)) \tag{6.6}$$

so that for pairs $x(\cdot), u(\cdot)$ in $\mathring{X} \times \mathring{U} \cap Q$ we have

$$\int_0^T f^0(x_0^k(t), u_0^k(t), t)\, dt \leqq P(r_k, x(\cdot), u(\cdot)).$$

By (A.2) we have, for every $\varepsilon > 0$, a pair $\hat{x}(\cdot), \hat{u}(\cdot)$ in $\mathring{X} \times \mathring{U} \cap Q$ such that

$$\begin{aligned} P(r_k, x_0^k(\cdot), u_0^k(\cdot)) & \leqq P(r_k, \hat{x}(\cdot), \hat{u}(\cdot)) \\ & \leqq \inf_{X \times U \cap Q} \int_0^T f^0(x(t), u(t), t)\, dt + \varepsilon/2 \\ & \quad + r_k \int_0^T \left(\sum_{i=1}^{p} \frac{1}{g_i(\hat{u}(t), t)} + \sum_{j=1}^{q} \frac{1}{h_j(\hat{x}(t), t)} \right) dt. \end{aligned} \tag{6.7}$$

Now we can choose k large enough so that

$$r_k \int_0^T \left(\sum_{j=1}^{p} \frac{1}{g_j(\hat{u}(t), t)} + \sum_{j=1}^{q} \frac{1}{h_j(\hat{x}(t), t)} \right) dt < \varepsilon/2 \tag{6.8}$$

since $r_k \downarrow 0$ as $k \to \infty$. From (6.6), (6.7) and (6.8) we get

$$\int_0^T f^0(x_0^k(t), u_0^k(t), t)\, dt \leqq \inf_{X \times U \cap Q} \int_0^T f^0(x(t), u(t), t)\, dt + \varepsilon \tag{6.9}$$

and we see that $\int_0^T f^0(x_0^k(t), u_0^k(t), t)\, dt$ is bounded above and below. As a consequence of this we have that

$$\|x_0^k(0) - x_0\|^2 \to 0 \tag{6.10}$$

and

$$\int_0^T \|\dot{x}_0^k(t) - f(x_0^k(t), u_0^k(t), t)\|^2\, dt \to 0 \qquad \text{as } k \to \infty. \tag{6.11}$$

Now we observe that for each k, $u_0^k(\cdot) \in \mathring{U}$ and $x_0^k(\cdot) \in \mathring{X}$ so by a repetition of the same argument used in the proof of (a) we infer the existence of a weak limit $u^*(\cdot)$ of a subsequence of $\{u_0^k(\cdot)\}$ and a limit $x^*(\cdot)$ of $\{x_0^k(\cdot)\}$ which has a derivative $\dot{x}^*(\cdot)$ belonging to $L_2[0, T]$. We observe that $u^*(\cdot) \in U$ and $x^*(\cdot) \in X$. To prove that $(x^*(\cdot), u^*(\cdot)) \in Q$ we merely observe, using (5.1), that we have

$$\int_0^T \|\dot{x}^*(t) - f(x^*(t), u^*(t), t)\|^2\, dt \leqq \liminf_k \int_0^T \|\dot{x}_0^k(t) - f(x_0^k(t), u_0^k(t), t)\|^2\, dt$$

and by (6.11) the right-hand side is zero. Therefore we have that $x^*(t) = f(x^*(t), u^*(t), t)$ a.e. $[0, T]$ and by (6.10) we have

$$\|x^*(0) - x_0\| \leqq \|x^*(0) - x_0^k(0)\| + \|x_0^k(0) - x_0\|.$$

Hence $x^*(0) = x_0$ and

$$\lim_{k \to \infty} P(r_k, x_0^k(\cdot), u_0^k(\cdot)) = \int_0^T f^0(x^*(t), u^*(t), t)\, dt = \inf_{X \times U \cap Q} \int_0^T f^0(x, u, t)\, dt,$$

for by (5.2) we have

$$\int_0^T f^0(x^*(t), u^*(t), t)\, dt \leqq \liminf_k \int_0^T f^0(x_0^k(t), u_0^k(t), t)\, dt$$

and by (6.9) we have, for $\varepsilon > 0$,

$$\int_0^T f^0(x_0^k(t), u_0^k(t), t)\, dt \leqq \inf_{X \times U \cap Q} \int_0^T f^0(x, u, t)\, dt + \varepsilon;$$

hence

$$\int_0^T f^0(x^*(t), u^*(t), t)\, dt \leqq \inf_{X \times U \cap Q} \int_0^T f^0(x, u, t)\, dt + \varepsilon$$

and we obtain

$$\int_0^T f^0(x^*(t), u^*(t), t)\, dt \leqq \inf_{X \times U \cap Q} \int_0^T f^0(x, u, t),$$

which yields the desired equality; this proves (c).

7. Conclusions. The purpose of this paper was to extend, to certain optimal control problems, methods which have enjoyed success in nonlinear programming. To this end, a convergence theorem has been proved which allows for the incorporation of explicit control vector inequality constraints, explicit state vector constraints as well as the system differential equations and initial conditions. This represents the first work, to the authors' knowledge, in which a state and control constrained optimal control problem is replaced by a sequence of truly unconstrained optimization problems. As of now, no effort has been made to implement these results from a computational point of view.

Acknowledgment. The authors gratefully acknowledge many discussions with Dr. J. E. Falk and Dr. A. V. Fiacco. In particular, we would like to acknowledge Dr. Fiacco's constant encouragement of this research.

REFERENCES

[1] A. V. BALAKRISHNAN, *On a new computing method in optimal control*, this Journal, 6 (1968), pp. 149–173.

[2] R. COURANT, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49 (1943), pp. 1–23.

[3] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[4] J. CULLUM, *Penalty functions and nonconvex continuous optimal control problems*, Rep. RC 2154, IBM, Yorktown Heights, New York, 1968.

[5] K. OKAMURA, *Some mathematical theory of the penalty method for solving optimum control problems*, this Journal, 2 (1965), pp. 317–331.

[6] A. V. FIACCO AND A. P. JONES, *Generalized penalty methods in topological spaces*, SIAM J. Appl. Math., 5 (1969), pp. 996–1000.

[7] D. L. RUSSELL, *Penalty functions and bounded phase coordinate control*, this Journal, 2 (1965), pp. 409–422.

[8] E. B. LEE AND L. MAKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[9] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis*, Frederick Ungar, New York, 1955.

[10] R. E. KOPP AND H. G. MOYER, *Trajectory optimization techniques*, Advances in Control Systems, vol. 4, C. T. Leondes, ed., Academic Press, New York, 1966.

[11] H. J. KELLEY, *Method of gradients*, Optimization Techniques with Applications to Aerospace Systems, G. Leitmann, ed., Academic Press, New York, 1962.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

MEASURABILITY PROPERTIES OF SET-VALUED MAPPINGS IN A BANACH SPACE*

RICHARD DATKO†

1. Introduction. Let X and Y be two point sets. A mapping P is called a set-valued mapping from X into Y if to every x in X the map P associates a subset of Y. In [1] Aumann made a measure theoretic study of set-valued mappings whose domain is the real line and whose range is Euclidean n-space. He introduced notions of measurability and integrability for such mappings and established important properties for them such as the Lebesgue dominated convergence theorem. Hermes [2] extended the work of Aumann. In addition he obtained necessary and sufficient conditions for a general class of mappings to be the integrals of uniformly bounded set-valued mappings.

Aumann's work was extended in another direction by Debreu [3] who replaced the assumption that the graphs of the mappings be analytic sets by other criteria which are in practice sometimes easier to handle. He also extended the study of measurable set-valued mappings to mappings with values in the nonempty compact convex subsets of a real Banach space. Using an embedding theorem of Rådström [4] Debreu embedded the mappings he studied as the points of a convex cone in a real normed linear space. In so doing he was able to preserve the Hausdorff set distance as an isometry, set addition as vector addition and the operation of set multiplication by positive scalars as multiplication of vectors by positive scalars. In this way he reduced the theory of integration for set-valued mappings with values in the nonempty compact convex sets of a real Banach space to the theory of integration for vector-valued functions in a Banach space.

The purpose of this paper is to consider Banach space analogues of some properties of set-valued mappings which were obtained by Aumann [1] and Hermes [2]. Here we treat the case where the range space is a real separable reflexive Banach space and the domain is a locally compact Polish space furnished with a positive nonatomic regular Borel measure. The main results are Theorem 1 which is a Blaschke-type selection theorem for measurable integrably bounded set-valued mappings, and Theorem 2 which gives necessary and sufficient conditions for a class of set-valued mappings to be representable as the indefinite integrals of set-valued mappings. One interesting by-product of the techniques developed in proving these theorems is that for the class of mappings under consideration their integration properties are largely dependent on those of their closed convex hulls; and the integration properties of the latter can be studied by examining their support functionals. This reduces the problem to a consideration of scalar functions which are amenable to the usual techniques of measure theory.

The most important new mathematical results utilized in this paper are those obtained by K. Kuratowski and C. Ryll-Nardzewski in [19] and C. Castaing in [5]. More accessible sources of Castaing's work are references [6]–[10]. However, all reference will be made to [5] since complete proofs appear there. Reference [10]

* Received by the editors March 18, 1969, and in revised form October 11, 1969.

† Department of Mathematics, Georgetown University, Washington, D.C. 20007.

is also a good source. A general reference for functional analytic concepts is [11], for properties of set-valued mappings [12] and for measure theoretic properties [13]. For completeness an Appendix is added which contains statements of results from [5] which are used in this paper and also the statement of a result from [14].

2. Notations and conventions. $\varnothing$ will denote the empty set.

X will denote a separable reflexive Banach space over the real numbers and X' its topological dual. The norm in X and X' will be denoted by $\|\cdot\|$ and, as usual, if $x' \in X'$ then $\|x'\| = \sup\{|x'(x)| : x \in X, \|x\| = 1\}$. S' will stand for the surface of the unit ball in X'.

The symbol 2^X is the family of all subsets of X, $\mathscr{K}(X)$ is the family of closed bounded subsets of X and $\overline{\text{co}}\,\mathscr{K}(X)$ the family of closed bounded convex subsets of X.

If $P \subset X$ then co P is the convex hull of P and $\overline{\text{co}}\,P$ the closed convex hull of P. $\bar{P}$ or cl (P) will denote the closure of P in the normed topology.

T will always be a locally compact Polish space (see, e.g., [15, p. 121] for the definition of a Polish space). It will be assumed that there exists a nonatomic positive regular Borel measure μ on T and that $\mu(T) < +\infty$.

$L^p(T, R)$, $1 \leqq p < \infty$, will denote the equivalence class of μ-measurable functions from T into the real numbers whose pth powers are μ-integrable. $L^p(T, X)$ will denote the collection of equivalence classes of μ-measurable mappings from T into X whose norms are in $L^p(T, R)$.

Let $x' \in X'$ and $K \subset X$. Then

$$x'(K) = \sup\{x'(x) : x \in K\}.$$

Let $\{q_n\}$ be a sequence of elements in X which converges weakly to an element q in X. This convergence will be denoted by either $\omega\text{-lim}\, q_n = q$ or $q_n \overset{\omega}{\to} q$.

3. Statement of definitions and basic lemmas. The following definition can be found in Castaing [5].

DEFINITION 1. Let $P: T \to 2^X$ be a set-valued mapping. If for each closed subset $A \subset X$ the set

$$P^-(A) = \{t \in T : P(t) \cap A \neq \varnothing\}$$

is measurable, then P is said to be a *measurable set-valued mapping.*

DEFINITION 2. A set-valued mapping $P: T \to 2^X$ is said to be *p-power integrably bounded* if there exists a $g \in L^p(T, R)$ such that

$$\sup\{\|x\| : x \in P(t)\} \leqq g(t) \quad \text{a.e. on } T.$$

Notice that if $\mu(T) < +\infty$, then $g \in L^1(T, R)$ if $g \in L^p(T, R)$, $1 < p < \infty$.

DEFINITION 3. Let P be a mapping from $T \to 2^X$. A measurable mapping $\sigma: T \to X$ is called a *measurable cross section* of P if $\sigma(t) \in P(t)$ a.e. on T.

DEFINITION 4. Let $P: T \to 2^X$ and let $\mathscr{F}$ denote the family of measurable cross sections of P. Then for any measurable set $A \subset T$ we define

$$\int_A P(t)\, d\mu(t) = \left\{\int_A \sigma(t)\, d\mu(t) : \sigma \in \mathscr{F}\right\}.$$

Remark 1. If $P: T \to 2^X$ is measurable and integrably bounded with values a.e. in the nonempty closed subsets of X, then $\int_A P(t)\,d\mu(t) \neq \varnothing$ for any measurable $A \subset T$. This is a consequence of Theorem 5.2 in [5] and the Lebesgue bounded convergence theorem.

Also notice that since X is a separable reflexive Banach space the same is true for its topological dual X'. Hence the surface of the unit ball in X' will contain a countably dense set of points $\nu = \{x_i'\}$.

DEFINITION 5. Let $\nu = \{x_i'\}$ be any countably dense subset of S'. For any two nonempty sets Q and R in $\overline{\text{co}}\,\mathscr{K}(X)$ we define the *distance function*

$$d_\nu(Q, R) = \sum_{i=1}^{\infty} \frac{1}{2^i} \frac{|x_i'(Q) - x_i'(R)|}{1 + |x_i'(Q) - x_i'(R)|}, \qquad x_i' \in \nu, \quad i = 1, 2, \cdots.$$

DEFINITION 6. The *graph* $\mathscr{G}$ of a set-valued mapping $P: T \to 2^X$ is the set $\mathscr{G} = \{(t, u) \in T \times X : u \in P(t)\}$. $\mathscr{G}$ is said to be *closed* if it is closed in the topological product $T \times X$.

4. Some properties of convex sets and measurable set-valued mappings.

LEMMA 1. *Let γ be any countably dense subset of S' and let K be a nonempty closed bounded convex set in X. To each $x_i' \in \gamma$ associate the closed half-space*

$$H_i = \{x \in X : x_i'(x) \leqq x_i'(K)\}.$$

Then $K = \bigcap_{i=1}^{\infty} H_i$. Hence any two closed bounded convex sets K_1 and K_2 are equal if and only if for all $x_i' \in \gamma$, $x_i'(K_1) = x_i'(K_2)$.

Proof. Each closed half-space H_i contains K. Hence $K \subset \bigcap_{i=1}^{\infty} H_i$. Suppose there exists $p \in \bigcap_{i=1}^{\infty} H_i - K$. Then there is an $x' \in S'$ and constants c and $\varepsilon > 0$ such that

$$c = x'(p) \geqq c - \varepsilon \geqq x'(K)$$

(see, e.g., [11, p. 417]). Since K is bounded, there exists a positive constant $M < +\infty$ such that $\|p\| \leqq M$ and $\|k\| \leqq M$ for all $k \in K$. Choose $x_i' \in \gamma$ such that $\|x_i' - x'\| < \varepsilon/(4M)$. Then the inequalities

$$|x_i'(k) - x'(k)| \leqq \|k\|\,\|x_i' - x'\| < \varepsilon/4, \qquad k \in K,$$

and

$$|x_i'(p) - x'(p)| < \varepsilon/4$$

hold.

Thus, for all $k \in K$,

$$x_i'(k) < x'(k) + \frac{\varepsilon}{4} \leqq x'(p) - \frac{3\varepsilon}{4} < x_i'(p) + \frac{\varepsilon}{4} - \frac{3\varepsilon}{4} = x_i'(p) - \frac{\varepsilon}{2}.$$

This shows that x_i' also strictly separates p and K. But this is impossible since $x_j'(K) \geqq x_j'(p)$ for all $x_j' \in \gamma$. This proves the lemma.

COROLLARY. *The distance function d_ν of Definition 5 makes the nonempty sets in $\overline{\text{co}}\,\mathscr{K}(X)$ a metric space.*

Proof. If P, Q and R are nonempty sets in $\overline{\text{co}}\ \mathcal{K}(X)$, then it is immediate from Definition 5 that:

(a) $d_v(R, Q) = d_v(Q, R)$;

(b) $d_v(R, Q) \leqq d_v(R, P) + d_v(P, Q)$;

(c) $0 \leqq d_v(P, Q) < +\infty$.

By Lemma 1, $R = Q$ if and only if $x_i'(R) = x_i'(Q)$ for all $x_i' \in v$. Hence

(d) $d_v(R, Q) = 0$ if and only if $R = Q$.

Properties (a)–(d) establish that d_v is a metric on the nonempty subsets of $\overline{\text{co}}\ \mathcal{K}(X)$.

Remark 2. Observe that a *uniformly* bounded sequence $\{K_n\}$ of nonempty sets from $\overline{\text{co}}\ \mathcal{K}(X)$ converges in terms of the metric d_v to a nonempty set K of $\overline{\text{co}}\ \mathcal{K}(X)$ if and only if $\lim_{n\to\infty} x_i'(K_n) = x_i'(K)$ for all $x_i' \in v$.

LEMMA 2. *Let $\{K_n\}$ be a sequence of nonempty closed convex subsets of X which are uniformly bounded. Let γ be any countably dense subset of S'. For each $x_i' \in \gamma$ define the numbers*

$$h_i^n = x_i'(K_n), \qquad n = 1, 2, \cdots,$$

and suppose that for each i, $\lim_{n\to\infty} h_i^n = h_i$ exists. Define the closed half-space $H_i = \{x : x_i'(x) \leqq h_i\}$ and the set $K = \bigcap_{i=1}^{\infty} H_i$. Then K is a nonempty bounded closed convex set in X and is hence in $\overline{\text{co}}\ \mathcal{K}(X)$.

Proof. To prove the lemma it is only necessary to show that $K \neq \varnothing$. For each integer n select an element $k_n \in K_n$. Since X is a reflexive Banach space and the sequence $\{k_n\}$ is uniformly bounded there exists a subsequence $\{k_q\} \subset \{k_n\}$ and an element k in X such that $k_q \xrightarrow{\omega} k$. But this implies that for each $x_i' \in v$,

$$x_i'(k) = \lim_{q\to\infty} x_i'(k_q) \leqq \lim_{q\to\infty} x_i'(K_q) = h_i.$$

This shows that $k \in \bigcap_{i=1}^{\infty} H_i$. Hence K is nonempty.

COROLLARY 1. *Let the hypotheses of Lemma* 2 *be satisfied. Then in terms of the metric d_v of Definition* 5 $\lim_{n\to\infty} d_v(K_n, K) = 0$.

Proof. Let $x_i' \in v$. Since the sets $\{K_n\}$ are uniformly bounded nonempty closed subsets of a reflexive Banach space there exists for each integer n an element $k_n \in K_n$ such that $x_i'(k_n) = h_i^n$. Again making use of the reflexivity of X we can find a weakly convergent subsequence $\{k_q\} \subset \{k_n\}$ with limit point k in K. But then $x_i'(k) = \lim_{q\to\infty} x_i'(k_q) = \lim_{q\to\infty} x_i'(K_q) = \lim_{n\to\infty} x_i'(K_n) = \lim_{n\to\infty} h_i^n = h_i = x_i'(K)$. Since this is true for all $x_i' \in v$ it follows from Remark 2 above that $\lim_{n\to\infty} d_v(K_n, K) = 0$.

COROLLARY 2 (Blaschke-type selection theorem). *Let $\{K_n\}$ be a sequence of closed nonempty convex sets in X which are uniformly bounded. Then there exists a subsequence $\{K_q\} \subset \{K_n\}$ which converges in the metric d_v to a nonempty closed bounded convex set K.*

Proof. For each $x_i' \in \gamma$ define the numbers $h_i^n = x_i'(K_n)$. Since the $\{h_i^n\}$ are uniformly bounded for all indices (i, n) the Cantor diagonalization process can be used to find a subsequence $\{q\} \subset \{n\}$ and a sequence $\{h_i\}$ such that for each $x_i' \in \gamma$, $\lim_{q\to\infty} h_i^q = h_i$. We then apply Lemma 2 and Corollary 1 above to the sets $\{K_q\}$ and the numbers $\{h_i^q\}$ to obtain a nonempty convex set K which satisfies the conclusion of this corollary.

Remark 3. The conclusion of Corollary 2 is not new. It is essentially a special case of Proposition 3.5 and Theorem 4.2 of [16]. To see this observe that if B is a

closed ball containing the sets $\{K_n\}$ and K then the topological space $(B, \sigma(X, X')|B)$ ($\sigma(X, X')$ denotes the weak topology on X) is metrizable (see, e.g., [11, p. 434]). The metric on this space gives rise to a Hausdorff metric (see, e.g., [17]) on the set $\overline{\text{co}}\,\mathscr{K}(X \cap B)$ (i.e., the nonempty closed convex subsets of X which are contained in B). It can be shown that the Hausdorff metric on $\overline{\text{co}}\,\mathscr{K}(X \cap B)$ makes $\overline{\text{co}}\,\mathscr{K}(X \cap B)$ into a compact metric space and determines a stronger topology on it than that induced by the metric d_v of Definition 5. Since $\overline{\text{co}}\,\mathscr{K}(X \cap B)$ is compact in the Hausdorff metric the same must be true of $\overline{\text{co}}\,\mathscr{K}(X \cap B)$ with the metric d_v.

The reason for using d_v in this paper is that because of its analytic representation it is easier to handle than the Hausdorff metric. Moreover it is defined on the entire space $\overline{\text{co}}\,\mathscr{K}(X)$ whereas the Hausdorff metric is not, unless X is finite-dimensional.

LEMMA 3. *Let $p: X' \to R$ be a continuous positively homogeneous sublinear functional. For each $x' \in S'$ define the half-space in X:*

$$H_{x'} = \{x : x'(x) \leqq p(x')\}.$$

Then $\bigcap_{x' \in S'} H_{x'} = K$ is a nonempty closed bounded convex set in X.

Proof. K is closed and convex since each of the sets $H_{x'}$ are. The boundedness of K is a consequence of the fact that p is positively homogeneous, continuous and subadditive. Thus it is only necessary to prove that $K \neq \varnothing$.

Choose $x_0' \in S'$ and consider the one-dimensional subspace $M' = \{\beta x_0' : \beta \in R\}$. On M' define the continuous linear functional x_0'' which is given by $x_0''(\beta x_0') = \beta p(x_0')$. Since p is sublinear and positively homogeneous $p(\beta x_0' - \beta x_0') = 0 \leqq p(\beta x_0') + p(-\beta x_0')$, which shows that

$$x_0''(\beta x_0') = \beta p(x_0') \leqq p(\beta x_0') \quad \text{if } \beta < 0$$

and

$$x_0''(\beta x_0') = \beta p(x_0') = x_0''(\beta x_0') \quad \text{if } \beta \geqq 0.$$

Thus p majorizes x_0'' on M' and by the Hahn-Banach theorem x_0'' can be extended to a continuous linear functional X_0'' on X' such that $X_0''(x') \leqq p(x')$ for all $x' \in X'$ and $X_0''(x_0') = p(x_0')$. However, since X is reflexive, this means there exists an $x_0 \in X$ such that $X_0''(x') = x'(x_0)$ for all $x' \in X'$. Thus $x'(x_0) \leqq p(x')$ for all $x' \in S'$ and $x_0'(x_0) = p(x_0)$ which proves the intersection is nonempty.

COROLLARY. *Let the hypothesis of Lemma 3 hold; then a by-product of its proof is that for each $x' \in X'$ there is a $k \in K$ such that $x'(k) = p(x')$. Hence $x'(K) = p(x')$ for all $x' \in S'$.*

LEMMA 4. *If K is a nonempty closed bounded convex subset of X, then there exists a continuous positively homogeneous sublinear functional $p: X' \to R$ such that*

$$K = \bigcap_{x' \in S'} \{x : x'(x) \leqq p(x')\}.$$

Proof. Define $p: S' \to R$ as follows:

$$p(x') = x'(K).$$

It is immediate that p satisfies all the conclusions of the lemma.

LEMMA 5. *Suppose* $P: T \to 2^X$ *is measurable. Let* $x' \in S'$ *be fixed and define*

$$h(t) = x'(P(t)),$$
$$H(t) = \{x : x'(x) \leqq h(t)\},$$
$$\hat{H}(t) = \{x : x'(x) \geqq h(t)\}.$$

Then the function h *and the set-valued mappings* H, $\hat{H}$ *and* $H \cap \hat{H}$ *are measurable.*

Proof. For any $r \in R$ define

$$H_r = \{x : x'(x) \leqq r\} \quad \text{and} \quad \hat{H}_r = \{x : x'(x) \geqq r\}.$$

Then since H_r and $\hat{H}_r$ are closed subsets of X and P is by assumption measurable, it follows that

$$\{t : P(t) \cap H_r \neq \varnothing\}, \qquad \{t : P(t) \cap \hat{H}_r \neq \varnothing\},$$
$$\{t : P(t) \cap H_r = \varnothing\}, \qquad \{t : P(t) \cap \hat{H}_r = \varnothing\}$$

are all measurable sets. Hence the set

$$\begin{aligned} E(q) &= \{t : h(t) = q\} \\ &= \{t : P(t) \cap H_q \neq \varnothing\} \cap \left[\bigcap_{n=1}^{\infty} \{t : P(t)\hat{H}_{q+1/n} = \varnothing\}\right] \\ &\qquad \cap \left[\bigcap_{n=1}^{\infty} \{t : P(t)\hat{H}_{q-1/n} \neq \varnothing\}\right] \end{aligned}$$

is measurable and thus the set

$$\{t : q > h(t)\} = \{t : P(t) \cap H_q \neq \varnothing\} \cap \{t : P(t) \cap \hat{H}_q = \varnothing\} - E(q)$$

is measurable. This proves that h is a measurable function.

To see that H is measurable we proceed as follows.

The function h is measurable and μ is a positive regular Borel measure. Hence given any $\varepsilon > 0$ there exists a closed set $A \subset T$ such that $h|A$ is continuous and $\mu(T - A) < \varepsilon$ (this is the Lusin property of regular Borel measures). We shall show that $H|A$ has a closed graph and this is accordingly Souslin (see, e.g., [15, p. 125]).

Consider $t_0 \in A$ and $x_0 \notin H(t_0)$. Then by the definition of $H(t_0)$, $x'(x_0) - h(t_0) = \varepsilon_0 > 0$. Let $V(x_0) = \{x : \|x - x_0\| < \varepsilon_0/2\}$ and choose $U(t_0)$ to be any open set in the relative topology of A such that $|h(t_0) - h(t)| < \varepsilon_0/2$ if $t \in U(t_0)$. We claim that if $t \in U(t_0)$ then $V(x_0) \cap H(t) = \varnothing$ and hence $H|A$ has a closed graph.

To see this suppose the contrary. Then there exists $\bar{x} \in V(x_0)$ such that $\bar{x} \in H(t)$. This implies that $x'(\bar{x}) \leqq h(t) \leqq x'(x_0) - \varepsilon_0/2$. However,

$$\begin{aligned} \varepsilon_0 &= x'(x_0) - h(t_0) \leqq x'(x_0) - h(t) + |h(t_0) - h(t)| \\ &< x'(x_0) - x'(\bar{x}) + \frac{\varepsilon_0}{2} \leqq \|x'\| \, \|x_0 - \bar{x}\| + \frac{\varepsilon_0}{2} < \frac{\varepsilon_0}{2} + \frac{\varepsilon_0}{2} = \varepsilon_0. \end{aligned}$$

The above contradiction shows that $H|A$ has a closed graph.

Application of Theorem 1(v) of [18] then shows that H is measurable.

It can be shown in a similar manner that $\hat{H}$ and $H \cap \hat{H}$ are measurable.

LEMMA 6. *Let γ be any countably dense subset of S'. Let $P: T \to 2^X$ be a measurable set-valued mapping. For each $x_n' \in \gamma$ define $h_n(t) = x_n'(P(t))$ and $H_n(t) = \{x : x_n'(x) \leqq h_n(t)\}$. Then the mapping $Q = \bigcap_{n=1}^{\infty} H_n$ is a measurable mapping from T into the closed convex subsets of X.*

Proof. Let $\varepsilon > 0$ be given. Applying the same argument used in proving Lemma 5 we can find for each n a closed set $A_n \subset T$ such that $\mu(T - A_n) < \varepsilon/2^n$ and $H_n | A_n$ has a graph which is Souslin in $T \times X$. By Proposition 8 in [15, p. 125], $Q \times \bigcap_{n=1}^{\infty} A_n$ is also Souslin in $T \times X$. But $\mu(T - \bigcap_{n=1}^{\infty} A_n) < \varepsilon$ and $\bigcap_{n=1}^{\infty} A_n$ is a closed and hence Souslin subset of T. Hence by Theorem 1(v) of [18] Q is measurable. The last assertion of the lemma follows from the definition of Q.

In the proof of Lemma 6 only the fact that $\{H_n\}$ is measurable was used. Hence with the same proof, the following corollary is valid.

COROLLARY. *Assume a sequence of functions $\{h_n\}$ is measurable and define $\{H_n\}$ as in Lemma 6. Then the set-valued mapping $Q = \bigcap_{n=1}^{\infty} H_n$ is measurable and maps T into the closed convex subsets of X.*

LEMMA 7. *Let $P: T \to \mathcal{K}(X)$ be an integrably bounded measurable set-valued mapping and let g in $L^p(T, R)$, for some p, $1 < p < \infty$, be the bounding function. Then for any measurable $A \subset T$ with $\mu(A) \neq 0$,*

$$\operatorname{cl}\left(\int_A P(t)\, d\mu(t)\right) = \int_A \overline{\operatorname{co}}\, P(t)\, d\mu(t).$$

Moreover the set-valued mapping $t \to \overline{\operatorname{co}}\, P(t)$ is measurable.

Proof. By Theorem 1 in [14], $\operatorname{cl}\left(\int_A P(t)\, d\mu(t)\right)$ is a closed bounded convex subset of X. (Theorem 1 in [14] is stated for bounded mappings. However it remains valid for integrably bounded mappings by a slightly extended but straightforward argument.)

We want to prove that the mapping $t \to \overline{\operatorname{co}}\, P(t)$ is measurable.

For each $x_i' \in \gamma$ (γ is any countable dense set on S') define the function $h_i(t) = x_i'(P(t))$ and the set-valued mapping $t \to H_i(t) = \{x : x_i'(x) \leqq h(t)\}$. By Lemma 1, $\overline{\operatorname{co}}\, P(t) = \bigcap_{i=1}^{\infty} H_i(t)$, and by Lemma 6, $\overline{\operatorname{co}}\, P$ is measurable.

We now show that $\int_A \overline{\operatorname{co}}\, P(t)\, d\mu(t)$ is closed for all measurable $A \subset T$. Let $r \in \operatorname{cl}\left(\int_A \overline{\operatorname{co}}\, P(t)\, d\mu(t)\right)$. Then there exists a sequence $\{r_n\} \subset \int_A \overline{\operatorname{co}}\, P(t)\, d\mu(t)$ which converges weakly to r (this is a consequence of the fact that strong convergence implies weak convergence). Each of the r_n has a representation $r_n = \int_A p_n(t)\, d\mu(t)$ where $\{p_n\} \subset \overline{\operatorname{co}}\, P$. Since $\overline{\operatorname{co}}\, P$ is integrably bounded by $g \in L^p(T, R)$, where $1 < p < \infty$, there exists a subsequence $\{p_q\} \subset \{p_n\}$ which converges weakly in $L^p(T, X)$ to a limit $\hat{p}$ which is also integrably bounded by g (see, e.g., [13, p. 282]). But $\overline{\operatorname{co}}\, P(t)$ is closed and convex for each $t \in T$ and this implies that $\hat{p}(t) \in \overline{\operatorname{co}}\, P(t)$ a.e. and hence, by Lemma 1, that $r = \int_A \hat{p}(t)\, d\mu(t) \in \int_A \overline{\operatorname{co}}\, P(t)\, d\mu(t)$. This proves that $\int_A \overline{\operatorname{co}}\, P(t)\, d\mu(t)$ is closed.

Clearly $\int_A \overline{\text{co}}\, P(t)\, d\mu(t) \supset \text{cl}\left(\int_A P(t)\, d\mu(t)\right)$. Suppose there exists an $\bar{r} \in \int_A \overline{\text{co}}\, P(t)\, d\mu(t)$ such that $\bar{r} \notin \text{cl}\left(\int_A P(t)\, d\mu(t)\right)$. Applying the argument used in proving Lemma 1, we can find an $x_i' \in \gamma$ and an $\varepsilon > 0$ such that $x_i'(\bar{r}) \geqq x_i' \text{cl}\left(\int_A P(t)\, d\mu(t)\right) + \varepsilon$. Notice that $\bar{r} = \int_A \overline{p(t)}\, d\mu(t)$ for some $\bar{p} \in \overline{\text{co}}\, P$. Hence $x_i'(\bar{r}) = \int_A x_i'(\overline{p(t)})\, d\mu(t)$. But $x_i'(\overline{p(t)}) \leqq h_i(t)$ for all t in T. Hence $x_i'(\bar{r}) \leqq \int_A h_i(t)\, d\mu(t)$.

By Theorem 5.4 in [5] there exists a denumerable family of measurable cross sections $\{\varphi_n\}$ in P such that $\{\varphi_n(t)\}$ is dense in $P(t)$ for each $t \in T$. For each integer n define the set

$$E_n = \left\{ t : x_i'(\varphi_n(t)) \geqq h_i(t) - \frac{\varepsilon}{2\mu(A)} \right\}.$$

Then $T = \bigcup_{n=1}^{\infty} E_n$. We now construct from the $\{E_n\}$ a measurable partition of T as follows:

$$\begin{aligned} A_1 &= E_1, \\ A_2 &= E_2 - E_1, \\ &\vdots \\ A_k &= E_k - \bigcup_{j=1}^{k-1} E_j, \\ &\vdots \end{aligned}$$

Define the mapping $\varphi = \sum_{n=1}^{\infty} \chi_{A_n} \varphi_n$ (χ_A is the characteristic function of a set $A \subset T$). Since $\varphi \in P$, we can write the chain of inequalities

$$\begin{aligned} x_i'\left[\int_A \varphi(t)\, d\mu(t)\right] &\geqq \int_A h_i(t)\, d\mu(t) - \frac{\varepsilon}{2} \\ &\geqq x_i'(\bar{r}) - \frac{\varepsilon}{2} \geqq x_i'\left[\int_A \varphi(t)\, d\mu(t)\right] + \frac{\varepsilon}{2}, \end{aligned}$$

which is a contradiction. This proves $\int_A \overline{\text{co}}\, P(t)\, d\mu(t) \subset \text{cl}\left(\int_A P(t)\, d\mu(t)\right)$ and thus that the two sets are equal.

COROLLARY 1. *If $P_i : T \to \mathcal{K}(X)$, $i = 1, 2$, are measurable p-power integrably bounded set-valued mappings and if* $\text{cl}\left(\int_A P_1(t)\, d\mu(t)\right) = \text{cl}\left(\int_A P_2(t)\, d\mu(t)\right)$ *for all measurable $A \subset T$, then* $\overline{\text{co}}\, P_1 = \overline{\text{co}}\, P_2$ *a.e. in T.*

COROLLARY 2. *For all $x' \in X'$ and measurable $A \subset T$,*

$$x'\left(\int_A \overline{\text{co}}\, P(t)\, d\mu(t)\right) = x'\,\text{cl}\left(\int_A P(t)\, d\mu(t)\right)$$
$$= \int_A x'(P(t))\, d\mu(t)$$
$$= \int_A x'(\overline{\text{co}}\, P(t))\, d\mu(t)$$

if $P: T \to \mathscr{K}(X)$ is an integrably bounded measurable set-valued mapping.

Proof. Let $h(t) = x'(P(t))$. By Lemma 5, h is measurable, and by Lemma 7, so is $\overline{\text{co}}\, P$. There exists a sequence of measurable cross sections $\{\varphi_i\}$ of $\overline{\text{co}}\, P$ such that $\{\varphi_i(t)\}$ is dense in $\overline{\text{co}}\, P(t)$ for all t in T (see, e.g., Theorem 5.2 in [5]). Let $\varepsilon > 0$. Define for each natural number n the set

$$E_n(\varepsilon) = \{t : x'(\varphi_n(t)) \geqq h(t) - \varepsilon/\mu(T)\}.$$

Let

$$A_1(\varepsilon) = E_1(\varepsilon),$$
$$\vdots$$
$$A_k(\varepsilon) = E_k(\varepsilon) - \bigcup_{j=1}^{k-1} E_j(\varepsilon).$$

Let $\varphi = \sum_{i=1}^{\infty} \chi_{A_n} \varphi_n$. Then $\varphi \in \overline{\text{co}}\, P$. If A is a measurable set in T,

$$\int_A x'(P(t))\, d\mu(t) = \int_A x'(\overline{\text{co}}\, P(t))\, d\mu(t) \geqq x'\left[\int_A \overline{\text{co}}\, P(t)\, d\mu(t)\right]$$
$$\geqq x'\left[\int_A \varphi(t)\, d\mu(t)\right]$$
$$= \int_A x'(\varphi(t))\, d\mu(t)$$
$$= \sum_{i=1}^{\infty} \int_{A \cap A_i} x'(\varphi_i(t))\, d\mu(t)$$
$$\geqq \sum_{i=1}^{\infty} \int_{A \cap A_i} h(t)\, d\mu(t) - \varepsilon \sum_{i=1}^{\infty} \frac{\mu(A \cap A_i)}{\mu(T)}$$
$$\geqq \int_A h(t)\, d\mu(t) - \varepsilon$$
$$= \int_A x'(\overline{\text{co}}\, P(t))\, d\mu(t) - \varepsilon$$
$$= \int_A x'(P(t))\, d\mu(t) - \varepsilon.$$

Since ε is arbitrary this proves that

$$x'\left(\int_A \overline{\text{co}}\, P(t)\, d\mu(t)\right) = x'\left(\text{cl}\left(\int_A P(t)\, d\mu(t)\right)\right)$$

$$= \int_A x'(P(t))\, d\mu(t)$$

$$= \int_A x'(\overline{\text{co}}\, P(t))\, d\mu(t).$$

LEMMA 8. *Let K_1 and K_2 be any two closed bounded convex sets in X and let $x' \in X'$. Define $K_1 + K_2 = \{k \in X : k_1 + k_2 = k,$ where $k_1 \in K_1$ and $k_2 \in K_2\}$. Then $x'(K_1 + K_2) = x'(K_1) + x'(K_2)$.*

Proof. Trivial.

5. The main results.

THEOREM 1. *Let $\{P^n\}$ be a sequence of nonempty measurable mappings from $T \to \mathscr{K}(X)$ which are integrably bounded by some function g in $L^p(T, R)$, $1 < p < \infty$, and let $v = \{x_i'\}$ be any countably dense subset of S'. Then there exists a subsequence $\{P^q\} \subset \{P^n\}$ and a measurable mapping $P : T \to \overline{\text{co}}\, \mathscr{K}(X)$ such that for any measurable set A in T,*

$$\lim_{q \to \infty} d_v\left(\int_A P^q(t)\, d\mu(t), \int_A P(t)\, d\mu(t)\right) = 0.$$

Proof. One consequence of Lemma 7 is that for each n and measurable set A in T,

$$\int_A \overline{P^n(t)\, d\mu(t)} = \int_A \overline{\text{co}}\, P^n(t)\, d\mu(t).$$

Let $\gamma = \{x_i'\}$ be a countably dense subset of S' and define the mappings $t \to h_i^n(t) = x_i'(\overline{\text{co}}\, P^n(t)) = x_i'(P^n(t))$. By Lemma 5, the functions $\{h_i^n\}$ are measurable, and since $\{P^n\}$ is integrably bounded by $g \in L^p(T, R)$, so is $\{h_i^n\}$. Thus the $\{h_i^n\}$ are weakly compact in $L^p(T, R)$. This means we can find a subsequence $\{q\} \subset \{n\}$ and measurable mappings $\{h_i\}$ such that for each i, $h_i^q \xrightarrow{\omega} h_i$. Define the sequence $\{H_i\}$ by $H_i(t) = \{x : x_i'(x) \leqq h_i(t)\}$ and the set-valued mapping $P : T \to \overline{\text{co}}\, \mathscr{K}(X)$ by

$$P(t) = \bigcap_{i=1}^{\infty} H_i(t) = \overline{\text{co}}\, P(t).$$

By Lemma 5, each of H_i is measurable, and by the corollary to Lemma 6, P is measurable. Using Corollary 2 to Lemma 7 and the corollary to Lemma 3, we can write for each $x_i' \in \gamma$ and measurable $A \subset T$,

$$\lim_{q \to \infty} x_i'\left[\int_A P^q(t)\, d\mu(t)\right] = \lim_{q \to \infty} x_i'\left[\text{cl}\left(\int_A P^q(t)\, d\mu(t)\right)\right]$$

$$= \lim_{q \to \infty} x_i'\left(\int_A \overline{\text{co}}\, P^q(t)\, d\mu(t)\right) = \lim_{q \to \infty} \int_A x_i'(\overline{\text{co}}\, P^q(t))\, d\mu(t)$$

$$= \lim_{q\to\infty} \int_A h_i^q(t)\,d\mu(t) = \int_A h_i(t)\,d\mu(t)$$

$$= x_i'\left(\int_A P(t)\,d\mu(t)\right).$$

Now application of Corollary 1 of Lemma 1 and Remark 2 establishes the desired result.

COROLLARY (Hermes [2]). *Let $\{P^n\}$ be a sequence of measurable closed nonempty set-valued mappings defined from $T \to \mathcal{K}(E^n)$ with values lying in some fixed bounded subset of E^n. For each measurable $E \subset T$ define $Q^k(E) = \int_E P^k(s)\,d\mu(s)$ and assume that for all such E, $\{Q^k(E)\}$ converges in the Hausdorff set topology to a set-valued mapping $A(E)$. Then there exists a measurable set-valued mapping P such that for all measurable $E \subset T$,*

$$Q(E) = \int_E P(t)\,d\mu(t).$$

Proof. The proof is an immediate consequence of Theorem 1 once it is noted that if $X = E^n$, then any measurable integrably bounded mapping $P: T \to \mathcal{K}(X)$ satisfies the equality

$$\int_E \overline{\text{co}}\, P(t)\,d\mu(t) = \int_E P(t)\,d\mu(t)$$

for all measurable $E \subset T$ (see, e.g., Theorem 4 in [1], or Theorem 7.1 in [5]).

THEOREM 2. *Let $E \to Q(E)$ be a set-valued mapping from $2^T \to \overline{\text{co}}\,\mathcal{K}(X)$ such that $Q(\varnothing) = K_0$, a nonempty closed bounded convex set in X. In order that Q be representable in the form $Q(A) = K_0 + \int_A P(t)\,d\mu(t)$, where A is any measurable set in T and $P: T \to \overline{\text{co}}\,\mathcal{K}(X)$ is a measurable integrably bounded set-valued mapping, it is necessary and sufficient that there exist a functional $h: X' \times T \to R$ such that:*

(i) *For t a.e. in T, $h(\cdot, t): X' \to R$ is a continuous positively homogeneous subadditive functional.*

(ii) *For all $x' \in X'$, $h(x, \cdot)$ is a measurable function satisfying the inequality $|h(x', t)| \leqq g(t)\|x'\|$ for some fixed g in $L^p(T, R)$, where p satisfies the condition $1 < p < \infty$.*

(iii) *For any $x' \in S'$ and any measurable set $A \subset T$, $x'[Q(A)] = x'(K_0) + \int_A h(x', s)\,d\mu(s)$.*

Proof of necessity. Suppose for all measurable $A \subset T$, $Q(A) = K_0 + \int_A P(t)\,d\mu(t)$, where $P: T \to \overline{\text{co}}\,\mathcal{K}(X)$ is measurable and integrably bounded by some fixed g in $L^p(T, R)$. Then by Lemma 8 and Corollary 2 of Lemma 7,

$$x'[Q(A)] = x'(K_0) + x'\left[\int_A P(t)\,d\mu(t)\right]$$

$$= x'(K_0) + \int_A x'(P(t))\,d\mu(t).$$

For each $x' \in X'$ let $h(x', t) = x'(P(t))$. By Lemma 5, $h(x, \cdot)$ is measurable and integrably bounded by $g(\cdot)\|x'\|$. Thus conditions (ii) and (iii) above are satisfied. It is trivial to verify condition (i) if we apply Lemma 4.

Proof of sufficiency. Assume conditions (i)–(iii) are satisfied. Let $\gamma = \{x_i'\}$ be any countably dense subset of S'. For each i define

$$H_i(t) = \{q : x_i'(q) \leqq h(t, x_i')\}.$$

By the corollary to Lemma 6 there exists a measurable mapping P from T into the closed convex subsets of X given by the expression $P(t) = \bigcap_{i=1}^{\infty} H_i(t)$. Lemma 3 shows that $P(t) \neq \varnothing$ for all $t \in T$. Since

$$\sup\{\|q\| : q \in P(t)\} = \sup\{x_i'(q) : x_i' \in \gamma, q \in P(t)\} \leqq g(t),$$

it follows that P is integrably bounded by g and that $P(t) \in \overline{\text{co}}\, \mathscr{K}(X)$ a.e. on T. Applying Lemma 8, Corollary 2 to Lemma 7, the corollary to Lemma 3 and condition (iii) we can write for any measurable set $A \subset T$ and any $x_i' \in \gamma$,

$$\begin{aligned} x_i'\left[K_0 + \int_A P(t)\, d\mu(t)\right] &= x_i'(K_0) + x_i'\left(\int_A P(t)\, d\mu(t)\right) \\ &= x_i'(K_0) + \int_A x_i'(P(t))\, d\mu(t) \\ &= x_i'(K_0) + \int_A h(x_i', t)\, d\mu(t) \\ &= x_i'(Q(A)). \end{aligned}$$

By Lemma 1, this shows that

$$Q(A) = K_0 + \int_A P(t)\, d\mu(t)$$

and completes the proof of sufficiency.

Appendix. The following definition and results are found in [5].

DEFINITION. Let T be a locally compact space and μ a Randon measure on T. A set-valued mapping Γ from T into a topological space will be said to be μ-measurable if the set $\Gamma^{-}A = \{t \in T : \Gamma(t) \cap A \neq \varnothing\}$ is μ-measurable for all closed sets A in E.

THEOREM 5.2. *Let T be a locally compact space, μ a measure on T and Γ a set-valued mapping from T into a Polish space E with values in the nonempty closed subsets of E. Then Γ admits a μ-measurable section.*

THEOREM 5.4. *Let T be locally compact, μ a measure on T and Γ a μ-measurable mapping from T into a Polish space E with values in the nonempty closed subsets of E. Then there exists a denumerable family $\{\sigma_i\}$ of μ-measurable sections of Γ such that the collection $\{\sigma_i(t)\}$ is dense in $\Gamma(t)$ for all t in T.*

Let Γ be an integrably bounded set-valued mapping from a locally compact space T into a finite-dimensional space F which is μ-measurable and takes its values in the compact subsets of F. Let $\beta_\Gamma = \{f : f$ is measurable and $f(t) \in \Gamma(t)$

for $t \in T\}$ and $\hat{\beta}_\Gamma = \{f : f$ is measurable and $f(t) \in \operatorname{co} \Gamma(t)$ for $t \in T\}$.

THEOREM 7.1. *Let* $\int \hat{\Gamma}\, d\mu = \left\{\int f\, d\mu : f \in \beta_\Gamma\right\}$ *and* $\int \Gamma\, d\mu = \left\{\int f\, d\mu : f \in \beta_\Gamma\right\}$. *Then if* μ *is atomless and if* μ *is bounded then* $\int \hat{\Gamma}\, d\mu = \int \Gamma\, d\mu$.

The following result can be found in [14] and is used in the proof of Lemma 7. This result is essentially a rephrasing of Theorem 1 in that paper.

THEOREM. *Let* T *be a measure space endowed with a finite positive nonatomic measure* μ *and let* X *be a Banach space. Suppose* P *is a measurable set-valued mapping from* T *into* X *which is uniformly bounded. Then for any measurable set* $A \subset T$, $\operatorname{cl}\left(\int_A P(t)\, d\mu(t)\right)$ *is a closed bounded convex subset of* X.

It should be remarked that in [14] the assumption that P is uniformly bounded can be replaced by the assumption that P is integrably bounded, since the boundedness of P is used only to establish that every measurable cross section in P is integrable.

Acknowledgment. I should like to thank the referee for his many useful suggestions, in particular the need to correct one serious error in the statement and proof of Lemma 7 and for calling to my attention references [3], [16], [18] and [19].

REFERENCES

[1] R. J. AUMANN, *Integrals of set-valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1–12.
[2] H. HERMES, *Calculus of set valued functions and control*, J. Math. Mech., 18 (1968), pp. 47–59.
[3] G. DEBREU, *Integration of correspondences*, Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. II, Part I, Univ. of Calif. Press, Berkeley and Los Angeles, 1967, pp. 351–372.
[4] H. RÅDSTRÖM, *An embedding theorem for spaces of convex sets*, Proc. Amer. Math. Soc., 3 (1952), pp. 165–169.
[5] C. CASTAING, *Sur les multi-applications mesurables*, Doctoral thesis, L'Universite de Caen, Caen, 1967.
[6] ———, *Sur une extension du theoreme de Liapunov*, C. R. Acad. Sci. Paris Sér. A-B, 260 (1965), pp. 3838–3841.
[7] ———, *Quelques problemes de mesurabilite lies à la theorie de la commande*, Ibid., 262 (1966), pp. 409–411.
[8] ———, *Sur les equations differentielles multivoques*, Ibid., 263 (1966), pp. 63–66.
[9] ———, *Sur une nouvelle extension du theoreme de Liapunov*, Ibid., 264 (1967), pp. 336–339.
[10] ———, *Sur les multi-applications mesurables*, Rev. Française Informatique et Recherche Operationnelle, 1 (1967), pp. 91–126.
[11] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. Part I*, John Wiley, New York, 1958.
[12] C. BERGE, *Espaces topologiques functions multivoques*, Dunod, Paris, 1959.
[13] N. DINCULEANU, *Vector Measures*, vol. 95, Pergamon Press, London, 1967.
[14] R. DATKO, *Convexity properties of some integral operators*, J. Differential Equations, to appear.
[15] N. BOURBAKI, *Topologie generale*, Elements de mathematiques, Livre III, Hermann, Paris, 1958, Chap. 9.
[16] E. MICHAEL, *Topologies on spaces of subsets*, Trans. Amer. Math. Soc., 71 (1951), pp. 152–182.
[17] J. L. KELLEY, *General Topology*, Van Nostrand, Princeton, 1955.
[18] C. J. HIMMELBERG, M. Q. JACOBS AND F. S. VAN FLECK, *Measurable Multifunctions, Selectors and Filippov's Implicit Functions Lemma*, J. Math. Anal. Appl., 25 (1969), pp. 276–284.
[19] K. KURATOWSKI AND C. RYLL-NARDZEWSKI, *A general theorem on selectors*, Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys., 13 (1965), no. 6, pp. 397–403.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

ON CERTAIN CONVERGENCE QUESTIONS IN SYSTEM IDENTIFICATION*

M. AOKI AND P. C. YUE†

Abstract. This paper examines the asymptotic properties of the maximum likelihood estimates of the unknown parameters and the unknown initial state of linear, stable, constant coefficient, discrete time dynamic systems where plant noise and observation noise are present. Necessary and sufficient conditions are obtained for the system parameter estimates to converge with probability one, to be asymptotically normal and to converge in mean square. These conditions require that the system representation be unique and impose a simple constraint on the input sequence. Under these conditions, the initial state estimate is shown to be asymptotically unbiased and have finite covariance.

1. Introduction. Problems of estimating the parameters and the initial state of a dynamic system are known as identification (or system determination) problems. The method of maximum likelihood has been widely accepted as a standard estimation technique and it is generally believed that this method provides estimates with many attractive asymptotic properties. These properties were first treated by Wald [1] specifically for identification problems where the observables, in general, are not independent samples from the same distribution. Wald's sufficient conditions for the estimates' consistency are overly restrictive for most systems of interest. Levin [2] considered deterministic systems with noisy observations of input and output signals. The estimation method he adopted only gave an approximate maximum likelihood estimate. Levin referred to Koopmans' early work [3] for the asymptotic properties but Koopmans' discussion was not very precise, and the presentation rather obscure. Recently, a complete analysis of this method has been given [4]. Computational experience is reported by many authors (see, e.g., [4], [5], [6]).

Astrom et al. [7] treated the case where the random disturbance may also be present in the system itself. They used a combined model which directly relates the system's input to the noisy output observations without explicitly distinguishing between plant noise and observation noise. (See (46)–(48) for exact mathematical models used by Astrom in comparison to this paper.) Since for both prediction and control purposes the main interest lies in how the input affects the output (rather than the noisy observations), it is more suitable to separate the plant dynamics from the observation equation. However, their discussion on the asymptotic properties of the combined model parameters are fairly complete; and they established sufficient conditions for convergence with probability one.

This paper considers the formulation which involves explicit equations for the plant dynamics and the observations, and shows that even though the likelihood function in the formulation of the paper differs from Astrom's, analogous results on the asymptotic properties of the m.l.e. (maximum likelihood estimate) can be obtained. In particular, a set of conditions are given which are not only sufficient but also necessary for the system parameter estimates to converge in probability, with

* Received by the editors July 1, 1969, and in revised form October 17, 1969.

† School of Engineering and Applied Science, University of California, Los Angeles, California 90024. This work was supported in part by the National Science Foundation under Grant GK-2032.

probability one and in mean square. The necessity part is important in optimal input synthesis for identification since in general the least amount of constraint on the input is desired. The significance of various assumptions is also discussed. Furthermore, the initial state estimate is shown to be asymptotically unbiased and bounds are obtained for its covariance.

2. System representation. First, we consider dynamic systems representable by

$$\begin{aligned} \mathbf{z}(t+1) &= \Phi\mathbf{z}(t) + \mathbf{b}u_t, \\ x_t &= \langle \mathbf{h}, \mathbf{z}(t) \rangle, \end{aligned} \tag{1}$$

where the state $\mathbf{z}(t)$ is a k-vector. The reduced state space (reduced in the sense that no distinct states are equivalent, following Zadeh [12]) is known a priori to be k-dimensional. Thus, by a suitable choice of basis, the pair $\{\Phi, \mathbf{h}\}$ has the following completely observable companion form:

$$\Phi = \begin{pmatrix} -a_1 & 1 & & \\ -a_2 & & 1 & \\ \vdots & & & \ddots \\ -a_k & 0 & 0 \cdots & 0 \end{pmatrix}_{k\times k}, \qquad h = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{k\times 1}$$

and $\mathbf{b} = (b_1, b_2, \cdots, b_k)^{\mathsf{T}}$. The initial state, $\mathbf{z}(0) = \mathbf{z}_0$, is an unknown constant k-vector. $\{a_i, b_i, 1 \leqq i \leqq k\}$ are unknown constants.

Observations are made on the output such that

$$y_t = x_t + \eta_t, \tag{2}$$

where the additive noise $\{\eta_t\}$ is assumed for simplicity to be independent and identically distributed as $N(0; \sigma^2)$. The input sequence is known exactly and is assumed to be uniformly bounded.

It is straightforward to verify that (1) has a simple input-output relation:

$$x_t + \sum_{i=1}^{k} a_i x_{t-i} = \sum_{i=1}^{k} b_i u_{t-i},$$

and for the interval $\{t = 0, 1, 2, \cdots, N-1\}$, a compact notation using Toeplitz matrices can be adopted to represent the system's behavior as follows. Let

$$\begin{aligned} \mathbf{u}_N &= (u_0, u_1, \cdots, u_{N-1})^{\mathsf{T}}, & \boldsymbol{\eta}_N &= (\eta_0, \eta_1, \cdots, \eta_{N-1})^{\mathsf{T}}, \\ \mathbf{x}_N &= (x_0, x_1, \cdots, x_{N-1})^{\mathsf{T}}, & \mathbf{y}_N &= (y_0, y_1, \cdots, y_{N-1})^{\mathsf{T}}. \end{aligned}$$

Then,

$$A_N\mathbf{x}_N = B_N\mathbf{u}_N + E_N\mathbf{z}_0, \tag{1.A}$$

where

$$A_N = I_N + \sum_{i=1}^{k} a_i S^i \qquad (N \times N),$$

$$B_N = \sum_{i=1}^{k} b_i S^i \qquad (N \times N), \tag{3}$$

$$E_N = \begin{pmatrix} I_k \\ \hline 0_{N-k,k} \end{pmatrix} \qquad (N \times k), \tag{4}$$

and S is the $N \times N$ shift matrix whose (i, j) element is equal to $\delta_{i,j+1}$.

Since the set of lower triangular Toeplitz matrices form a commutative ring with identity under the ordinary matrix operations (with the usual zero and identity) and the subset of nonsingular lower triangular Toeplitz matrices are the group of units of the ring, this notation will prove to be very convenient for algebraic manipulations.

Another way of expressing the input-output behavior of (1) is in terms of the parameter vector $\boldsymbol{\theta}$, the initial state $\mathbf{z}_0$ and a matrix H_N such that

$$\mathbf{x}_N = H_N\boldsymbol{\theta} + E_N\mathbf{z}_0, \tag{1.B}$$

where

$$\boldsymbol{\theta} = (a_1, a_2, \cdots, a_k, b_1, b_2, \cdots, b_k)^{\mathsf{T}}, \tag{5}$$

and

$$H_N = (-S\mathbf{x}_N, -S^2\mathbf{x}_N, \cdots, -S^k\mathbf{x}_N, S\mathbf{u}_N, S^2\mathbf{u}_N, \cdots, S^k\mathbf{u}_N)_{N\times 2k} \tag{6}$$

since, by (1.A),

$$\mathbf{x}_N = -\sum_{i=1}^{k} a_i S^i \mathbf{x}_N + \sum_{i=1}^{k} b_i S^i \mathbf{u}_N + E_N\mathbf{z}_0. \tag{7}$$

This notation is useful in stating many major results in later discussions.

Both (1.A) and (1.B) are equivalent to system (1) with respect to $\mathbf{u}_N$ and $\mathbf{x}_N$ for arbitrary $\mathbf{z}_0$. There remains, however, the question whether these representations have a unique set of values for the parameter vector, $\boldsymbol{\theta}$.

PROPOSITION 1. *For any initial state* $\mathbf{z}_0$, $\boldsymbol{\theta}$ *is uniquely determined from* $\mathbf{u}_N$ *and* $\mathbf{x}_N$ *if and only if* $H_N^{\mathsf{T}}H_N > 0$, *whereupon*

$$\boldsymbol{\theta} = (H_N^{\mathsf{T}}H_N)^{-1}H_N^{\mathsf{T}}(\mathbf{x}_N - E_N\mathbf{z}_0). \tag{8}$$

Proof. It follows immediately from (1.B). This result was first stated in [9].

PROPOSITION 2. *Suppose* $\mathbf{z}_0 = \mathbf{0}$; *then* $\boldsymbol{\theta}$ *is uniquely determined from* $\mathbf{u}_N$, $\mathbf{x}_N$ *if and only if*:

(i) $N \geqq 2k$;

(ii) $\{b_i\}$ *are not all zero,* $1 \leqq i \leqq k$;

(iii) $\{u_i\}$ *are not identically zero for* $0 \leqq i \leqq N - 2k$;

(iv) *the polynomials* $A(z)$ *and* $B(z)$ *do not have a common divisor where*

$$A(z) = 1 + \sum_{i=1}^{k} a_i z^i, \qquad B(z) = \sum_{i=1}^{k} b_i z^i.$$

Proof of necessity.

(i) H_N is $N \times 2k$. $H_N^{\mathsf{T}}H_N > 0$ implies $N \geqq 2k$.

(ii) If $b_i \equiv 0$, then $B_N = 0$, $\mathbf{x}_N = A_N^{-1}B_N\mathbf{u}_N = \mathbf{0}$.
Consequently, any A_N satisfies (1.A).

(iii) If $u_i \equiv 0$ for $0 \leqq i \leqq N - 2k$, then by (1), $x_i \equiv 0$ also for $0 \leqq i \leqq N - 2k$. Thus, in (7), the first $N - 2k + 1$ rows of H_N become zero, rank $H_N < 2k$, and $H_N^{\mathsf{T}} H_N \not> 0$.

(iv) If $A(z)$ and $B(z)$ have a common divisor, then

$$A_N = \tilde{A}_N D_N, \qquad B_N = \tilde{B}_N D_N, \tag{9}$$

where

$$\tilde{A}_N = I_N + \sum_{i=1}^{k_1} \tilde{a}_i S^i, \qquad \tilde{B}_N = \sum_{i=1}^{k_1} \tilde{b}_i S^i, \qquad k_1 < k,$$

and

$$D_N = d_0 I_N + \sum_{i=1}^{k-k_1} d_i S^i, \qquad d_0 \neq 0.$$

Substituting (9) into (1.A) and multiplying both sides by D_N^{-1} since $|D_N| \neq 0$, we obtain

$$\tilde{A}_N \mathbf{x}_N = \tilde{B}_N \mathbf{u}_N;$$

then $\tilde{\boldsymbol{\theta}} = (\tilde{a}_1, \tilde{a}_2, \cdots, \tilde{a}_{k_1}, 0, \cdots, 0, \tilde{b}_1, \cdots, \tilde{b}_{k_1}, 0, \cdots, 0)$ would satisfy (1.A) and contradict the uniqueness assumption.

Proof of sufficiency. Let $\tilde{\boldsymbol{\theta}} = (\tilde{a}_1, \tilde{a}_2, \cdots, \tilde{a}_k, \tilde{b}_1, \tilde{b}_2, \cdots, \tilde{b}_k)^{\mathsf{T}}$ be any vector such that the corresponding matrices $\tilde{A}_N$, $\tilde{B}_N$ satisfy (1.A). Then $\tilde{A}_N \mathbf{x}_N = \tilde{B}_N \mathbf{u}_N$, $A_N \mathbf{x}_N = B_N \mathbf{u}_N$. Therefore,

$$\begin{aligned} A_N \tilde{B}_N \mathbf{u}_N &= A_N \tilde{A}_N \mathbf{x}_N \\ &= \tilde{A}_N A_N \mathbf{x}_N = \tilde{A}_N B_N \mathbf{u}_N. \end{aligned} \tag{10}$$

Let

$$\begin{aligned} C_N &= A_N \tilde{B}_N - \tilde{A}_N B_N \\ &= \sum_{i=1}^{2k} c_i S^i. \end{aligned} \tag{11}$$

Let

$$U_{N,2k} = (S\mathbf{u}_N, S^2\mathbf{u}_N, \cdots, S^{2k}\mathbf{u}_N), \tag{12}$$

$$\mathbf{c} = (c_1, c_2, \cdots, c_{2k})^{\mathsf{T}}. \tag{13}$$

Combining (10)–(13), we obtain

$$U_{N,2k}\mathbf{c} = C_N \mathbf{u}_N = 0. \tag{14}$$

By (i) and (iii), $U_{N,2k}$ has rank $2k$. Hence $\mathbf{c} = \mathbf{0}$, and by (11),

$$A_N \tilde{B}_N = \tilde{A}_N B_N.$$

This implies $B(z)/A(z) = \tilde{B}(z)/\tilde{A}(z)$; and by (ii), (iv), $\mathbf{a} = \tilde{\mathbf{a}}$, $\mathbf{b} = \tilde{\mathbf{b}}$. This completes the proof.

The parameter being unique means that u_i, x_i are realizable as the input and output sequences of a dynamic system (1/1.A/1.B) such that the realization has

minimal order, k. The conditions (ii)–(iii) are equivalent to complete controllability of the pair $\{\Phi, \mathbf{b}\}$ (see, e.g., [12]). See also [11] for a more general discussion of the realization problem in the case of multiple input and multiple output.

When the parameter in the representation is unique, all discussions on parameter estimates can be simplified greatly since there exists a unique set of values that can be qualified as the true parameter values. Otherwise, the parameter estimation problem is solvable only up to an equivalence class (input-output equivalence) and the question of canonical form would inevitably complicate the treatment.

3. Characterization of the maximum likelihood estimates. In the sequel we will restrict ourselves to the case where the true parameter $\boldsymbol{\theta}^0$ is known to lie in the interior of a given compact subset $\textcircled{H}_S$ of R_{2k} and where every system with $\boldsymbol{\theta}$ in $\textcircled{H}_S$ is stable, i.e., $A(z)$ has zeros outside the unit circle. The assumption of stability is made to facilitate the study of asymptotic properties. The assumption of compactness, however, is a practical one since usually, from a priori knowledge of the system, the parameter values must fall within a certain range thus allowing us to contain the parameter set in a compact region.

From (1.A) and (2), the observables are related to the unknown parameters and the unknown initial state by

$$\begin{aligned} \mathbf{y}_N &= \boldsymbol{\eta}_N + \mathbf{x}_N \\ &= \boldsymbol{\eta}_N + A_N^{-1}(B_N\mathbf{u}_N + E_N\mathbf{z}_0), \end{aligned} \tag{15}$$

$$p(\mathbf{y}_N|\boldsymbol{\theta}, \mathbf{z}_0) = \text{const.} \cdot \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}_N - A_N^{-1}(B_N\mathbf{u}_N + E_N\mathbf{z}_0)\|^2 \right). \tag{16}$$

Let $\hat{\boldsymbol{\theta}}_N$, $\hat{\mathbf{z}}_{0N}$ be the maximum likelihood estimates (m.l.e.) of $\boldsymbol{\theta}$ and $\mathbf{z}_0$ from $\mathbf{u}_N$ and $\mathbf{y}_N$; that is,

$$l(\hat{\boldsymbol{\theta}}_N, \hat{\mathbf{z}}_{0N}) = \max_{\boldsymbol{\theta} \in \textcircled{H}_S, \mathbf{z}_0 \in R_k} l(\boldsymbol{\theta}, \mathbf{z}_0), \tag{17}$$

where $l(\boldsymbol{\theta}, \mathbf{z}_0) = \log p(\mathbf{y}_N|\boldsymbol{\theta}, \mathbf{z}_0)$. For any $\boldsymbol{\theta} \in \textcircled{H}_S$, $\max_{\mathbf{z}_0 \in R_k} l(\boldsymbol{\theta}, \mathbf{z}_0)$ is achieved by

$$\tilde{\mathbf{z}}_{0N}(\boldsymbol{\theta}) = (E_N^T A_N^{-T} A_N^{-1} E_N)^{-1} E_N^T A_N^{-T}(\mathbf{y}_N - A_N^{-1} B_N \mathbf{u}_N). \tag{18}$$

Thus, $\hat{\boldsymbol{\theta}}_N$ is obtained by

$$\min_{\boldsymbol{\theta} \in \textcircled{H}_S} J_N(\boldsymbol{\theta}) = J_N(\hat{\boldsymbol{\theta}}_N), \tag{19}$$

where

$$\begin{aligned} J_N(\boldsymbol{\theta}) &= \|\mathbf{y}_N - A_N^{-1}(B_N\mathbf{u}_N + E_N\tilde{\mathbf{z}}_{0N}(\boldsymbol{\theta}))\|^2 \\ &= \|A_N\mathbf{y}_N - B_N\mathbf{u}_N - E_N\tilde{\mathbf{z}}_{0N}(\boldsymbol{\theta})\|^2_{(A_N A_N^T)^{-1}}, \end{aligned} \tag{20}$$

and

$$\hat{\mathbf{z}}_{0N} = \tilde{\mathbf{z}}_{0N}(\hat{\boldsymbol{\theta}}_N). \tag{21}$$

Astrom et al. [7] suggested that Wald's classical proof [8] for the consistency of m.l.e. could be easily modified if the almost sure (a.s.) convergence of the likelihood function could be established.

PROPOSITION 3. *For all* $\boldsymbol{\theta} \in \textcircled{H}_S$, $J_N(\boldsymbol{\theta})/N \to J(\boldsymbol{\theta})$ *with probability one, where*

$$J(\boldsymbol{\theta}) = \lim_{N\to\infty} \frac{1}{N} EJ_N(\boldsymbol{\theta}) \tag{22}$$
$$= \sigma^2 + \lim_{N\to\infty} \frac{1}{N} \|A_N \mathbf{x}_N - B_N \mathbf{u}_N^T\|^2_{(A_N A_N^T)^{-1}}.$$

Proof. In (20), the vector $E_N \tilde{\mathbf{z}}_{0N}$ has only a finite number, k, of nonzero elements. As $N \to \infty$, it contributes nothing to $J_N(\boldsymbol{\theta})/N$ in the limit. Therefore, $\tilde{\mathbf{z}}_{0N}$ can be dropped from (20) without any loss of generality.

$$\frac{1}{N} J_N(\boldsymbol{\theta}) = \frac{1}{N} \|A_N \boldsymbol{\eta}_N + A_N \mathbf{x}_N - B_N \mathbf{u}_N\|^2_{(A_N A_N^T)^{-1}} \tag{23}$$

$$= \frac{1}{N} \|A_N \mathbf{x}_N - B_N \mathbf{u}_N\|^2_{(A_N A_N^T)^{-1}} + \frac{2}{N} \langle \boldsymbol{\eta}_N, \mathbf{x}_N - A_N^{-1} B_N \mathbf{u}_N \rangle + \frac{1}{N} \|\boldsymbol{\eta}_N\|^2. \tag{23a}$$

The first term is deterministic. In the second term, both $\mathbf{x}_N$ and $A_N^{-1} B_N \mathbf{u}_N$ represent output sequences of some stable system with a bounded input sequence. Thus $\mathbf{x}_N - A_N^{-1} B_N \mathbf{u}_N$ is uniformly bounded. The limit in (22) exists, and

$$\frac{1}{N} \langle \boldsymbol{\eta}_N, \mathbf{x}_N - A_N^{-1} B_N \mathbf{u}_N \rangle = \frac{1}{N} \sum_{i=0}^{N-1} \alpha_i \eta_i, \tag{24}$$

where $\alpha_i \leqq \alpha < \infty$ for all i. Since $\{\eta_i\}$ are independent random variables with

$$E\eta_i = 0, \qquad E\eta_i^2 = \sigma^2 < \infty$$

and

$$\sum_i \frac{\alpha_i^2}{i^2} \leqq \alpha^2 \sum_i \frac{1}{i^2} < \infty,$$

the strong law of large numbers applies. Hence, with probability one,

$$\frac{1}{N} \sum_i \alpha_i \eta_i \to 0, \tag{25}$$

$$\frac{1}{N} \sum_i \eta_i^2 \to \sigma^2, \tag{26}$$

and substituting (24)–(26) into (23a), we have

$$\frac{1}{N} J_N(\boldsymbol{\theta}) \to J(\boldsymbol{\theta})$$

with probability one. This completes the proof.

Note that $J(\boldsymbol{\theta}^0) = \sigma^2 = \min J(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}^0$ which satisfies (1.A), namely, the true parameter vector if $\boldsymbol{\theta}^0$ is unique in the representation (1.A). Throughout this paper, A_N^0, B_N^0, C_N^0, $A^0(z)$, $B^0(z)$, $C^0(z)$ denote A_N, B_N, C_N, $A(z)$, $B(z)$, $C(z)$, respectively, with $\boldsymbol{\theta}^0$ as parameter. As we see in the next proposition, only those $\boldsymbol{\theta}$ which give rise to $J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}^0) = \sigma^2$ are of interest to us.

PROPOSITION 4. *With probability one,* $\hat{\boldsymbol{\theta}}_N$ *converges to* $\hat{\boldsymbol{\theta}} \in \textcircled{H}^0 \cap \textcircled{H}_S$, *where*

$$\textcircled{H}^0 = \{\boldsymbol{\theta} | J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}^0)\}. \tag{27}$$

Proof. This is a modification of Wald's theorem [8] with the "independent observables" assumption replaced by the results of Proposition 3. The proof for dependent random noises has been given by Astrom et al. [7], and is valid for our formulation with slight refinement. See Appendix A.

PROPOSITION 5. $J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}^0)$ *if and only if*

$$\lim_{N\to\infty} \frac{1}{N} \|(A_N B_N^0 - A_N^0 B_N)\mathbf{u}_N\|^2 = 0. \tag{28}$$

Proof. From (22),

$$J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}^0) + \lim_{N\to\infty} \frac{1}{N} \|A_N \mathbf{x}_N - B_N \mathbf{u}_N\|^2_{(A_N A_N^{\mathrm{T}})^{-1}}.$$

$$\lim_{N\to\infty} \frac{1}{N} \|A_N \mathbf{x}_N - B_N \mathbf{u}_N\|^2_{(A_N A_N^{\mathrm{T}})^{-1}}$$

$$= \lim_{N\to\infty} \frac{1}{N} \|A_N A_N^{0^{-1}} (B_N^0 \mathbf{u}_N + E_N \mathbf{z}_0) - B_N \mathbf{u}_N\|^2_{(A_N A_N^{0\mathrm{T}})^{-1}}.$$

$$= \lim_{N\to\infty} \frac{1}{N} \|B_N^0 \mathbf{u}_N + E_N \mathbf{z}_0 - A_N^{-1} A_N^0 B_N \mathbf{u}_N\|^2_{A_N^{0\mathrm{T}-1} A_N^{0-1}}.$$

From Appendix B, $\rho_1 I_N \leqq A_N A_N^{\mathrm{T}} \leqq \rho_2 I_N$ for all $\boldsymbol{\theta} \in \textcircled{H}_S$, where

$$0 < \rho_1 < \rho_2 < \infty. \tag{29}$$

By repeated application of (29), we deduce that $J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}^0)$ if and only if

$$0 = \lim_{N\to\infty} \frac{1}{N} \|B_N^0 \mathbf{u}_N + E_N \mathbf{z}^0 - A_N^{-1} A_N^0 B_N \mathbf{u}_N\|^2,$$

or

$$0 = \lim_{N\to\infty} \frac{1}{N} \|B_N^0 \mathbf{u}_N - A_N^{-1} A_N^0 B_N \mathbf{u}_N\|^2$$

$$= \lim_{N\to\infty} \frac{1}{N} \|A_N B_N^0 \mathbf{u}_N - A_N^0 B_N \mathbf{u}_N\|^2_{(A_N A_N^{\mathrm{T}})^{-1}};$$

and by (29), this is true if and only if

$$\lim_{N\to\infty} \frac{1}{N} \|A_N B_N^0 \mathbf{u}_N - A_N^0 B_N \mathbf{u}_N\|^2 = 0.$$

4. Consistency and mean-square convergence.

PROPOSITION 4a. *Given that* $\boldsymbol{\theta}^0$ *is unique, the m.l.e.* $\hat{\boldsymbol{\theta}}_N$ *converges to* $\boldsymbol{\theta}^0$ *with probability one if and only if* $\textcircled{H}^0 \cap \textcircled{H}_S$ *is a singleton.*

Proof. It follows immediately from Proposition 4.

A necessary and sufficient condition is now given to ensure that the condition of Proposition 4a is always satisfied.

THEOREM 1. *Given the linear dynamic system* (1/1.A/1.B) *such that* $\mathbf{b} \neq \mathbf{0}$ *and* $\{\Phi, \mathbf{b}\}$ *is completely controllable, the m.l.e.* $\hat{\boldsymbol{\theta}}_N$ *converges to* $\boldsymbol{\theta}^0$ *with probability one if and only if*

$$\lim_{N\to\infty} \frac{1}{N} U_{N,2k}^{\mathsf{T}} U_{N,2k} > 0, \tag{30}$$

where

$$U_{N,2k} = (S\mathbf{u}_N, S^2\mathbf{u}_N, \cdots, S^{2k}\mathbf{u}_N). \tag{31}$$

Proof of sufficiency. Define the matrix C_N and the associated vector $\mathbf{c}$ by

$$\begin{aligned} C_N &= A_N B_N^0 - A_N^0 B_N \\ &= \sum_{i=1}^{2k} c_i S^i, \end{aligned} \tag{32}$$

$$\mathbf{c} = (c_1, c_2, \cdots, c_{2k})^{\mathsf{T}}. \tag{32a}$$

Then,

$$0 = \lim_{N\to\infty} \|(A_N B_N^0 - A_N^0 B_N)\mathbf{u}_N\|^2$$

if and only if

$$\lim_{N\to\infty} \|C_N\mathbf{u}_N\|^2 = \langle \mathbf{c}, \lim_{N\to\infty} \frac{1}{N} U_{N,2k}^{\mathsf{T}} U_{N,2k}\mathbf{c}\rangle = 0. \tag{33}$$

Condition (30) implies that $\mathbf{c} = \mathbf{0}$, or, equivalently,

$$C(z) = A(z)B^0(z) - A^0(z)B(z) = 0,$$

$$B(z)/A(z) = B^0(z)/A^0(z).$$

Hence, by controllability, $\mathbf{b} = \mathbf{b}^0$, $\mathbf{a} = \mathbf{a}^0$, $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ and $\textcircled{H}_S \cap \textcircled{H}^0$ is a singleton. This result is essentially due to Astrom [7].

Proof of necessity. It suffices to show that if (30) is not satisfied, then there exists $\boldsymbol{\theta} = (\mathbf{a}^0 + \delta\mathbf{a}, \mathbf{b}^0 + \delta\mathbf{b})$ such that $\delta\mathbf{a} \neq \mathbf{0}$, $\delta\mathbf{b} \neq \mathbf{0}$, the condition (28) is satisfied, and $\boldsymbol{\theta} \in \textcircled{H}^0 \cap \textcircled{H}_S$.

Note that the vector $\mathbf{c}$ as defined in (32), (32a) can be reexpressed as

$$\mathbf{c} = T_{b^0}\mathbf{a} + T_{a^0}\mathbf{b} + E_{2k}\mathbf{b}^0, \tag{34}$$

where

$$T_{a^0} = -A_{2k}^0 E_{2k} \qquad (2k \times k),$$

$$T_{b^0} = B_{2k}^0 E_{2k} \qquad (2k \times k),$$

$$E_{2k} = \begin{pmatrix} I_k \\ \hline 0_{k,k} \end{pmatrix} \qquad (2k \times k).$$

Let $V = \lim_{N\to\infty} U_{N,2k}^{\mathsf{T}} U_{N,2k}/N$. If the matrix V is not positive definite, then there

exists a nontrivial solution to the equation

$$\mathbf{0} = V(T_{b^0}, T_{a^0})\begin{pmatrix}\delta\mathbf{a}\\ \delta\mathbf{b}\end{pmatrix}. \tag{35}$$

It is immediate from the definition of C_N that $C_N^0 = 0$ and thus

$$\mathbf{0} = T_{b^0}\mathbf{a}^0 + T_{a^0}\mathbf{b}^0 + E_{2k}\mathbf{b}^0. \tag{36}$$

Therefore, letting $\mathbf{a} = \mathbf{a}^0 + \alpha \cdot \delta\mathbf{a}$, $\mathbf{b} = \mathbf{b}^0 + \alpha \cdot \delta\mathbf{b}$ for any scalar α, we obtain by combining (34)–(36) that $\langle \mathbf{c}, V\mathbf{c}\rangle = 0$ which by (33) implies that the condition (28) is satisfied and $\boldsymbol{\theta} \in \textcircled{H}^0$. It only remains to show $\boldsymbol{\theta} \in \textcircled{H}_S$.

Suppose $\{\lambda_i(A)\}$, $1 \leqq i \leqq k$, are the roots of $A(z)$. Then $\lambda_i(A^0)$ are exterior points of the unit disc $D = \{z : |z| \leqq 1\}$ on the complex plane by stability. Since the roots of $A(z)$ are continuous in $\mathbf{a}$ at $\mathbf{a}^0$ in the sense that there exists a neighborhood $\textcircled{H}_a$ where $\textcircled{H}_a = \{\mathbf{a} : \|\mathbf{a} - \mathbf{a}^0\| < \varepsilon\}$ such that $\lambda_i(A) \in D$ for all $\mathbf{a} \in \textcircled{H}_a$, clearly $\mathbf{a} = \mathbf{a}^0 + \alpha \cdot \delta\mathbf{a}$ is stable if $\alpha < \varepsilon/\|\delta\mathbf{a}\|$. Furthermore, $\boldsymbol{\theta}^0$ is an exterior point of $\textcircled{H}_S$. Thus, α can be chosen to have

$$\boldsymbol{\theta} = (\mathbf{a}^0 + \alpha \cdot \delta\mathbf{a}, \mathbf{b}^0 + \alpha \cdot \delta\mathbf{b}) \in \textcircled{H}_S.$$

The necessary and sufficient condition of Theorem 1 can also be stated in various forms for the purpose of different applications.

COROLLARY 1.1. *Given the linear dynamic system* (1/1.A/1.B) *such that* $\mathbf{b} \neq \mathbf{0}$ *and* $\{\Phi, \mathbf{b}\}$ *is completely controllable, the m.l.e.* $\hat{\boldsymbol{\theta}}_N$ *converges to* $\boldsymbol{\theta}^0$ *with probability one if and only if*

$$\lim_{N\to\infty} \frac{1}{N} H_N^T H_N > 0 \tag{30a}$$

or, equivalently,

$$\lim_{N\to\infty} \frac{1}{N} M_N > 0, \tag{30b}$$

where

$$M_N = \frac{1}{\sigma^2} H_N^T (A_N A_N^T)^{-1} H_N.$$

Proof. It suffices to recognize that Proposition 5 can be restated in many equivalent forms as follows. From (22), $J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}^0)$ if and only if

$$\lim_{N\to\infty} \frac{1}{N} \|A_N\mathbf{x}_N - B_N\mathbf{u}_N\|^2_{(A_N A_N^T)^{-1}} = 0,$$

or, equivalently,

$$\lim_{N\to\infty} \frac{1}{N} \|A_N\mathbf{x}_N - B_N\mathbf{u}_N\|^2 = 0$$

by (29), or, equivalently,

$$\lim_{N\to\infty}\frac{1}{N}\|H_N\boldsymbol{\theta}-\mathbf{x}_N\|^2=0$$

by the equivalence of (1.A) and (1.B). Therefore $J(\boldsymbol{\theta})=J(\boldsymbol{\theta}^0)$ implies $\boldsymbol{\theta}=\boldsymbol{\theta}^0$ if and only if $\lim_{N\to\infty}H_N^{\mathrm{T}}H_N/N>0$, or $\lim_{N\to\infty}M_N/N=(1/\sigma^2)\lim_{N\to\infty}H_N^{\mathrm{T}}(A_NA_N^{\mathrm{T}})^{-1}H_N/N>0$ by (29).

Theorem 1 can be viewed as a stochastic version of Proposition 2, and Corollary 1.1 corresponds to Proposition 1. The matrix M_N is actually the information matrix. The derivation is implicit in the proof of Proposition 6. See [9] for details.

Note that the a.s. convergence of $J_N(\boldsymbol{\theta})/N$ does not require any conditions other than the basic assumptions of bounded-input and stability. Since $p\text{-}\lim_{N\to\infty}J_N(\boldsymbol{\theta})/N=J(\boldsymbol{\theta})$ if $J(\boldsymbol{\theta})$ exists, either condition (30) or (30a) or (30b) is necessary and sufficient for consistency, i.e., for $\hat{\boldsymbol{\theta}}_N$ to converge to $\boldsymbol{\theta}^0$ in probability.

PROPOSITION 6. *If the conditions of Theorem* 1 *are satisfied, then* $\hat{\boldsymbol{\theta}}_N$ *is asymptotically normal, with covariance matrix* $M(\boldsymbol{\theta}^0)^{-1}$ *where* $M(\boldsymbol{\theta}^0)=\lim_{N\to\infty}M_N(\boldsymbol{\theta}^0)$.

Proof. By (20), $J_N(\boldsymbol{\theta})/N=\|\mathbf{y}_N-A_N^{-1}B_N\mathbf{u}_N\|^2/N$, with the term $\tilde{\mathbf{z}}_N$ omitted since it does not affect the value of $J_N(\boldsymbol{\theta})/N$ as $N\to\infty$. Define

$$\phi_N(\boldsymbol{\theta})=\nabla_{\boldsymbol{\theta}}J_N(\boldsymbol{\theta})/2N\sigma^2. \tag{37}$$

For all N,

$$0=\phi_N(\hat{\boldsymbol{\theta}}_N)=\phi_N(\boldsymbol{\theta}^0)+\nabla_{\boldsymbol{\theta}}\phi_N(\boldsymbol{\theta}_N^*)(\hat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}^0), \tag{38}$$

where $\|\boldsymbol{\theta}_N^*-\boldsymbol{\theta}^0\|\leqq\|\hat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}^0\|$. $\phi_N(\boldsymbol{\theta}_0)$ is normal since by direct calculation,

$$\frac{\partial}{\partial a_j}\frac{1}{2N\sigma^2}J_N(\boldsymbol{\theta}^0)=\langle\boldsymbol{\eta}_N,A_N^{0-1}S^j\mathbf{x}_N\rangle/N\sigma^2,$$

$$\frac{\partial}{\partial b_j}\frac{1}{2N\sigma^2}J_N(\boldsymbol{\theta}^0)=-\langle\boldsymbol{\eta}_N,A_N^{0-1}S^j\mathbf{u}_N\rangle/N\sigma^2\qquad\text{for } N \text{ large}.$$

Thus

$$\phi_N(\boldsymbol{\theta}^0)\sim\text{Normal}\left(0;\frac{1}{N^2}M_N(\boldsymbol{\theta}^0)\right), \tag{39}$$

where

$$M_N=\frac{1}{\sigma^2}H_N^{\mathrm{T}}A_N^{0\mathrm{T}-1}A_N^{0-1}H_N\qquad(k\times k).$$

By Theorem 1 and Corollary 1.1, $M_N>0$ for all N.

$\nabla_{\boldsymbol{\theta}}\phi_N(\boldsymbol{\theta}_N)$ can also be calculated directly.

$$\frac{\partial^2}{\partial a_i\partial a_j}\frac{1}{2N\sigma^2}J_N(\boldsymbol{\theta})=-2\langle\boldsymbol{\eta}_N+\mathbf{x}_N-A_N^{-1}B_N\mathbf{u}_N,A_N^{-3}B_NS^{i+j}\mathbf{u}_N\rangle/N\sigma^2$$
$$+\langle A_N^{-2}B_NS^i\mathbf{u}_N,A_N^{-2}B_NS^j\mathbf{u}_N\rangle/N\sigma^2,$$

$$\frac{\partial^2}{\partial a_i \partial b_j}\frac{1}{2N\sigma^2}J_N(\boldsymbol{\theta}) = \langle \boldsymbol{\eta}_N + \mathbf{x}_N - A_N^{-1}B_N\mathbf{u}_N, A_N^{-2}S^{i+j}\mathbf{u}_N\rangle/N\sigma^2$$

$$- \langle A_N^{-2}B_N S^i\mathbf{u}_N, A_N^{-1}S^j\mathbf{u}_N\rangle/N\sigma^2,$$

$$\frac{\partial^2}{\partial b_i \partial b_j}\frac{1}{2N\sigma^2}J_N(\boldsymbol{\theta}) = \langle A_N^{-1}S^i\mathbf{u}_N, A_N^{-1}S^j\mathbf{u}_N\rangle/N\sigma^2 \qquad \text{for } N \text{ large.}$$

Since $\hat{\boldsymbol{\theta}}_N \overset{\text{a.s.}}{\to} \boldsymbol{\theta}^0$ by Theorem 1, $\boldsymbol{\theta}_N^* \overset{\text{a.s.}}{\to} \boldsymbol{\theta}^0$. The law of large numbers implies

$$\nabla_{\boldsymbol{\theta}}\phi_N(\boldsymbol{\theta}_N^*) \overset{\text{a.s.}}{\to} E\nabla_{\boldsymbol{\theta}}\phi_N(\boldsymbol{\theta}^0) = \lim_{N\to\infty}\frac{1}{N}M_N(\boldsymbol{\theta}^0). \tag{40}$$

Combining (38)–(40), we conclude that $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^0$ converges with probability one to $-\lim_{N\to\infty} NM_N^{-1}(\boldsymbol{\theta}^0)\phi_N(\boldsymbol{\theta}^0)$ which is normal, with zero mean and covariance $\lim_{N\to\infty} M_N(\boldsymbol{\theta}^0)^{-1}$ since

$$\lim_{N\to\infty} EN^2M_N^{-1}(\boldsymbol{\theta}^0)\phi_N(\boldsymbol{\theta}^0)\phi_N^{\mathsf{T}}(\boldsymbol{\theta}^0)M_N^{-1}(\boldsymbol{\theta}^0) = \lim_{N\to\infty} M_N^{-1}(\boldsymbol{\theta}^0).$$

THEOREM 2. *Given the system* (1/1.A/1.B) *with* $\mathbf{b} \neq \mathbf{0}$ *and* $\{\Phi, \mathbf{b}\}$ *completely controllable,* $\hat{\boldsymbol{\theta}}_N$ *converges to* $\boldsymbol{\theta}^0$ *in mean square if and only if condition* (30)/(30a)/(30b) *is satisfied.*

Proof. As noted before, the three conditions (30), (30a) and (30b) are equivalent.

Sufficiency. By Proposition 6, $E(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^0)(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^0)^{\mathsf{T}} \to \lim_{N\to\infty} M_N^{-1}(\boldsymbol{\theta}^0)$. Equation (30b) implies that $\lim_{N\to\infty} M_N(\boldsymbol{\theta}^0)/N > 0$. Thus, all the eigenvalues of $M_N(\boldsymbol{\theta}^0)/N$ are bounded from below by some positive number ρ for large N, and $\operatorname{tr} M_N^{-1}(\boldsymbol{\theta}^0) \leqq 2k/(\rho N)$. As $N \to \infty$, $\operatorname{tr}\lim_{N\to\infty} M_N^{-1}(\boldsymbol{\theta}^0) = 0$, $E\|\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^0\|^2 \to 0$.

Necessity. Since (30)/(30a)/(30b) is necessary for the consistency of $\hat{\boldsymbol{\theta}}_N$, it is also necessary for convergence in mean square.

5. Convergence of the initial state estimate. We have seen that the initial state does not affect the convergence of the parameter estimates, $\hat{\boldsymbol{\theta}}_N$. The initial state estimate itself, however, is given uniquely by (18) and (21) as a function of $\hat{\boldsymbol{\theta}}_N$ and $\mathbf{y}_N$:

$$\hat{\mathbf{z}}_{0N} = (E_N^{\mathsf{T}}\hat{A}_N^{\mathsf{T}-1}\hat{A}_N^{-1}E_N)^{-1}E_N^{\mathsf{T}}\hat{A}_N^{\mathsf{T}-1}(\mathbf{y}_N - \hat{A}_N^{-1}\hat{B}_N\mathbf{u}_N). \tag{41}$$

The asymptotic properties of $\hat{\mathbf{z}}_{0N}$ depend heavily upon $\hat{\boldsymbol{\theta}}_N$ and are much weaker than those of $\hat{\boldsymbol{\theta}}_N$.

THEOREM 3. *If the conditions of Theorem* 1 *are satisfied, then* $E\hat{\mathbf{z}}_{0N} \to \mathbf{z}_0$ *as* $N \to \infty$ *and the covariance of* $\hat{\mathbf{z}}_{0N}$ *converges to* $R_0 = \lim_{N\to\infty}(E_N^{\mathsf{T}}A_N^{0\mathsf{T}-1}A_N^{0-1}E_N)^{-1}\sigma^2$, *which is finite,* $R_0 > 0$.

Proof. Let

$$\tilde{\mathbf{z}}_K(N) = F_K(N)(\mathbf{y}_K - \hat{A}_{K,N}^{-1}\hat{B}_{K,N}\mathbf{u}_K), \tag{41a}$$

where

$$F_K(N) = (E_K^{\mathsf{T}}\hat{A}_{K,N}^{\mathsf{T}-1}\hat{A}_{K,N}^{-1}E_K)^{-1}E_K^{\mathsf{T}}\hat{A}_{K,N}^{\mathsf{T}-1} \qquad (k \times K),$$

$$\hat{A}_{K,N} = A_K(\hat{\boldsymbol{\theta}}_N) \qquad (K \times K), \tag{42}$$

$$\hat{B}_{K,N} = B_K(\hat{\boldsymbol{\theta}}_N) \qquad (K \times K).$$

E_K is defined analogously to E_N of (4). For each K,

$$\tilde{\mathbf{z}}_K(N) = F_K(N)(\mathbf{x}_K - \hat{A}_{K,N}^{-1}\hat{B}_{K,N}\mathbf{u}_K) + F_K(N)\boldsymbol{\eta}_K. \tag{43}$$

If $E\tilde{\mathbf{z}}_K(N)$ converges to $\tilde{\mathbf{z}}_K$ uniformly in K as $N \to \infty$, then $E\hat{\mathbf{z}}_{0N} \to \lim_{K\to\infty} \tilde{\mathbf{z}}_K$ as $K \to \infty$, by (41), (41a). Since $\hat{\boldsymbol{\theta}}_N \overset{\text{a.s.}}{\to} \boldsymbol{\theta}^0$ as $N \to \infty$, we note that for every K,

$$\begin{aligned} &\hat{A}_{K,N} \overset{\text{a.s.}}{\to} A_K^0, \qquad \hat{B}_{K,N} \overset{\text{a.s.}}{\to} B_K^0, \\ &F_K(N) \overset{\text{a.s.}}{\to} F_K = (E_K^{\mathsf{T}} A_K^{0\mathsf{T}-1} A_K^{0-1} E_K)^{-1} E_K^{\mathsf{T}} A_K^{0\mathsf{T}-1}. \end{aligned} \tag{44}$$

Substituting (44) into (43), we obtain for every K, as $N \to \infty$,

$$\begin{aligned} \tilde{\mathbf{z}}_K(N) &\overset{\text{a.s.}}{\to} F_K(\mathbf{x}_K - A_K^{0-1}B_K^0\mathbf{u}_K) + F_K\boldsymbol{\eta}_K \\ &= F_K A_K^{0-1} E_K \mathbf{z}_0 + F_K\boldsymbol{\eta}_K \\ &= \mathbf{z}_0 + F_K\boldsymbol{\eta}_K \qquad \text{by (44).} \end{aligned}$$

Since

$$\tilde{\mathbf{z}}_K(N) - \mathbf{z}_0 \overset{\text{a.s.}}{\to} F_K\boldsymbol{\eta}_K \tag{45a}$$

and

$$\int F_K\boldsymbol{\eta}_K \, dP(\boldsymbol{\eta}_K) = \mathbf{0}, \tag{45b}$$

by the Lebesgue convergence theorem,

$$\begin{aligned} \lim_{N\to\infty} E(\tilde{\mathbf{z}}_K(N) - \mathbf{z}_0) &= \lim_{N\to\infty} \int (\tilde{\mathbf{z}}_K(N) - \mathbf{z}_0)\, dP(\tilde{\mathbf{z}}_K(N)) \\ &= \int F_K\boldsymbol{\eta}_K \, dP(\boldsymbol{\eta}_K) = \mathbf{0}; \end{aligned}$$

hence, $\lim_{K\to\infty} E\hat{\mathbf{z}}_{0K} = \lim_{K\to\infty} \lim_{N\to\infty} E\tilde{\mathbf{z}}_K(N) = \mathbf{z}_0$. Also,

$$\begin{aligned} E(\tilde{\mathbf{z}}_K(N) - \mathbf{z}_0)(\tilde{\mathbf{z}}_K(N) - \mathbf{z}_0)^{\mathsf{T}} &\to \int F_K\boldsymbol{\eta}_K\boldsymbol{\eta}_K^{\mathsf{T}}F_K^{\mathsf{T}}\, dP(\boldsymbol{\eta}_K) \\ &= F_K F_K^{\mathsf{T}}\sigma^2 \qquad \text{uniformly in } K; \end{aligned}$$

hence

$$\begin{aligned} \lim_{K\to\infty} E(\hat{\mathbf{z}}_{0K} - \mathbf{z}_0)(\hat{\mathbf{z}}_{0K} - \mathbf{z}_0)^{\mathsf{T}} &= \lim_{K\to\infty} F_K F_K^{\mathsf{T}}\sigma^2 \\ &= \lim_{K\to\infty} (E_K^{\mathsf{T}} A_K^{0\mathsf{T}-1} A_K^{0-1} E_K)^{-1}\sigma^2 = R_0. \end{aligned}$$

By the result of Appendix B,

$$\sigma^2\left(1 + \sum_{i=1}^{\infty} |g_i|\right)^{-2} I_k \leqq R_0 \leqq \sigma^2\left(1 + \sum_{i=1}^{k} |a_i|\right)^2 I_k.$$

6. Generalization and discussion. The above analysis extends easily to systems containing random disturbance with rational spectrum, i.e., systems which are

representable by the following Gauss–Markov model:

$$\begin{aligned} \mathbf{z}(t+1) &= \Phi\mathbf{z}(t) + \mathbf{b}u_t + \mathbf{d}\xi_t, \\ x_t &= \langle \mathbf{h}, \mathbf{z}(t) \rangle + b_0 u_t + \xi_t, \end{aligned} \tag{46}$$

where $\{\xi_t\}$ is Gaussian white noise identically distributed as $N(0; \lambda^2)$; the input sequence $\{u_t\}$ is given and uniformly bounded; $\mathbf{d} = (d_1\, d_2, \cdots, d_k)^{\mathsf{T}}$; $\Phi, \mathbf{h}, \mathbf{b}$ are defined as before. Thus, the unknown system parameters are $\mathbf{a}$, $\mathbf{b}$, $\mathbf{d}$, b_0, λ^2 and the input-output relation is

$$x_t + \sum_{i=1}^{k} a_i x_{t-i} = b_0 u_t + \sum_{i=1}^{k} b_i' u_{t-i} + \xi_t + \sum_{i=1}^{k} d_i' \xi_{t-i}.$$

The output is observed with additive noise:

$$y_t = x_t + \eta_t, \tag{47}$$

where $\{\eta_t\}$ is again assumed for simplicity to be independent and identically distributed as $N(0; 1)$, ξ_t and η_t are independent. In Toeplitz matrix notation, (46) and (47) become

$$A_N \mathbf{x}_N = B_N \mathbf{u}_N + D_N \boldsymbol{\xi}_N + E_N \mathbf{z}_0, \tag{46.A}$$

$$\mathbf{y}_N = \mathbf{x}_N + \boldsymbol{\eta}_N, \tag{47.A}$$

where

$$A_N = I_N + \sum_{i=1}^{k} a_i S^i,$$

$$B_N = b_0 I_N + \sum_{i=1}^{k} b_i' S^i,$$

$$D_N = I_N + \sum_{i=1}^{k} d_i' S^i.$$

When (46.A) and (47.A) are combined,

$$A_N \mathbf{y}_N = B_N \mathbf{u}_N + D_N \boldsymbol{\xi}_N + A_N \boldsymbol{\eta}_N + E_N \mathbf{z}_0. \tag{48}$$

Clearly, the combined model of Astrom is the special case of (48) with $\eta_i = 0$ for all i. Furthermore, if we attempt to replace (48) by a model with only one single noise source, for instance, by the proper canonical representation of Lévy (see, e.g., [13], [14]), then (48) can be rewritten as

$$A_N \mathbf{y}_N = B_N \mathbf{u}_N + W_N \mathbf{e}_N + E_N \tilde{\mathbf{z}}_0 \tag{49.A}$$

with some W_N defined analogously to D_N in (46.A), which corresponds to

$$\begin{aligned} \tilde{\mathbf{z}}(t+1) &= \Phi\tilde{\mathbf{z}}(t) + \mathbf{b}u_t + \mathbf{w}e_t, \\ y_t &= \langle \mathbf{h}, \tilde{\mathbf{z}}(t) \rangle + b_0 u_t + e_t, \end{aligned} \tag{49}$$

where $\{e_t\}$ is a white noise sequence known as the innovation process. In appearance, (49.A) resembles the form of Astrom's model. Note, however, $\mathbf{w} = (w_1,$

$w_2, \cdots, w_k)^T$ that appears in W_N is no longer constant, nor is the variance of e_t. In fact, the variance of e_t is given by a Ricatti equation with the unknown constants **a**, **b**, **d** as parameters, and is thus time varying and unknown. Clearly, this model is inconvenient to use in comparison with the two-noise model of (48). Even in the simple case of no plant noise wherefore $A_N = W_N$, this should be explicitly incorporated into the likelihood function; consequently **a** and **w** should not be treated as distinct quantities as in [7].

We now demonstrate that by using techniques similar to the previous sections, the various convergence questions can be studied for the general case. Only the results for convergence with probability one are presented since the other generalizations can be made in like manner.

$$p(\mathbf{y}_N|\mathbf{a}, \mathbf{b}, \mathbf{d}, \lambda, \mathbf{z}_0) = \frac{1}{2\pi}|\Gamma_N|^{-1/2} \exp -\tfrac{1}{2}\|\mathbf{y}_N - A_N^{-1}B_N\mathbf{u}_N - A_N^{-1}E_N\mathbf{z}_0\|^2_{\Gamma_N^{-1}},$$

where

$$\Gamma_N = \sigma^2 I_N + \lambda^2 A_N^{-1} D_N D_N^T A_N^{-1T}.$$

Astrom treated the case where $\eta_i = 0$. Thus, $\Gamma_N = \lambda^2 A_N^{-1} D_N D_N^T A_N^{-1T}$ and the determinant reduces simply to λ^2. The analysis of our previous sections treats the case $\xi_i = 0$. Thus, $\Gamma_N = I_N$ and the determinant reduces to 1. In the general case, the m.l.e. is obtained by minimizing J_N, where

$$J_N = \ln|\Gamma_N| + \|\mathbf{y}_N - A_N^{-1}B_N\mathbf{u}_N - A_N^{-1}E_N\mathbf{z}_0\|^2_{\Gamma_N^{-1}}.$$

Again, assume that $\boldsymbol{\theta} = (\mathbf{a}^0, \mathbf{b}^0, \mathbf{d}^0, \lambda^0)$ must lie in the interior of a given compact subset $\textcircled{H}_S$ of R_{3k+2} and that $A^0(z)$ and $D^0(z)$ are known to be stable.

As for computing the m.l.e., there are many existing algorithms for both unconstrained and constrained minimizations of $J_N(\boldsymbol{\theta})$. There are various versions of small and large step gradient methods and the second order methods such as the Newton-Raphson with possibly some modification to prevent divergence or the Fletcher-Powell-Davidon algorithm for unconstrained minimization and various versions of feasible direction methods, for example, for the constrained case where the parameter set is defined implicitly by a set of inequality and/or equality constraints. See, for example, [16] and [17] for further detail.

$$\lim_{N\to\infty} \frac{1}{N} J_N = J_1(\mathbf{a}, \mathbf{b}, \mathbf{d}, \lambda^2) + J_2(\mathbf{a}, \mathbf{d}, \lambda^2), \quad \text{a.s.},$$

where

$$J_1 = \lim_{N\to\infty} \frac{1}{N} \|(A_N B_N^0 - A_N^0 B_N)\mathbf{u}_N\|^2_{(A_N A_N^0 \Gamma_N A_N^{0T} A_N^T)^{-1}},$$

$$J_2 = \lim_{N\to\infty} \frac{1}{N} \{\ln \Gamma_N + \operatorname{tr} \Gamma_N^{-1}\Gamma_N^0\}.$$

By Proposition 4, $\hat{\boldsymbol{\theta}}_N$ converges with probability one to $\boldsymbol{\theta}^0$ if and only if the set

$\textcircled{H}^0 \cap \textcircled{H}_S$ is a singleton, where

$$\begin{aligned}\textcircled{H}^0 &= \{\boldsymbol{\theta} | J_1(\boldsymbol{\theta}) + J_2(\boldsymbol{\theta}) = J_1(\boldsymbol{\theta}^0) + J_2(\boldsymbol{\theta}^0)\} \\ &= \{\boldsymbol{\theta} | J_1(\boldsymbol{\theta}) = J_1(\boldsymbol{\theta}^0) \text{ and } J_2(\boldsymbol{\theta}) = J_2(\boldsymbol{\theta}^0)\}\end{aligned}$$

since $J_1 + J_2 \geqq J(\boldsymbol{\theta}^0) + J_2(\boldsymbol{\theta}^0)$ by definition of the m.l.e. and $J_2 \geqq J_2(\boldsymbol{\theta}^0) = 0$.

To establish results similar to Theorem 1, we need the following lemma.

LEMMA 1.

$$\frac{1}{N} \ln |\Gamma_N| + \frac{1}{N} \operatorname{tr} \Gamma_N^{-1} \Gamma_N^0 \geqq \frac{1}{N} \ln |\Gamma_N^0| + 1. \tag{50}$$

Equality is achieved if and only if $D(z)/A(z) = D^0(z)/A^0(z)$, $\lambda^2 = \lambda^{02}$.

Proof. The matrix $\sigma^2 I_N + \lambda^2 A_N^{0-1} D_N^0 D_N^{0T} A_N^{0-1T}$ is positive definite and is therefore equal to the product $R_N R_N^T$ where R_N is a nonsingular matrix.

Let $Q_N = R_N^T \Gamma_N^{-1} R_N$ and let $q_1, q_2, \cdots, q_N$ be the eigenvalues of Q_N. Then, $q_i > 0$ since Q_N is positive definite.

$$\ln |\Gamma_N^{-1} \Gamma_N^0| = \ln |Q_N| = \ln \prod_{i=1}^N q_i,$$

$$\operatorname{tr} (\Gamma_N^{-1} \Gamma_N^0) = \operatorname{tr} Q_N = \sum_{i=1}^N q_i.$$

Since $\exp (q_i - 1) \geqq q_i$ for $q_i > 0$,

$$\exp \sum_{i=1}^N (q_1 - 1) \geqq \prod_{i=1}^N q_i,$$

$$\ln |\Gamma_N^{-1} \Gamma_N^0| = \ln \prod_{i=1}^N q_i \leqq \sum_{i=1}^N (q_i - 1) = \operatorname{tr} (\Gamma_N^{-1} \Gamma_N^0) - N$$

from which (50) immediately follows. Equality is achieved if and only if $q_i = 1$ for all i; or, equivalently, $Q_N = I_N$, $\Gamma_N = \Gamma_N^0$ and hence $A_N^{-1} D_N = A_N^{0-1} D_N^0$, $\lambda^2 = \lambda^{02}$.

In situations where it is known a priori that the random disturbance is not confined to any invariant subspace of Φ^0, the plant noise actually has the effect of persistently exciting all the system modes. Consequently, a milder set of conditions on the input is required as seen in the following theorem.

THEOREM 4. *Given the system* (46) *with* $\mathbf{d} \neq \mathbf{0}$ *and* $\{\Phi, \mathbf{d}\}$ *completely controllable,* $\hat{\boldsymbol{\theta}}_N$ *converges with probability one to* $\boldsymbol{\theta}^0$ *if and only if*

$$V = \lim_{N \to \infty} \frac{1}{N} \tilde{U}_{N,k}^T \tilde{U}_{N,k} > 0,$$

where $\tilde{U}_N = (\mathbf{u}_N, S\mathbf{u}_N, \cdots, S^k \mathbf{u}_N)$.

Proof. $J_2 = J_2(\boldsymbol{\theta}^0) = 1 + \lim_{N \to \infty} \ln |\Gamma_N^0|/N$ if and only if $D(z)/A(z) = D^0(z)/A^0(z)$, $\lambda^2 = \lambda^{02}$, by Lemma 1; or, equivalently, $\mathbf{d} = \mathbf{d}^0$, $\mathbf{a} = \mathbf{a}^0$, $\lambda^2 = \lambda^{02}$, by controllability. $J_1 = J_1(\boldsymbol{\theta}^0) = 0$ if and only if $0 = \lim_{N \to \infty} \|(B_N^0 - B_N)\mathbf{u}_N\|^2/N$

by $\mathbf{a} = \mathbf{a}^0$ and by arguments similar to Appendix B; or

$$0 = (b_0^0 - b_0, \mathbf{b}^{0\mathsf{T}} - \mathbf{b}^{\mathsf{T}}) V \begin{pmatrix} b_0^0 - b_0 \\ \mathbf{b}_0 - \mathbf{b} \end{pmatrix}.$$

Therefore, $b_0 = b_0^0$ and $\mathbf{b} = \mathbf{b}^0$ if and only if $V > 0$. This completes the proof.

Note that if $\mathbf{b}^0 = \mathbf{0}$, i.e., without deterministic input, $\boldsymbol{\theta} = (\mathbf{a}, \mathbf{d}, \lambda)$, we need only the controllability of $\{\Phi, \mathbf{d}\}$ since the condition on $\{u_t\}$ is only required for the convergence of the estimates of $\mathbf{b}$ and b_0.

Lastly, when $\{\Phi, \mathbf{d}\}$ is not known to be controllable, we have the following theorem.

THEOREM 4a. *Given the system* (46) *such that* $\{\Phi, (\mathbf{b}, \mathbf{d})\}$ *is completely controllable,* $\hat{\boldsymbol{\theta}}_N$ *converges with probability one if*

$$\lim_{N \to \infty} \frac{1}{N} \tilde{U}_{N,2k}^{\mathsf{T}} \tilde{U}_{N,2k} > 0.$$

Proof. The proof invokes Lemma 1 to guarantee that $\lambda^2 = \lambda^{02}$, $D(z)/A(z) = D^0(z)/A^0(z)$ and uses the same techniques as in Theorem 1 to guarantee that $B(z)/A(z) = B^0(z)/A^0(z)$. The details can be found in [15] and are omitted here. The assumption of this theorem is a condition to ensure that the model in (46) is a minimal realization of the joint distribution of $\{x_t\}$.

Conclusion. The above discussions have exhibited the significance of a priori knowledge of the structural properties of a system's model on the convergence of the estimates. Under various assumptions, necessary and sufficient conditions have been derived for the maximum likelihood estimates to converge with probability one. Sufficient indications have also been given to show the mean-square convergence of the system-parameter estimates, to show the asymptotic unbiasedness of the initial state estimate and to obtain upper and lower bounds for the initial estimate covariance.

Appendix A.

Proof of Proposition 4 (Wald-Kendall-Astrom [7]). It suffices to show that for every $\hat{\boldsymbol{\theta}}_N$ defined by (19), the distance of $\hat{\boldsymbol{\theta}}_N$ from $\textcircled{H}^0$ tends to zero; that is,

$$\inf_{\boldsymbol{\theta} \in \textcircled{H}^0 \cap \textcircled{H}_S} \|\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}\| \to 0$$

with probability one as $N \to \infty$. Define the set $\textcircled{H}^\varepsilon$ as:

$$\textcircled{H}^\varepsilon = \bigcup_{\tilde{\boldsymbol{\theta}} \in \textcircled{H}^0} \left\{ \boldsymbol{\theta} \,\middle|\, \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\| < \varepsilon \right\},$$

and denote its complement by $\overline{\textcircled{H}^\varepsilon}$.

Clearly, $\overline{\textcircled{H}^\varepsilon} \cap \textcircled{H}_S$ is compact and $\textcircled{H}^\varepsilon \cap \textcircled{H}_S \supset \textcircled{H}^0 \cap \textcircled{H}_S$ if $\varepsilon > 0$. Let $\boldsymbol{\theta}^0$ be any vector in $\textcircled{H}_S$ which satisfies (1.A); hence $\boldsymbol{\theta}^0 \in \textcircled{H}^0 \cap \textcircled{H}_S$. For any $\varepsilon > 0$, if $\boldsymbol{\theta}^* \in \overline{\textcircled{H}^\varepsilon} \cap \textcircled{H}_S$, then $\boldsymbol{\theta}^* \notin \textcircled{H}^0 \cap \textcircled{H}_S$ and there exists $\delta(\boldsymbol{\theta}^*, \boldsymbol{\theta}^0) > 0$ such that

$$|J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}^0)| = \delta. \tag{A.1}$$

Since $J_N(\boldsymbol{\theta})/N \overset{\text{a.s.}}{\to} J(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \textcircled{H}_S$, there exists $N_0(\delta)$ such that for all $N \geqq N_0$,

$$\text{(A.2)} \qquad \begin{aligned} \left| \frac{1}{N} J_N(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}^*) \right| &< \frac{\delta}{2}, \quad \text{a.s.}, \\ \left| \frac{1}{N} J_N(\boldsymbol{\theta}^0) - J(\boldsymbol{\theta}^0) \right| &< \frac{\delta}{2}, \quad \text{a.s.}, \end{aligned}$$

(A.1) and (A.2) imply that

$$\left| \frac{1}{N} J_N(\boldsymbol{\theta}^*) - \frac{1}{N} J_N(\boldsymbol{\theta}^0) - \{J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}^0)\} \right| < |J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}^0)|,$$

and $J_N(\boldsymbol{\theta}^0) \neq J_N(\boldsymbol{\theta}^*)$, a.s. for all $N \geqq N_0$.

$$\text{(A.3)} \qquad \begin{aligned} E_{\mathbf{y}_N|\boldsymbol{\theta}^0}[J_N(\boldsymbol{\theta}^0) - J_N(\boldsymbol{\theta}^*)] &= E_{\mathbf{y}_N|\boldsymbol{\theta}^0} \log \frac{p(\mathbf{y}_N|\theta^*)}{p(\mathbf{y}_N|\boldsymbol{\theta}^0)} \\ &< \log E_{\mathbf{y}_N|\boldsymbol{\theta}^0} \frac{p(\mathbf{y}_N|\theta^*)}{p(\mathbf{y}_N|\boldsymbol{\theta}^0)} = 0 \end{aligned}$$

by Jensen's inequality. Since $J_N(\boldsymbol{\theta})/N \overset{\text{a.s.}}{\to} EJ_N(\boldsymbol{\theta})/N$ for all $\boldsymbol{\theta} \in \textcircled{H}_S$, if $\lim EJ_N(\boldsymbol{\theta})/N < \infty$ then $\lim_{N\to\infty} \{J_N(\boldsymbol{\theta}^0)/N - J_N(\boldsymbol{\theta}^*)/N\} < 0$, a.s. The same is true if $EJ_N(\boldsymbol{\theta}^*)/N \to \infty$ since $EJ_N(\boldsymbol{\theta}^0)/N = \sigma^2 < \infty$. Hence, there exists $N_1(\boldsymbol{\theta}^*, \boldsymbol{\theta}^0)$ such that

$$J_N(\boldsymbol{\theta}^0) < J_N(\boldsymbol{\theta}^*), \quad \text{a.s.} \quad \text{for all } N \geqq N_1(\boldsymbol{\theta}^*, \boldsymbol{\theta}^0).$$

By the compactness of $\textcircled{H}^0 \cap \textcircled{H}_S$ and $\overline{\textcircled{H}^\varepsilon} \cap \textcircled{H}_S$, and the continuity of $J_N(\boldsymbol{\theta})$, there exists N_ε which depends only on ε such that for all $N \geqq N_\varepsilon$,

$$\max_{\boldsymbol{\theta}^0 \in \textcircled{H}^0 \cap \textcircled{H}_S} J_N(\boldsymbol{\theta}^0) < \min_{\boldsymbol{\theta}^* \in \overline{\textcircled{H}^\varepsilon} \cap \textcircled{H}_S} J_N(\boldsymbol{\theta}^*), \quad \text{a.s.}$$

But, $J_N(\boldsymbol{\theta}^0) \geqq J_N(\hat{\boldsymbol{\theta}}_N)$. Therefore, for every $\varepsilon > 0$, $\hat{\boldsymbol{\theta}}_N \notin \overline{\textcircled{H}^\varepsilon} \cap \textcircled{H}_S$ or $\hat{\boldsymbol{\theta}}_N \in \textcircled{H}^\varepsilon \cap \textcircled{H}_S$, a.s.; and by the definition of $\textcircled{H}^\varepsilon$,

$$\inf_{\boldsymbol{\theta} \in \textcircled{H}^0 \cap \textcircled{H}_S} \|\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}\| < \varepsilon, \quad \text{a.s.} \quad \text{for all } N \geqq N_\varepsilon.$$

Appendix B. We give bounds on $A_N A_N^\top$, where

$$A_N = I_N + \sum_{i=1}^{k} a_i S^i.$$

For general systems, $A_N A_N^\top \leqq (1 + \sum_{i=1}^{k} |a_i|)^2 I_N$. For bounded input–bounded output systems, $A_N^{-1} = g_0 I_N + \sum_{i=1}^{N-1} g_i S^i$ such that $\sum_{i=0}^{\infty} |g_i| < \infty$. Hence,

$$(A_N A_N^\top)^{-1} = A_N^{-1\top} A_N^{-1} \leqq \left(\sum_{i=0}^{N-1} |g_i| \right)^2 I_N \leqq \left(\sum_{i=0}^{\infty} |g_i| \right)^2 I_N$$

and

$$A_N A_N^\top \geqq \rho_1 I_N,$$

where $\rho_1 = 1/(\sum_{i=0}^{\infty} |g_i|)^2$.

REFERENCES

[1] A. WALD, *Asymptotic properties of the maximum-likelihood estimate of an unknown parameter of a discrete stochastic process*, Ann. Math. Statist., 19 (1948), pp. 40–46.

[2] M. LEVIN, *Estimation of system pulse transfer function in the presence of noise*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 229–235.

[3] T. J. KOOPMANS, *Linear Regression Analysis of Economic Time Series*, D. F. Bohn, N. V. Haarlem, The Netherlands, 1937.

[4] M. AOKI AND P. C. YUE, *On a priori estimates of some identification methods*, to appear.

[5] F. W. SMITH AND W. B. HILTON, *Monte Carlo evaluation of methods for pulse transfer function estimation*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 568–576.

[6] A. E. ROGERS, *Maximum likelihood estimation of dynamic system parameters*, Doctoral thesis, Department of Electrical Engineering, Princeton University, Princeton, 1968.

[7] K. T. ASTROM, T. BOHLIN AND S. WENSMARK, *Automatic construction of linear stochastic dynamic models for stationary industrial processes with random disturbances using operating records*, Rep. TP18.150, IBM Nordic Laboratory, Lidingö, Sweden, 1965.

[8] A. WALD, *Note on the consistency of the maximum likelihood estimate*, Ann. Math. Statist., 22 (1949), pp. 595–601.

[9] R. M. STALEY, *Input signal synthesis in identification problems*, Doctoral thesis, Department of Engineering, University of California, Los Angeles, 1969.

[10] M. AOKI AND R. M. STALEY, *Input signal synthesis in parameter identification*, presented at the 4th IFAC Congress, Warsaw, Poland, 1969.

[11] B. L. HO, *An effective construction of realizations from input/output descriptions*, Doctoral thesis, Stanford University, Stanford, Calif., 1966.

[12] L. A. ZADEH AND C. A. DESOER, *Linear Systems Theory, the State-Space Approach*, McGraw-Hill, New York, 1963.

[13] R. GEESEY AND T. KAILATH, *Comments on the relationship of alternate state-space representations in linear filtering problems*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 113–114.

[14] T. KAILATH, *An innovations approach to least-squares estimation. Part I*, Ibid., AC-13 (1968), pp. 646–655.

[15] P. C. YUE, *The identification of constrained systems with high dimension*, Doctoral thesis, University of California, Los Angeles, 1970, forthcoming.

[16] J. ABADIE, *Nonlinear Programming*, North-Holland, Amsterdam, 1967.

[17] T. BOHLIN, *On maximum likelihood method of identification*, Rep. TP 18.191, System Development Division, IBM Nordic Laboratory, Lidingö, Sweden, 1969.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

ON THE CONTROL OF A LINEAR FUNCTIONAL-DIFFERENTIAL EQUATION WITH QUADRATIC COST*

HAROLD J. KUSHNER† AND DANIEL I. BARNEA‡

1. Introduction. Let H be the space of bounded measurable n-vector-valued functions[1] $y(\varphi) = (y_1(\varphi), \cdots, y_n(\varphi))'$ on the finite interval $[-r, 0]$, $r > 0$, whose components are continuous on $[-r, 0]$. Suppose $x(t)$ is an n-vector-valued bounded measurable function defined on the interval $[-r, T]$, $T > 0$. Fix $t \in [0, T]$. Let x_t denote the element of H which takes the value $x(t + \varphi)$ at φ, $\varphi \in [-r, 0]$. Let $x(\cdot)$ be the solution of the delay equation[2]

$$\dot{x}(t) = A(t)x(t) + B(t)x(t - r) + \int_{-r}^{0} C(t, \varphi)x(t + \varphi)\, d\varphi + D(t)u(t), \tag{1}$$

where $A(t)$, $B(t)$, $C(t, \varphi)$, $D(t)$, and the derivatives of $B(t)$ and $C(t, \varphi)$ are given continuous functions of (t, φ) in $[0, T] \times [-r, 0]$, and the initial condition, x_0, is in H.

$u(\cdot)$ is a control which may take several forms. The majority of the paper is devoted to showing that a linear *feedback* form (5) (given by the construction of Theorem 5) is optimal, with cost (2). All the work up to Theorem 6 uses the linear form (5). In Theorem 6, it is shown that the linear form (5), constructed in Theorem 5, is optimal with respect to the class of

(a) linear feedback controls (5), with bounded measurable coefficient matrices,
(b) any control $u(t)$ which is square integrable on $[0, T]$,
(c) any control $u(x, t)$ for which there is a continuous solution $x(t)$ to (1), and for which $u(x(t), t)$ is square integrable on $[0, T]$.

$$V^u(x_t, t) = \int_t^T [x'(s)M(s)x(s) + u'(s)N(s)u(s)]ds, \tag{2}$$

where $M(s)$ and $N(s)$ are continuous,[3] $M(s) \geqq 0$, and $N(s) > 0$ for each s in $[0, T]$. Special forms have been considered by other authors, e.g., Krasovskii [1], [2]; however, that work is quite vague and, in particular, the crucial fact that the relevant Ricatti-like equation has a solution of the proper form or even some solution is not shown. Since the Ricatti equation is a rather complicated coupled set of first order partial differential equations, the questions of existence and uniqueness

* Received by the editors November 7, 1968, and in revised form July 10, 1969.

† Division of Applied Mathematics and Engineering, Brown University, Providence, Rhode Island 02912. The work of this author was supported in part by the National Aeronautics and Space Administration under Grant NGR 40-002-015, in part by the Air Force Office of Scientific Research under Grant AF-AFOSR 693-67 and in part by the National Science Foundation under Grant GK 2788.

‡ Division of Engineering, Brown University, Providence, Rhode Island 02912. The work of this author was supported by the National Science Foundation under Grant GK 2788.

[1] The prime ′ denotes transpose.

[2] Equation (1) is treated for simplicity; it will be obvious that replacing the term $Bx(t - r)$ by $\sum B_i x(t - r_i)$ demands few changes in the development.

[3] $M \geqq 0$, $N > 0$ denote that M is nonnegative definite and N is positive definite.

require some treatment. In fact, this is the most difficult part of the entire problem. Theorems 1 and 2 give representation of $V(x_t, t)$ as a quadratic functional of x_t, when u is in the linear feedback form (5). Theorem 3 proves the smoothness of solutions to certain partial differential equations, and Theorems 4 and 5 contain the basic result on iteration in policy space. Theorem 6 is the final optimization theorem. Unfortunately, as is common with works on functional-differential equations, some of the calculations are somewhat tedious. Although the problem has an intrinsic interest of its own, owing to the appearance of delays in many situations, the authors' interest in it stemmed from an attempt to analyze a problem where $u(t)$ was actually a functional of noise corrupted observations taken on the interval $[t - r, t]$. This was part of an attempt to use the theory of stochastic delay equations to study certain approximations to nonlinear filters, and to stabilize a system when only noise corrupted observations are available. The latter investigation led to the consideration of the problem of the paper. See Barnea [3]. A rather thorough bibliography of time lag control problems appears in Oguztoreli [5].

2. Preliminary lemma.

LEMMA 1. *Let $u = 0$ and let $A(t)$, $B(t)$, $\partial B(t)/\partial t$, $\partial C(t, \Phi)/\partial t$ and $C(t, \varphi)$ be continuous. Then the solution $x(s)$ has the representation, for $s \geqq t$,*

$$x(s) = K(s, t)x(t) + \int_{-r}^{0} \tilde{K}(s, t, \varphi)x(t + \varphi)\, d\varphi, \tag{3}$$

where $K(s, t) = 0$ for $s < t$, $K(t, t) = I$, the identity, and $K(s, t)$ is continuous in (s, t) for $s \geqq t$. For fixed t, it satisfies (1), as a function of s (with $u = 0$). For fixed s, it satisfies (as a function of t) the adjoint of (1) (with $u = 0$) for $t \leqq s$. The terms $\partial K(s, t)/\partial s$ and $\partial K(s, t)/\partial t$ are continuous for $s \geqq t$ except for a finite discontinuity at $s = t + r$. Also

$$\begin{aligned} \tilde{K}(s, t, \varphi) &= K(s, t + r + \varphi)B(t + r + \varphi) \\ &\quad + \int_{-\varphi}^{r} K(s, t + \varphi + \rho)C(t + \varphi + \rho, -\rho)\, d\rho. \end{aligned} \tag{4}$$

(The upper limit r can be replaced by $\min(s - t - \varphi, r)$.) The first term on the right of (4) is zero for $s < t + r + \varphi$, continuous in (s, t, φ) for $s \geqq t + r + \varphi$, and its derivatives with respect to s, t, φ are continuous for $s \geqq t + r + \varphi$, except at $s = t + 2r + \varphi$, where there is a finite discontinuity. The second term of (4) is zero for $s < t$ and is continuous together with its derivatives with respect to[4] *s, t, φ for $T \geqq s \geqq t \geqq 0$, $-r \leqq \varphi \leqq 0$.*

Note. $\tilde{K}(s, t, \varphi) = 0$ for $s < t$. For the computations of Theorem 1, it is convenient to redefine $\tilde{K}(s, t, \varphi)$ for $s < t$ so that (3) gives the solution for $s \geqq t - r$. Then define $\hat{K}(s, t, \varphi) = \tilde{K}(s, t, \varphi)$ for $s \geqq t$ and, for $t - \varphi \leqq s < t$, define the symbol $\int_{-r}^{0} \hat{K}(s, t, \varphi)x(t + \varphi)d\varphi$ to mean $x(s)$; i.e., for $s < t$, $\hat{K}(s, t, \varphi)$ is the Dirac δ-function

[4] By convention, if $s = t + r + \varphi$, the derivative with respect to s is a right-hand derivative, and with respect to t and φ a left-hand derivative; i.e., the limits are taken within the segment $s \geqq t + r + \varphi$.

$\delta(s - (t + \varphi))$. Thus for $s \geqq t - r$,

$$x(s) = K(s, t)x(t) + \int_{-r}^{0} \hat{K}(s, t, \varphi)x(t + \varphi)\, d\varphi. \tag{3'}$$

Proof. The forms (3), (4) and statements concerning $K(s, t)$ follow from Halanay [4, pp. 369–370]. The statements concerning $\hat{K}(s, t, \varphi)$ are straightforward consequences of the properties $K(s, t)$, by virtue of the representation (4).

Remark. In (1) let $u(t)$ take the form

$$u(t) = E_u(t)x(t) + \int_{-r}^{0} F_u(t, \varphi)x(t + \varphi)\, d\varphi. \tag{5}$$

Then

$$\dot{x}(t) = A_u(t)x(t) + B(t)x(t - r) + \int_{-r}^{0} C_u(t, \varphi)x(t + \varphi), \tag{1'}$$

where

$$A_u(t) = A(t) + D(t)E_u(t),$$
$$C_u(t, \varphi) = C(t, \varphi) + D(t)F_u(t, \varphi).$$

Let $D(t)$, $E_u(t)$, $F_u(t, \varphi)$, $\partial D(t)/\partial t$ and $\partial F_u(t, \varphi)/\partial t$ be continuous. Then Lemma 1 remains valid, where we replace K, $\hat{K}$ by K_u, $\hat{K}_u$, respectively, the kernels corresponding to (1').

3. Representations for the cost. By substituting (5) into (2), we obtain

$$\begin{aligned} V^u(x_t, t) = &\int_t^T \{x'(s)M_u(s)x(s)\}\, ds \\ &+ \int_t^T ds \left\{ \int_{-r}^{0} d\varphi x'(s)L_u(s, \varphi)x(s + \varphi) \right\} \\ &+ \int_t^T ds \left\{ \int_{-r}^{0} d\varphi x'(s + \varphi)L_u'(s, \varphi)x(s) \right\} \\ &+ \int_t^T ds \left\{ \int_{-r}^{0} d\varphi \int_{-r}^{0} d\rho x'(s + \varphi)G_u(s, \varphi, \rho)x(s + \rho) \right\} \\ \equiv\ & T_1 + T_2 + T_3 + T_4, \end{aligned} \tag{6}$$

where the T_i are the terms on the right of (6), and

$$\begin{aligned} M_u(s) &= M(s) + E_u'(s)N(s)E_u(s), \\ L_u(s, \varphi) &= E_u'(s)N(s)F_u(s, \varphi), \\ G_u(s, \varphi, \rho) &= F_u'(s, \varphi)N(s)F_u(s, \rho). \end{aligned} \tag{7}$$

THEOREM 1. *Let $u(t)$ take the form* (5), *and assume the conditions of Lemma* 1 *and the remark following it. In addition, let $\partial C(t, \varphi)/\partial\varphi$ and $\partial F_u(t, \varphi)/\partial\varphi$ be continuous and $F_u(t, \varphi)$ and $E_u(t)$ tend to zero as $t \to T$. Let $M(s)$ and $N(s)$ be symmetric and continuously differentiable for $s \in [0, T]$. Then*[5,6]

[5] The S_i, $\hat{S}_i$ are defined as the terms on the right of (8).

[6] If (2) contained a terminal cost term $x'(T)Zx(T)$, then (9), (10), (11) would each contain one additional term (which is not of an integral form). However, we have not been able to show that the additional terms have the smoothness that we will require (i.e., be differentiable).

$$
\begin{aligned}
V^u(x_t, t) &= S_1 + S_2 + \tilde{S}_2 + S_3 \\
&= x'(t)P_u(t)x(t) + x'(t)\int_{-r}^{0} Q_u(t,\varphi)x(t+\varphi)\,d\varphi \\
&\quad + \int_{-r}^{0} x'(t+\varphi)Q_u'(t,\varphi)x(t)\,d\varphi \\
&\quad + \int_{-r}^{0} d\varphi \int_{-r}^{0} d\rho x'(t+\varphi)R_u(t,\varphi,\rho)x(t+\rho).
\end{aligned} \tag{8}
$$

The $P_u(t)$, $Q_u(t,\varphi)$, $R_u(t,\varphi,\rho)$ are sums of the terms in (9), (10), (11), *respectively.*

$$P_{u1}(t) = \int_t^T K_u'(s,t)M_u(s)K_u(s,t)\,ds, \tag{9a}$$

$$P_{u2}(t) = \int_t^T ds \int_{-r}^{0} d\tau K_u'(s,t)L_u(s,\tau)K_u(s+\tau,t), \tag{9b}$$

$$P_{u3}(t) = P_{u2}'(t), \tag{9c}$$

$$P_{u4}(t) = \int_t^T ds \int_{-r}^{0} d\varphi \int_{-r}^{0} d\rho K_u'(s+\varphi,t)G_u(s,\varphi,\rho)K_u(s+\rho,t), \tag{9d}$$

$$
\begin{aligned}
Q_{u1}(t,\varphi) &= \int_t^T K_u'(s,t)M_u(s)\hat{K}_u(s,t,\varphi)\,ds \\
&= \int_t^T K_u'(s,t)M_u(s)\tilde{K}_u(s,t,\varphi),
\end{aligned} \tag{10a}
$$

$$
\begin{aligned}
Q_{u2}(t,\varphi) &= \int_t^T ds \int_{-r}^{0} d\tau K_u'(s,t)L_u(s,\tau)\hat{K}_u(s+\tau,t,\varphi) \\
&= \int_t^{\min[t+r+\varphi,T]} ds K_u'(s,t)L_u(s,t-s+\varphi) \\
&\quad + \int_t^T ds \int_{-r}^{0} d\tau K_u'(s,\tau)L_u(s,\tau)\tilde{K}_u(s+\tau,t,\varphi),
\end{aligned} \tag{10b}
$$

$$
\begin{aligned}
Q_{u3}(t,\varphi) &= \int_t^T ds \int_{-r}^{0} d\tau K_u'(s+\tau,t)L_u'(s,\tau)\hat{K}_u(s,t,\varphi) \\
&= \int_t^T ds \int_{-r}^{0} d\tau K_u'(s+\tau,t)L_u'(s,\tau)\tilde{K}_u(s,t,\varphi),
\end{aligned} \tag{10c}
$$

$$
\begin{aligned}
Q_{u4}(t,\varphi) &= \int_t^T ds \int_{-r}^{0} d\alpha \int_{-r}^{0} d\rho K_u'(s+\alpha,t)G_u(s,\alpha,\rho)\hat{K}_u(s+\rho,t,\varphi) \\
&= \int_t^{\min[t+r+\varphi,T]} ds \int_{-r}^{0} d\alpha K_u'(s+\alpha,t)G_u(s,\alpha,t-s+\varphi) \\
&\quad + \int_t^T ds \int_{-r}^{0} d\alpha \int_{-r}^{0} d\rho K_u'(s+\alpha,t)G_u(s,\alpha,\rho)\tilde{K}_u(s+\rho,t,\varphi),
\end{aligned} \tag{10d}
$$

$$
\begin{aligned}
R_{u1}(t,\varphi,\rho) &= \int_t^T \hat{K}_u'(s,t,\varphi)M_u(s)\hat{K}_u(s,t,\rho)\,ds \\
&= \int_t^T \tilde{K}_u'(s,t,\varphi)M_u(s)\tilde{K}_u(s,t,\rho)\,ds,
\end{aligned} \tag{11a}
$$

$$R_{u2}(t,\varphi,\rho) = \int_t^T ds \int_{-r}^0 d\tau \hat{K}_u'(s,t,\varphi)L_u(s,\tau)\hat{K}_u(s+\tau,t,\rho)$$

$$= \int_t^{\min[t+r+\rho,T]} ds\tilde{K}_u'(s,t,\varphi)L_u(s,t-s+\rho) \tag{11b}$$

$$+ \int_t^T ds \int_{-r}^0 d\tau \tilde{K}_u'(s,t,\varphi)L_u(s,\tau)\tilde{K}_u(s+\tau,t,\rho),$$

$$R_{u3}(t,\varphi,\rho) = R_{u2}'(t,\rho,\varphi), \tag{11c}$$

$$R_{u4}(t,\varphi,\rho) = \int_t^T ds \int_{-r}^0 d\alpha \int_{-r}^0 d\beta \hat{K}_u'(s+\alpha,t,\varphi)G_u(s,\alpha,\beta)\hat{K}_u(s+\beta,t,\rho)$$

$$= \int_t^{\min[t+r+\varphi,t+r+\rho,T]} G_u(s,t-s+\varphi,t-s+\rho)\,ds$$

$$+ \int_t^{\min[t+r+\rho,T]} ds \int_{-r}^0 d\alpha \tilde{K}_u'(s+\alpha,t,\varphi)G_u(s,\alpha,t-s+\rho) \tag{11d}$$

$$+ \int_t^{\min[t+r+\varphi,T]} ds \int_{-r}^0 d\alpha G_u(s,\alpha,t-s+\varphi)\tilde{K}(s+\alpha,t,\rho)$$

$$+ \int_t^T ds \int_{-r}^0 d\alpha \int_{-r}^0 d\beta \tilde{K}_u'(s+\alpha,t,\varphi)G_u(s,\alpha,\beta)$$

$$\cdot \tilde{K}_u(s+\beta,t,\rho).$$

Furthermore, the T_i have the form (8) *where P_u, Q_u and R_u are replaced by P_{ui}, Q_{ui} and R_{ui}, respectively; P_u, Q_u and R_u have bounded derivatives in their arguments for*[7] $0 \leqq t \leqq T$, $-r \leqq \varphi \leqq 0$, $-r \leqq \rho \leqq 0$, *and satisfy* (12). *The derivatives are continuous, except for the φ or ρ derivative of $R_u(t,\varphi,\rho)$ at $\varphi = \rho$ where there may be finite discontinuity.*[8]

$$P_u(T) = Q_u(T,\varphi) = R_u(T,\varphi,\rho) = 0, \tag{12a}$$

$$\frac{dP_u(t)}{dt} + A_u'(t)P_u(t) + P_u(t)A_u(t) + Q_u(t,0) + Q_u'(t,0)$$

$$= -M(t) - E_u'(t)N(t)E_u(t) = -M_u(t), \tag{12b}$$

$$P_u(t)C_u(t,\varphi) + A_u'(t)Q_u(t,\varphi) + \frac{\partial Q(t,\varphi)}{\partial t} - \frac{\partial Q(t,\varphi)}{\partial \varphi} + R_u(t,0,\varphi)$$

$$= -E_u'(t)N(t)F_u(t,\varphi) = -L_u(t,\varphi). \tag{12c}$$

$$C_u'(t,\varphi)Q_u(t,\rho) + Q_u'(t,\varphi)C_u(t,\rho) + \frac{\partial R_u}{\partial t}(t,\varphi,\rho) - \frac{\partial R_u}{\partial \varphi}(t,\varphi,\rho) - \frac{\partial R_u}{\partial \rho}(t,\varphi,\rho)$$

$$= -F_u'(t,\varphi)N(t)F_u(t,\rho) = -G_u(t,\varphi,\rho), \tag{12d}$$

[7] At $\varphi = 0$ or $\varphi = r$ or $\rho = 0$ or $\rho = r$ or $t = 0$, the derivatives are replaced by the appropriate one-sided derivatives.

[8] For future reference, we note that the discontinuity in R_u is in the terms R_{u2} and R_{u3}. However, it is easy to verify that R_{u2} and R_{u3} are differentiable in the $(1,-1,-1)$ direction in the (t,φ,ρ) set $[0,T] \times [-r,0]^2$.

$$B'(t)P_u(t) - Q_u'(t, -r) = 0, \tag{12e}$$

$$B'(t)Q_u(t, \varphi) - R_u(t, -r, \varphi) - R_u'(t, \varphi, -r) + Q_u'(t, \varphi)B(t) = 0.$$

Finally, the solution $P_u(t)$, $Q_u(t, \varphi)$, $R_u(t, \varphi, \rho)$ is unique within the class of symmetric[9] *differentiable $P_u(t)$, $R_u(t, \varphi, \rho)$ and differentiable $Q_u(t, \varphi)$.*

Proof. The evaluation of the T_i-terms on the right of (6) is straightforward by merely substituting the expressions for $x(s)$, $x(s + \varphi)$ and $x(s + \rho)$ from (3) into the T_i and separating the result into a sum of the form of the right side of (8), where the P_{ui}, Q_{ui} and R_{ui} are given by (9)–(11). The right sides of (9)–(11) are obtained from the center expressions by replacing $\hat{K}$ by its definition in terms of $\tilde{K}$ and the δ-function, and noting that $\tilde{K}(s, t, \varphi) = 0$ for $s < t$. Then (8) follows by merely summing the T_i. The statement concerning the continuity of the derivatives of P_u, Q_u and R_u follow from Theorem 3 and the differentiability of $M_u(s)$, $L_u(s, \varphi)$ and $G_u(s, \varphi, \rho)$ for $0 \leqq s \leqq T$, $-r \leqq \varphi \leqq 0$, $-r \leqq \rho \leqq 0$.

Now, we evaluate

$$\frac{d}{dt}[x'(t)P_u(t)x(t)] = \left[A_u(t)x(t) + B(t)x(t - r) + \int_{-r}^{0} C_u(t, \varphi)x(t + \varphi)\, d\varphi\right]P_u(t)x(t)$$

$$+ x'(t)\left(\frac{dP_u(t)}{dt}\right)x(t) + x'(t)P_u(t)\left[A_u(t)x(t) + B(t)x(t - r)\right. \tag{13a}$$

$$\left. + \int_{-r}^{0} C_u(t, \varphi)x(t + \varphi)\, d\varphi\right],$$

$$\frac{d}{dt}\left[x'(t) \int_{-r}^{0} Q_u(t, \varphi)x(t + \varphi)\, d\varphi\right] = \frac{d}{dt}\left[x'(t) \int_{t-r}^{t} Q_u(t, \tau - t)x(\tau)\, d\tau\right]$$

$$= \left[A_u(t)x(t) + B(t)x(t - r) + \int_{-r}^{0} C_u(t, \varphi)x(t + \varphi)\right]' \int_{-r}^{0} Q_u(t, \varphi)x(t + \varphi)\, d\varphi \tag{13b}$$

$$+ x'(t)\left[Q_u(t, 0)x(t) - Q_u(t, -r)x(t - r) + \int_{t-r}^{t} \frac{\partial Q_u(t, \tau - t)}{\partial t}x(\tau)\, d\tau\right],$$

where

$$\int_{t-r}^{t} \frac{\partial Q_u(t, \tau - t)}{\partial t}x(\tau)\, d\tau = \int_{-r}^{0}\left[\frac{\partial Q_u(t, \varphi)}{\partial t} - \frac{\partial Q_u(t, \varphi)}{\partial \varphi}\right]x(t + \varphi)\, d\varphi. \tag{13c}$$

Similarly,

$$\frac{d}{dt}\left[\int_{-r}^{0} \int_{-r}^{0} d\varphi\, d\rho x'(t + \varphi)R_u(t, \varphi, \rho)x(t + \rho)\right]$$

$$= \frac{d}{dt} \int_{t-r}^{t} \int_{t-r}^{t} d\tau\, d\sigma x'(\tau)R_u(t, \rho - t, \sigma - t)x(\sigma)$$

[9] By symmetric M we mean $M'(t) = M(t)$; by symmetric $G(t, \rho, \varphi)$, we mean $G(t, \varphi, \rho) = G'(t, \rho, \varphi)$.

$$= \int_{t-r}^{t} d\sigma[x(t)'R_u(t,0,\sigma-t) - x'(t-r)R_u(t,-r,\sigma-t)]x(\sigma)$$
$$+ \int_{t-r}^{t} d\tau x'(\tau)[R_u(t,\tau,0)x(t) - R_u(t,\tau,-r)x(t-r)]$$
$$+ \int_{t-r}^{t}\int_{t-r}^{t} d\tau\, d\sigma x'(\tau)\left[\frac{\partial}{\partial t}R_u(t,\tau-t,\sigma-t)x(\sigma)\right]$$

(13d)
$$= \int_{-r}^{0} d\rho[x'(t)R_u(t,0,\rho) - x'(t-r)R_u(t,-r,\rho)]x(t+\rho)$$
$$+ \int_{-r}^{0} d\varphi x'(t+\varphi)[R_u(t,\varphi,0)x(t) - R_u(t,\varphi,-r)x(t-r)]$$
$$+ \int_{-r}^{0}\int_{-r}^{0} x'(t+\varphi)\left[\frac{\partial}{\partial t} - \frac{\partial}{\partial\varphi} - \frac{\partial}{\partial\rho}\right]R_u(t,\varphi,\rho)x(t+\rho)\,d\varphi\,d\rho.$$

Note (for reference in Theorems 5, 6) that the representations (13b), (13c), (13d) are valid if $Q_u(t,\varphi)$ only has a uniformly bounded derivative almost everywhere along each line in the $(1,-1)$ direction in the set $\varphi\in[-r,0], t\in[0,T]$, and if $R_u(t,\varphi,\rho)$ has only a uniformly bounded derivative almost everywhere along each line in the $(1,-1,-1)$ direction in the set $t\in[0,T], \varphi,\rho\in[-r,0]$. These conditions and the differentiability of $P_u(t)$ assure the differentiability (in t) of $V^u(x_t,t)$. Next, adding (13a), twice (13b) and (13d), and using the substitution (13c), yields an expression for $\partial V^u(x_t,t)/\partial t$. However, $\partial V^u(x_t,t)/\partial t$ also equals the negative of the sum of the bracketed integrands in (6), evaluated at $s=t$. The equality of these two forms[10] of $\partial V^u(x_t,t)/\partial t$ for all $x_t\in H$ and $0\leqq t\leqq T$ implies that the coefficients of like terms in $x(t)$, $x(t+\varphi)$, etc. in each form must be equal. This yields (12). Note that, by *construction* and Theorem 3, (12) has a smooth symmetric solution; i.e., the terms have continuous derivatives and $P_u(s)=P_u'(s)$, $R_u(t,\varphi,\rho)=R_u(t,\rho,\varphi)$ (except that the φ, ρ derivatives of R_u are discontinuous at $\varphi=\rho$).

Let $\hat{P}(t)$, $\hat{Q}(t,\varphi)$, $\hat{R}(t,\varphi,\rho)$ be differentiable solutions[11] to (12) with $\hat{P}(t)$, $\hat{R}(t,\varphi,\rho)$ symmetric and define $Z(x_t,t)$ by (14). Then, by reversing the argument leading to (12), we get $d/dt[Z(x_t,t)] = -x'(t)M(t)x(t) - u'(t)N(t)u(t)$.

(14)
$$x'(t)\hat{P}(t)x(t) + x'(t)\int_{-r}^{0}\hat{Q}(t,\varphi)x(t+\varphi)\,d\varphi + \int_{-r}^{0}x'(t+\varphi)\hat{Q}'(t,\varphi)x(t)\,d\varphi$$
$$+ \int_{-r}^{0}d\varphi\int_{-r}^{0}d\rho x'(t+\varphi)\hat{R}(t,\varphi,\rho)x(t+\varphi) = Z(x_t,t).$$

However,

$$Z(x_T,T) = V^u(x_T,T) = 0$$

[10] Note that $\partial V^u(x_t,t)/\partial t$ also equals $-x'(t)M(t)x(t) - u'(t)N(t)u(t)$.

[11] In fact, it is readily verified that we only need that $\hat{Q}(t,\varphi)$ and $\hat{R}(t,\varphi,\rho)$ have uniformly bounded derivatives, i.e., in the $(1,-1)$ and $(1,-1,-1)$ directions on the sets $t\in[0,T]$, $\Phi\in[-r,0]$ and $t\in[0,T]$, $\varphi,\rho\in[-r,0]$, respectively. More generally, for uniqueness we only need that $\partial R_u(t,\varphi-t,\rho-t)/\partial t$ and $\partial Q_u(t,\varphi-t)/\partial t$ be uniformly bounded for almost all φ,ρ.

and

$$Z(x_t, t) - Z(x_T, T) = \int_t^T [x'(s)M(s)x(s) + u'(s)N(s)u(s)]\, ds$$
$$= V^u(x_t, t) - V^u(x_T, T)$$

or, equivalently,

$$Z(x_t, t) = V^u(x_t, t). \tag{15}$$

Using the identity (15), the representations (14) and (8), and the continuity of the $P, \hat{P}, Q, \hat{Q}, R, \hat{R}$, and symmetry of $P, \hat{P}$ and $R, \hat{R}$, it is easily shown that[12] $P_u(t) = \hat{P}(t)$, $Q_u(t, \varphi) = \hat{Q}(t, \varphi)$, $R_u(t, \varphi, \rho) = \hat{R}(t, \varphi, \rho)$; thus the uniqueness is proved.

In the sequel, it will be helpful to separate out the u-dependent terms in the coefficients of P_u, Q_u and R_u in (12b, c, d) and to eliminate the u-dependence of the kernels K_u and $\tilde{K}_u$ in (10). Write (12b, c, d) as

$$\frac{dP_u(t)}{dt} + A'(t)P_u(t) + P_u(t)A(t) + Q_u(t, 0) + Q_u'(t, 0) = -\hat{M}_u(t), \tag{12b'}$$

$$P_u(t)C(t, \varphi) + A'(t)Q_u(t, \varphi) + \frac{\partial Q_u(t, \varphi)}{\partial t} - \frac{\partial Q_u(t, \varphi)}{\partial \varphi} + R_u(t, 0, \varphi) = -\hat{L}_u(t, \varphi), \tag{12c'}$$

$$C'(t, \varphi)Q_u(t, \rho) + Q'(t, \varphi)C(t, \rho) + \frac{\partial R(t, \varphi, \rho)}{\partial t} - \frac{\partial R(t, \varphi, \rho)}{\partial \varphi} - \frac{\partial R(t, \varphi, \rho)}{\partial \rho} = -\hat{G}_u(t, \varphi, \rho), \tag{12d'}$$

where

$$\hat{M}_u(t) = M_u(t) + E_u'(t)D'(t)P_u(t) + P_u(t)D(t)E_u(t) \tag{16a}$$

$$\hat{L}_u(t, \varphi) = L_u(t, \varphi) + P_u(t)D(t)F_u(t, \varphi) + \tfrac{1}{2}[E_u'(t)D'(t)Q_u(t, \varphi) + Q_u'(t, \varphi)D(t)E_u(t)], \tag{16b}$$

$$\hat{G}_u(t, \varphi, \rho) = G_u(t, \varphi, \rho) + F_u'(t, \varphi)D'(t)Q_u(t, \rho) + Q_u'(t, \varphi)D(t)F_u(t, \rho). \tag{16c}$$

The boundary conditions (13a, e) do not depend on u.

THEOREM 2. *Suppose the conditions of Theorem* 1. *Define* $\hat{P}_{ui}$, $\hat{Q}_{ui}$ *and* $\hat{R}_{ui}$ *as the terms in* (9′)–(11′), *or, equivalently, the respective terms in* (9)–(11) *with* $K, \tilde{K}$, $\hat{M}_u$, $\hat{L}_u$ *and* $\hat{G}_u$ *replacing* K_u, $\tilde{K}_u$, M_u, L_u *and* G_u, *respectively. Then*

$$P_u(t) = \sum_1^4 \hat{P}_{ui}(t), \quad Q_u(t, \varphi) = \sum_1^4 \hat{Q}_{ui}(t, \varphi), \quad R_u(t, \varphi, \rho) = \sum_1^4 \hat{R}_{ui}(t, \varphi, \rho), \tag{17}$$

$$\hat{P}_{u1}(t) = \int_t^T K'(s, t)\hat{M}_u(s)K(s, t)\, ds, \tag{9a'}$$

[12] In fact, under the weaker hypothesis of the last footnote, the equalities hold between Q_u, $\hat{Q}$ and R_u, $\hat{R}$ almost everywhere in (φ, ρ) for each t.

$$\text{(9b')} \qquad \hat{P}_{u2}(t) = \int_t^T ds \int_{-r}^0 d\tau K'(s,t)\hat{L}_u(s,\tau)K(s+\tau,t),$$

$$\text{(9c')} \qquad \hat{P}_{u3}(t) = P'_{u2}(t),$$

$$\text{(9d')} \qquad \hat{P}_{u4}(t) = \int_t^T ds \int_{-r}^0 d\varphi \int_{-r}^0 d\rho K'(s+\varphi,t)\hat{G}_u(s,\varphi,\rho)K(s+\rho,t),$$

$$\text{(10a')} \quad \hat{Q}_{u1}(t,\varphi) = \int_t^T dsK'(s,t)\hat{M}_u(s)\tilde{K}(s,t,\varphi),$$

$$\text{(10b')} \quad \begin{aligned}\hat{Q}_{u2}(t,\varphi) = {} & \int_t^T ds \int_{-r}^0 d\tau K'(s,t)\hat{L}_u(s,\tau)\tilde{K}(s+\tau,t,\varphi) \\ & + \int_t^{\min[t+r+\varphi,T]} K'(s,t)\hat{L}_u(s,t-s+\varphi)\,ds,\end{aligned}$$

$$\text{(10c')} \quad \hat{Q}_{u3}(t,\varphi) = \int_t^T ds \int_{-r}^0 d\tau K'(s+\tau,t)\hat{L}'_u(s,\tau)\tilde{K}(s,t,\varphi),$$

$$\text{(10d')} \quad \begin{aligned}\hat{Q}_{u4}(t,\varphi) = {} & \int_t^T ds \int_{-r}^0 d\alpha \int_{-r}^0 d\rho K'(s+\alpha,t)\hat{G}_u(s,\alpha,\rho)K(s+\rho,t,\varphi) \\ & + \int_t^{\min[t+\varphi+r,T]} ds \int_{-r}^0 d\alpha K'(s+\alpha,t)\hat{G}_u(s,\alpha,t-s+\varphi),\end{aligned}$$

$$\text{(11a')} \quad \hat{R}_{u1}(t,\varphi,\rho) = \int_t^T dsK'(s,t,\varphi)\hat{M}_u(s)\tilde{K}(s,t,\rho),$$

$$\text{(11b')} \quad \begin{aligned}\hat{R}_{u2}(t,\varphi,\rho) = {} & \int_t^T ds \int_{-r}^0 d\tau \tilde{K}'(s,t,\varphi)\hat{L}_u(s,\tau)\tilde{K}(s+\tau,t,\rho) \\ & + \int_t^{\min[t+r+\rho,T]} \tilde{K}(s,t,\varphi)\hat{L}_u(s,t-s+\rho)\,ds,\end{aligned}$$

$$\text{(11c')} \quad \hat{R}_{u3}(t,\varphi,\rho) = R'_{u2}(t,\rho,\varphi),$$

$$\text{(11d')} \quad \begin{aligned}\hat{R}_{u4}(t,\varphi,\rho) = {} & \int_t^T ds \int_{-r}^0 d\alpha \int_{-r}^0 d\beta \tilde{K}'(s+\alpha,t,\varphi)\hat{G}_u(s,\alpha,\beta)\tilde{K}(s+\beta,t,\rho) \\ & + \int_t^{\min[t+r+\varphi,t+r+\rho,T]} \hat{G}_u(s,t-s+\varphi,t-s+\rho)\,ds \\ & + \int_t^{\min[t+r+\rho,T]} ds \int_{-r}^0 d\alpha \tilde{K}'(s+\alpha,t,\varphi)\hat{G}_u(s,\alpha,t-s+\rho) \\ & + \int_t^{\min[t+r+\varphi,T]} ds \int_{-r}^0 d\alpha \hat{G}_u(s,\alpha,t-s+\varphi)\tilde{K}(s+\alpha,t,\rho).\end{aligned}$$

Proof. In the integrals (9) in the expression $\sum_1^4 P_{ui}(t)$, replace K_u and $\tilde{K}_u$ by K and $\tilde{K}$, respectively, and M_u, L_u, G_u by $\hat{M}_u$, $\hat{L}_u$, $\hat{G}_u$, respectively. In Theorem 1,

let $u \equiv 0$, $L_0 = \hat{L}_u$, $M_0 = \hat{M}_u$, $G_0 = \hat{G}_u$. With this replacement, the P_{ui} terms in (9) become the $\hat{P}_{ui}$ terms in (9′). Then, by Theorem 1, the $\hat{P}_{ui}(t)$ are differentiable, and $\sum_1^4 \hat{P}_{ui}(t) = \hat{P}_u(t)$ satisfies (12b′) (or, equivalently, (12b)). This follows similarly for $\sum_1^4 \hat{Q}_{ui}(t, \varphi) = \hat{Q}_u(t, \varphi)$ and $\sum_1^4 \hat{R}_{ui}(t, \varphi, \rho) = \hat{R}_u(t, \varphi, \rho)$. Then, by the symmetry of $\hat{P}_u(t)$ and $\hat{R}_u(t, \varphi, \rho)$ and the uniqueness part of Theorem 1, we have (17).

THEOREM 3. *Suppose that $N(t)$, $M(t)$, $A(t)$, $B(t)$, $C(t, \varphi)$, $D(t)$, $E_u(t)$ and $F_u(t, \varphi)$ satisfy the condition of Theorem* 1. *Then the $P_{ui}(t)$, $Q_{ui}(t, \varphi)$ and $R_{ui}(t, \varphi, \rho)$ of* (9)–(11) *are continuously differentiable in their arguments for $0 \leqq t \leqq T$, $-r \leqq \varphi \leqq 0$, $-r \leqq \rho \leqq 0$, except that the φ or ρ derivatives of $R_{u2}(t, \varphi, \rho)$ and $R_{u3}(t, \varphi, \rho)$ may be discontinuous at $\varphi = \rho$. However, $R_u(t, \varphi, \rho)$ has a derivative in the $(1, -1, -1)$ direction.*

Proof. Since the evaluations are tedious and straightforward, we give the details for one "typical" term only, namely $Q_{u2}(t, \varphi)$. We note only that the asserted discontinuity in R_{u2} arises from the latter term of (11b′) and that it is easy to verify that $(\partial/\partial t - \partial/\partial\varphi)$ applied to this latter term yields a continuous function. For future reference note that the discontinuity is uniformly bounded if the L_u are. Write

$$Q_{u2}(t, \varphi) = \int_t^T \int_{-r}^0 K'_u(s, t) L_u(s, \tau) \tilde{K}_u(s + \tau, t, \varphi)\, ds\, d\tau + \int_t^{\min[t+r+\varphi, T]} K'_u(s, t) L_u(s, t - s + \varphi)\, ds.$$

Recall that $L_u(t, \varphi) = E'_u(t) N(t) F_u(t, \varphi)$.

Denote the second term of $Q_{u2}(t, \varphi)$ by $\beta(t, \varphi)$. Observe that it is continuous in (t, φ). Let $t + r + \varphi > T$. Then

$$\frac{\partial \beta(t, \varphi)}{\partial \varphi} = \int_t^T K'_u(s, t) \frac{\partial L_u}{\partial \varphi}(s, t - s + \varphi) ds$$

which is continuous in (t, φ). For $t + r + \varphi < T$, we have

$$\frac{\partial \beta(t, \varphi)}{\partial \varphi} = K'_u(t + r + \varphi, t) L_u(t + r + \varphi, -r) + \int_t^T K'_u(s, t) \frac{\partial L_u}{\partial \varphi}(s, t - s + \varphi)$$

which is continuous in (t, φ) in the desired range. In addition, $L_u(t + r + \varphi, -r) \to 0$ as $t + r + \varphi \to T$, since $F_u(t, \varphi) \to 0$ as $t \to T$. Thus $\beta(t, \varphi)$ has continuous φ derivatives for $t, \varphi \in [0, T] \times [-r, 0]$. The details for $\partial\beta(t, \varphi)/\partial t$ are similar and are omitted.

Write the first term of $Q_{u2}(t, \varphi)$ as

$$\alpha(t, \varphi) = \int_t^T h(s, \varphi, t)\, ds,$$

where

$$h(s, \varphi, t) = \int_{\max[t - s + \varphi, -r]}^0 K'_u(s, t) L_u(s, \tau) \tilde{K}_u(s + \tau, t, \varphi)\, ds.$$

If $t - s + \varphi > 0$, the lower limit is replaced by zero.

For each fixed $t \geqq 0$, let $k(s, \varphi, t)$ satisfy: (a) $k(s, \varphi, t)$ is continuous on $[t, T] \times [-r, 0]$; (b) There is a bounded measurable function $k_\varphi(s, \varphi, t)$ so that for each t and each s, not in some null set in $[t, T]$, $k_\varphi(s, \varphi, t) = \partial k(s, \varphi, t)/\partial\varphi$ for almost all φ in $[-r, 0]$; (c) $\int_t^T k_\varphi(s, \varphi, t)\,ds$ is continuous on $[0, T] \times [-r, 0]$. Then $\int_t^T k_\varphi(s, \varphi, t)ds = \partial/\partial\varphi \int_t^T k(s, \varphi, t)ds$ and is continuous on $[0, T] \times [-r, 0]$. Let $k(s, \varphi, t) = h(s, \varphi, t)$, and note that $h(s, \varphi, t)$ is continuous for each fixed t. Let $t - s + \varphi < -r$. Then

$$\delta_1(s, \varphi, t) \equiv \frac{\partial h(s, \varphi, t)}{\partial\varphi} = \int_{-r}^{0} K_u'(s, t)L_u(s, \tau)\frac{\partial \tilde{K}_u(s + \tau, t, \varphi)}{\partial\varphi}\,ds$$

which is continuous in all three variables.

Now, let $0 > t - s + \varphi > -r$. Then

$$\partial_2(s, \varphi, t) \equiv \frac{\partial k(s, \varphi, t)}{\partial\varphi} = K_u'(s, t)L_u(s, t - s + \varphi)\tilde{K}_u(t + \varphi, t, \varphi) + \int_{t-s+\varphi}^{0} K_u'(s, t)L_u(s, \tau)\frac{\partial \tilde{K}_u}{\partial\varphi}(s + \tau, t, \varphi)d\tau.$$

The first term of $\delta_2(s, \varphi, t)$ is zero since $\tilde{K}_u(t + \varphi, t, \varphi) = 0$ and the second tends to $\delta_1(s, \varphi, t)$ as $t - s + \varphi \downarrow -r$. It can now easily be verified that (a)–(c) hold and that $\alpha(t, \varphi)$ has a continuous φ derivative on $[0, T] \times [-r, 0]$. The details for $\partial\alpha(t, \varphi)/\partial t$ are similar and are omitted.

4. Iteration in policy space. In Theorem 4, the basic result on "iteration in policy space," we will require the time derivative of the function $V^u(x_t, t)$ evaluated on the path corresponding to a control w (and written $\dot{V}^{u,w}(x_t, t)$); to be specific the time derivative of $V^u(x_t, t)$ along the path corresponding to w is defined by

$$\dot{V}^{u,w}(x_t, t) = \frac{\partial}{\partial t}\Big[x'(t)P_u(t)x(t) + 2x'(t)\int_{-r}^{0} Q_u(t, \varphi)x(t + \varphi)\,d\varphi + \int_{-r}^{0}\int_{-r}^{0} x'(t + \varphi)R_u(t, \varphi, \rho)x(t + \rho)\,d\rho\,d\varphi\Big], \tag{18}$$

where for $\dot{x}(t) \equiv \partial x(t)/\partial t$ we use the derivative evaluated along the trajectory corresponding to w; i.e.,

$$\dot{x}(t) = A(t)x(t) + B(t)x(t - r) + D(t)w(t) + \int_{-r}^{0} C(t, \varphi)x(t + \varphi)\,d\varphi. \tag{19}$$

Using (19) in the calculations (13), we have

$$\dot{V}^{u,w}(x_t, t) = 2w'(t)D(t)P_u(t)x(t) + 2w'(t)D'(t)\int_{-r}^{0} Q_u(t, \varphi)x(t + \varphi)\,d\varphi + x'(t)\Big[\frac{dP_u(t)}{dt} + A'(t)P_u(t) + P_u(t)A(t) + Q_u(t, 0) + Q_u'(t, 0)\Big]x(t)$$

$$
\begin{aligned}
&+ x'(t)\int_{-r}^{0}\Big[2\Big(\frac{\partial}{\partial t}-\frac{\partial}{\partial\varphi}\Big)Q_u(t,\varphi)+2P_u(t)C(t,\varphi)\\
&\qquad + A'(t)Q_u(t,\varphi)+Q_u'(t,\varphi)A(t)\\
&\qquad + R_u(t,\varphi,0)+R_u(t,0,\varphi)\Big]x(t+\varphi)\,d\varphi\\
&+\int_{-r}^{0}\int_{-r}^{0}x'(t+\varphi)\Big[\Big(\frac{\partial}{\partial t}-\frac{\partial}{\partial\varphi}-\frac{\partial}{\partial\rho}\Big)R_u(t,\varphi,\rho)\\
&\qquad + C'(t,\varphi)Q_u(t,\rho)+Q'(t,\varphi)C(t,\rho)\Big]x(t+\rho)\,d\varphi\,d\rho.
\end{aligned}
\tag{19a}
$$

THEOREM 4. *Let u have the form* (5), *and define $\dot{V}^{u,w}(x_t,t)$ by* (18). *Assume the conditions on A, B, C, D, E_u, F_u, N and M of Theorem* 1, *and let $N(s)$ be positive definite and $M(s)$ positive semidefinite in $[0,T]$, and let $D(t)$ be continuously differentiable in $[0,T]$. The control w which attains the minimum in* (22) *has the form* (5), *and*

$$
w(t)=E_w(t)x(t)+\int_{-r}^{0}F_w(t,\varphi)x(t+\varphi)d\varphi,
\tag{20a}
$$

where

$$
\begin{aligned}
E_w(t)&=-N^{-1}(t)D'(t)P_u(t),\\
F_w(t,\varphi)&=-N^{-1}(t)D'(t)Q_u(t,\varphi).
\end{aligned}
\tag{20b}
$$

$E_w(t)$ and $F_w(t,\varphi)$ satisfy the conditions on the $E_u(t)$ and $F_u(t,\varphi)$ in Theorem 5. *Also*

$$
V^w(x_t,t)\leqq V^u(x_t,t)
\tag{21}
$$

for all $x_t\in H$, and $t\in[0,T]$.

$$
H(x_t,t)=\min_w\,[\dot{V}^{u,w}(x_t,t)+x'(t)M(t)x(t)+w'(t)N(t)w(t)].
\tag{22}
$$

Remark. Note that, with $w=u$, the bracketed term in (22) is zero by the definition of $\dot{V}^{u,u}(x_t,t)=\partial V^u(x_t,t)/\partial t$.

Proof. In computing the minimum in (22), only the terms

$$
\dot{x}'(t)P_u(t)x(t)+x'(t)P_u(t)\dot{x}(t)+2\dot{x}'(t)\int_{-r}^{0}Q_u(t,\varphi)x(t+\varphi)\,d\varphi+w'(t)N(t)w(t)
\tag{23a}
$$

or, equivalently, only the terms

$$
2w'(t)D'(t)P_u(t)x(t)+2w'(t)D'(t)\int_{-r}^{0}Q_u(t,\varphi)x(t+\varphi)\,d\varphi+w'(t)N(t)w(t)
\tag{23b}
$$

need be taken into account. The other terms in the brackets in (22) do not contain w by (19a). The $w(t)$ minimizing (23b) is of the form (20a), whereas E_w and F_w satisfy (20b). By the hypothesis and by Theorem 1, the coefficients E_w and F_w satisfy the smoothness conditions required in Theorem 1 on the E_u, F_u there.

Now, for any w of the form (20), $V^u(x_T, T) = V^w(x_T, T) = 0$ and

$$\int_t^T \dot{V}^{u,v}(x_t, t)\, dt = V^u(x_T, T) - V^u(x_t, t).$$

The bracketed term in (22), with the minimizing w inserted, is nonpositive since the bracketed term is zero if w is replaced by u. Thus

$$0 \geqq \int_t^T \dot{V}^{u,v}(x_s, s)\, ds + \int_t^T [x'(s)M(s)x(s) + w'(s)N(s)w(s)]\, ds$$

or

$$0 \geqq V^u(x_T, T) - V^u(x_t, t) + V^w(x_t, t) - V^w(x_T, T) = -V^u(x_t, t) + V^w(x_t, t)$$

and (21) holds. This completes the proof.

Suppose the conditions on A, B, C, D, N and M of Theorem 4 hold. Let u_0 satisfy the conditions in the remark below Lemma 1. Define the improved control u_n recursively in terms of u_{n-1} by the method of Theorem 4. Then, by Theorem 4 (where we write $E_n = E_{u_n}, F_n = F_{u_n}, V^n = V^{u_n}$),

$$u_n = E_n(t)x(t) + \int_{-r}^0 F_n(t, \varphi)x(t + \varphi)\, d\varphi, \tag{24}$$

$$\begin{aligned} E_{n+1}(t) &= -N^{-1}(t)D'(t)P_n(t), \\ F_{n+1}(t, \varphi) &= -N^{-1}(t)D'(t)Q_n(t, \varphi), \end{aligned} \tag{25}$$

and, for all $t \in [0, T]$ and $x_t \in H$,

$$V^{n+1}(x_t, t) \leqq V^n(x_t, t). \tag{26}$$

Next, it is shown that (26) implies that the P_n, Q_n, R_n and u_n converge.

THEOREM 5. *Assume the conditions of Theorem* 4. *The* $P_n(t)$, $Q_n(t, \varphi)$, $R_n(t, \varphi, \rho)$, $E_n(t)$ *and* $F_n(t, \varphi)$ *are uniformly bounded and converge pointwise to functions* $P(t)$, $Q(t, \varphi)$, $R(t, \varphi, \rho)$, $E(t)$ *and* $F(t, \varphi)$, *respectively.* $P(t)$ *and* $R(t, \varphi, \rho)$ *are symmetric and*

$$\begin{aligned} V^u(x_t, t) = {} & x'(t)P(t)x(t) + x'(t)\int_{-r}^0 Q(t, \varphi)x(t + \varphi)\, d\varphi \\ & + \int_{-r}^0 x'(t + \varphi)Q'(t, \varphi)x(t)\, d\varphi \\ & + \int_{-r}^0 \int_{-r}^0 x'(t + \varphi)R(t, \varphi, \rho)x(t + \rho)\, d\varphi\, d\rho, \end{aligned} \tag{27}$$

where u *is the limit of the* u_n:

$$u(t) = E(t)x(t) + \int_{-r}^0 F(t, \varphi)x(t + \varphi)\, d\varphi. \tag{28}$$

Furthermore, the $\hat{M}_n$, $\hat{G}_n$ *and* $\hat{L}_n$ *in* (9′)–(11′) *converge pointwise and are uniformly bounded and the* P, Q *and* R *are the limits of the sums of the* P_{ni}, Q_{ni} *and* R_{ni}, *respectively.*

Finally, let v be the $(1, -1)$ direction in the (t, φ) set $[0, T] \times [-r, 0]$, and σ the $(1, -1, -1)$ direction in the (t, φ, ρ) set $[0, T] \times [-r, 0]^2$. Then the derivatives $\partial P(t)/\partial t$, $\partial Q(t, \varphi)/\partial v$, $\partial R(t, \varphi, \rho)/\partial \sigma$ exist and satisfy

$$\frac{\partial P(t)}{\partial t} + A'(t)P(t) + P(t)A(t) + Q(t, 0) + Q'(t, 0) = -\hat{M}(t), \tag{29a}$$

$$\sqrt{2}\frac{\partial Q(t, \varphi)}{\partial v} + P(t)C(t, \varphi) + A'(t)Q(t, \varphi) + R(t, 0, \varphi) = -\hat{L}(t, \varphi), \tag{29b}$$

$$\sqrt{3}\frac{\partial R(t, \varphi, \rho)}{\partial \sigma} + C'(t, \varphi)Q(t, \rho) + Q'(t, \varphi)C(t, \rho) = -\hat{G}(t, \varphi, \rho), \tag{29c}$$

where the $\hat{M}$, $\hat{L}$ and $\hat{G}$ are the $\hat{M}_n$, $\hat{G}_n$, $\hat{L}_n$, with E_n and F_n replaced by their limit. Also

$$\begin{aligned} B'(t)P_u(t) - Q'(t, -r) &= 0, \\ B'(t)Q(t, \varphi) - R(t, -r, \varphi) - R'(t, \varphi, -r) + Q'(t, \varphi)B(t) &= 0. \end{aligned} \tag{29d}$$

$\partial P_n(t)/\partial t$, $\partial Q_n(t, \varphi)/\partial v$ and $\partial R_n(t, \varphi, \rho)/\partial \sigma$ converge to $\partial P(t)/\partial t$, $\partial Q(t, \varphi)/\partial v$ and $\partial R(t, \varphi, \rho)/\partial \sigma$, respectively.

Proof. The other statements follow readily from the uniform boundedness and convergence of the P_n, Q_n and R_n and Theorems 1 and 2; hence only this will be shown.

We note only that

$$(\partial/\partial t - \partial/\partial \varphi)Q_n(t, \varphi) = \sqrt{2}\,\partial Q_n(t, \varphi)/\partial v,$$

and

$$(\partial/\partial t - \partial/\partial \varphi - \partial/\partial \rho)R_n(t, \varphi, \rho) = \sqrt{3}\,\partial R_n(t, \varphi, \rho)/\partial \sigma.$$

These derivatives converge if the P_n, Q_n and R_n do, and are uniformly bounded by (12) and (12′). If the P_n, Q_n and R_n and their (t, v, σ, respectively) derivatives all converge, then the (t, v, σ, respectively) derivatives of the limits are the limits of the (t, v, σ, respectively) derivatives. In (26), let $x(t + \varphi) = 0$ for $\varphi \neq 0$. Then (26) implies that $x'P_{n+1}(t)x \leqq x'P_n(t)x$ for any vector x. Hence, $P_n(t)$ converges pointwise to a symmetric measurable matrix $P(t)$. Since the diagonal elements $p_{n,ii}(t)$ are nonincreasing, and $|p_{n,ij}(t)| \leqq \max_i p_{n,ii}(t)$, the $P_n(t)$ are uniformly bounded.

Let $x(\varphi)$ be any continuous function on $[-r, 0]$ with $x(0) = 0$. Then, for such $x(\varphi)$, (26) implies that

$$\int_{-r}^{0}\int_{-r}^{0} x'(\varphi)R_{n+1}(t, \varphi, \rho)x(\rho)\,d\varphi\,d\rho \leqq \int_{-\sigma}^{0}\int_{-\sigma}^{0} x'(\varphi)R_n(t, \varphi, \rho)x(\rho)\,d\varphi\,d\rho. \tag{30}$$

By the continuity of the $R_n(t, \varphi, \rho)$, (30) holds if $x(\varphi)$ is a Dirac δ-function. In particular, if $-r < \varphi_0 < 0$, $-r < \rho_0 < 0$ and $x(\varphi) = x\delta(\varphi - \varphi_0) + y\delta(\varphi - \rho_0)$, then (30) and the fact that $R_n'(t, \varphi, \rho) = R_n(t, \rho, \varphi)$ yield

$$\begin{aligned} &x'R_{n+1}(t, \varphi_0, \varphi_0)x + y'R_{n+1}(t, \rho_0, \rho_0)y + 2x'R_{n+1}(t, \varphi_0, \rho_0)y \\ &\qquad \leqq x'R_n(t, \varphi_0, \varphi_0)x + y'R_n(t, \rho_0, \rho_0)y + 2x'R_n(t, \varphi_0, \rho_0)y. \end{aligned} \tag{31}$$

But, by continuity of the $R_n(t, \varphi, \rho)$, (31) holds for any φ_0, ρ_0 in $[-r, 0]$. Let $y = 0$. Then, as shown for the P_n, (31) implies that the $R_n(t, \varphi, \varphi)$ are uniformly bounded and converge to some $R(t, \varphi, \varphi)$. Using this and (31) and the arbitrariness of x, y implies that the $R_n(t, \varphi, \rho)$ are uniformly bounded and that $R_n(t, \varphi, \rho)$ converges to some $R(t, \varphi, \rho)$. By similar reasoning, (26) implies that, for each $\varphi_0 \in [-r, 0]$,

$$x'P_{n+1}(t)x + 2x'Q_{n+1}(t, \varphi_0)y + y'R_{n+1}(t, \varphi_0, \varphi_0)y \tag{32}$$
$$\leqq x'P_n(t)x + 2x'Q_n(t, \varphi_0)y + y'R_n(t, \varphi_0, \varphi_0)y.$$

Using (32) and the conclusions concerning P_n and R_n, we may deduce that the $Q_{n+1}(t, \varphi)$ converge to some $Q(t, \varphi)$ and are uniformly bounded.

COROLLARY. *For any control $w(t)$ which gives bounded continuous paths $x(t)$, and which is bounded for any bounded continuous initial condition, $\dot{V}^{u,w}(x_t, t)$ exists and $\dot{V}^{u_n,w}(x_t, t)$ converges to it for any continuous initial condition. The class of $w(t)$ includes all controls which are linear in x_t and have bounded coefficients.*

Note. Recall that $\dot{V}^{u,w}(x_t, t)$ is the *time derivative* of $V^u(x_t, t)$ along x_t paths corresponding to the control w.

Proof. Since $V^{u_n}(x_t, t)$ converges to $V^u(x_t, t)$ for any continuous initial condition, we only need to show that $\dot{V}^{u_n,w}(x_t, t)$ is uniformly bounded (in n) and converges for any continuous initial condition. $\dot{V}^{u_n,w}(x_t, t)$ is given by (19a) with u_n replacing u, and Theorem 5 implies that $\dot{V}^{u_n,w}(x_t, t)$ converges.

5. The optimality theorem.

THEOREM 6. *Let $w(x, t)$ be any control for which a solution to* (1) *is defined on $[0, T]$ for any initial condition, and let u be given by* (28). *Then $V^u(x_t, t) \leqq V^w(x_t, t)$ for all t and initial conditions x_t. Let $u = w$ and E_u and F_u be given by* (28). *Then the set of equations* (29) *has a unique solution (for symmetric $P(t)$ and $R(t, \varphi, \rho)$) and determines the optimal control w.*

Proof. Calculating the minimizing w in (33) (see Theorem 4 for terminology)

$$\min_w [\dot{V}^{u,w}(x_t, t) + x'(t)M(t)x(t) + w'(t)N(t)w(t)] \tag{33}$$

yields (see (19a))

$$w(x_t, t) = -N^{-1}(t)D'(t)\left[P(t)x(t) + \int_{-r}^{0} Q(t, \varphi)x(t + \varphi)\, d\varphi\right],$$

which is exactly u. Also the bracketed term in (33) is zero if u replaces w. Thus, for any $u \neq w$, we have

$$\dot{V}^{u,w}(x_t, t) + x'(t)M(t)x(t) + w'(t)N(t)w(t) \geqq 0$$

or

$$0 \leqq \int_t^T \dot{V}^{u,w}(x_s, s)\, ds + \int_t^T [x'(s)M(s)x(s) + w'(s)N(s)w(s)]ds$$
$$= -V^u(x_t, t) + V^u(x_T, T) + V^w(x_t, t) - V^w(x_T, T)$$

or, equivalently, $V^w(x_t, t) \geqq V^u(x_t, t)$. The last sentence of the theorem follows from Theorems 5 and 2.

REFERENCES

[1] N. N. KRASOVSKII, *Analytic construction of optimal controllers in systems with time lags*, J. Appl. Math. Mech., 26 (1962), pp. 39–51.

[2] ———, *Optimal processes in systems with time lags*, Proc. IFAC Congress, Pergamon Press, London, 1963.

[3] D. BARNEA, *Control and stabilization of stochastic functional-differential equations*, Doctoral thesis, Department of Engineering, Brown University, Providence, 1969.

[4] A. HALANAY, *Differential Equations; Stability, Oscillations, Time Lags*, Academic Press, New York, 1966.

[5] M. N. OGUZTORELI, *Time-Lag Control Systems*, Academic Press, New York, 1966.

SIAM J. CONTROL
Vol. 8, No. 2, May 1970

OPTIMAL CONTROL OF VIBRATING THIN PLATES*

VADIM KOMKOV†

Summary. This paper establishes some basic generalizations of Pontryagin's principle for vibrating thin, inhomogeneous plates, subject to mixed boundary conditions. (Part of the boundary is simply supported, part of it is free, and part of it is built in.) An instantly optimal control, as defined by the author in [7], is studied, and a corresponding principle is developed. Finally some comments are made regarding the nature of an optimal excitation for thin plates.

Introductory remarks. The results of this paper, concerning an optimal control of thin plates for a fixed time interval, are similar to those for the symmetric hyperbolic case, as given by Russell in [16]. In some cases they are a direct generalization of the optimal control theory for vibrating beams [7]. In some cases no obvious attempt to generalize the beam theory will succeed, and Pontryagin's principle assumes a very complex form.

1.1. Assumptions and notation. We assume the usual linear hypothesis of thin plate theory. These assumptions imply Duhamel's principle as given in § 1.9. The plate is assumed to occupy a compact simply connected region $\bar{\Omega}$ of the Euclidean space E^2. The interior of $\bar{\Omega}$ will be denoted by Ω, and the boundary of $\bar{\Omega}$ by $\partial\Omega$. The boundary $\partial\Omega$ consists of smooth Jordan curves. The problem of corner points will be considered in the final paragraphs of this paper. Prior to that the boundary curves will be assumed smooth (i.e., of the class C^1).

Notation and the physical meaning of symbols used here is as follows:

x, y, z will denote Cartesian coordinates of E^3.

t will denote time.

The plate occupies a region $\Omega \subset E^2$, the plane of Ω being spanned by the coordinates x, y.

u, v, w may denote the displacements in the directions of the axes x, y, z respectively.

v—transverse velocity $v = dw/dt$

E—Young's modulus ($E > 0$)

ν—Poisson's ratio ($0 < \nu < \frac{1}{2}$)

h—thickness of the plate ($h > 0$ in Ω)

D—flexural rigidity: $D = Eh^3/(12(1 - \nu^2))$

ρ—the mass density (mass per unit area)

n—the unit vector in the direction of the outward normal to $\partial\Omega$

τ—the unit vector in the direction tangential to $\partial\Omega$

ds—increment of length

$\epsilon_{xx}, \epsilon_{xy}, \epsilon_{yy}$—the linear strains

$\tau_{xx}, \tau_{xy}, \tau_{yy}$—the linear stresses (The stress system is assumed to be two-dimensional.)

* Received by the editors January 3, 1969.

† Florida State University, Tallahassee, Florida, and Texas Technological University, Lubbock, Texas 79409. This research was supported in part by the National Science Foundation under Grant GP 8921.

M_{ij}—the moments
Q_i—the shear forces (per unit length)
T—kinetic energy
U—strain energy
$\mathscr{E}$—total energy ($\mathscr{E} = T + U$)

The displacement functions $u(x, y, t)$, $v(x, y, t)$, $w(x, y, t)$ will satisfy the "elastica" hypothesis listed below:

(E1) For a fixed t the functions u, v, w and their time derivatives $\partial u/\partial t$, $\partial v/\partial t$, $\partial w/\partial t$ possess in Ω the partial derivatives of order one and two with respect to x and y.

(E2) The two-dimensional strain components are twice differentiable functions almost everywhere in Ω.

(E3) For each fixed x and y the displacement function $w(x, y, t)$ is a continuously differentiable function of t.

(E4) The functions $D(x, y)(\nabla^2 w)^2 - (1 - \nu)\diamondsuit^4(w, w)$ and $\rho(x, y)(\partial w/\partial t)^2$ are bounded and measurable functions of t and are square integrable in Ω (for every fixed value of $t \in [0, \infty]$). ∇^2 denotes the Laplace operator, $\diamondsuit^4$ is defined by the formula (1.12).

1.2. The basic equations. The assumptions suggest that the plate may be considered to be a subset of the Euclidean plane E^2, and the displacements u, v in the directions of the Cartesian coordinates x and y respectively (of E^2) are linearly varying with the distance z from the mid-plane of the plate.

The strain components are given by the usual linear approximations:

$$\epsilon_{xx} = \frac{\partial u}{\partial x} = -z\frac{\partial^2 w}{\partial x^2}, \quad \text{etc.} \tag{1.1}$$

The effects of the strain components ϵ_{xz}, ϵ_{yz}, ϵ_{zz} are going to be ignored.

We assume the correctness of Hooke's law:

$$\epsilon_{xx} = \frac{1}{E}(\tau_{xx} - \nu\tau_{yy}), \tag{1.2}$$

$$\epsilon_{xy} = \frac{2(1 + \nu)}{E}\tau_{xy}, \tag{1.3}$$

$$\epsilon_{yy} = \frac{1}{E}(\tau_{yy} - \nu\tau_{xx}). \tag{1.4}$$

The moments acting on the plate are given by the formulas:

$$M_{xx} = \int_{-h/2}^{+h/2} z \cdot \tau_{xx}\, dz = -D\left(\frac{\partial^2 w}{\partial x^2} + \nu\frac{\partial^2 w}{\partial y^2}\right), \tag{1.5a}$$

$$M_{xy} = -M_{yx} = \int_{-h/2}^{+h/2} z\tau_{xy}\, dx = -D(1 - \nu)\frac{\partial^2 w}{\partial x \partial y}, \tag{1.5b}$$

$$M_{yy} = \int_{-h/2}^{+h/2} z \cdot \tau_{yy}\, ds = -D\left(\frac{\partial^2 w}{\partial y^2} + \nu\frac{\partial^2 w}{\partial x^2}\right). \tag{1.5c}$$

The shear forces (i.e., forces normal to the plane of the plate) are expressed in terms of moments as follows:

$$Q_x = \frac{\partial M_{xx}}{\partial x} + \frac{\partial M_{xy}}{\partial y}, \tag{1.6a}$$

$$Q_y = \frac{\partial M_{yx}}{\partial x} + \frac{\partial M_{yy}}{\partial y}. \tag{1.6b}$$

Using the expressions (1.5a), (1.5b), and (1.5c) we have

$$\begin{aligned} Q_x &= \frac{\partial}{\partial x}\left[-D\left(\frac{\partial^2 w}{\partial x^2} + \nu\frac{\partial^2 w}{\partial y^2}\right)\right] + \frac{\partial}{\partial y}\left[-D(1-\nu)\frac{\partial^2 w}{\partial x \partial y}\right] \\ &= \left(-D\frac{\partial}{\partial x}\nabla^2 w\right) - \frac{\partial D}{\partial x}\left(\frac{\partial^2 w}{\partial x^2} + \nu\frac{\partial^2 w}{\partial y^2}\right) - \frac{\partial D}{\partial y}(1-\nu)\frac{\partial^2 w}{\partial x \partial y}. \end{aligned} \tag{1.7}$$

If $D = \text{const.}$ then this reduces to

$$Q_\xi = -D\frac{\partial}{\partial \xi}(\nabla^2 w). \tag{1.8}$$

The equation of equilibrium expressed in terms of the moments is

$$\frac{\partial^2 M_{xx}}{\partial x^2} - \frac{\partial^2 M_{xy}}{\partial x \partial y} + \frac{\partial^2 M_{yx}}{\partial x \partial y} + \frac{\partial^2 M_{yy}}{\partial y^2} + q = 0, \tag{1.9}$$

and substituting formulas (1.5a), (1.5b), and (1.5c) we obtain the well-known deflection equation of small deflection, thin plate theory:

$$\begin{aligned} L(w) = \frac{\partial^2}{\partial x^2}\left[D\left(\frac{\partial^2 w}{\partial x^2} + \nu\frac{\partial^2 w}{\partial y^2}\right)\right] + 2(1-\nu)\frac{\partial^2}{\partial x \partial y}\left(D\frac{\partial^2 w}{\partial x \partial y}\right) \\ + \frac{\partial^2}{\partial y^2}\left[D\frac{\partial^2 w}{\partial y^2} + \nu\frac{\partial^2 w}{\partial x^2}\right] = q. \end{aligned} \tag{1.10}$$

This can be rewritten in the form

$$\nabla^2(D\nabla^2 w) - (1-\nu)\diamondsuit^4(D, w) - q = 0, \tag{1.11}$$

$$\diamondsuit^4(A, B) \stackrel{\text{def}}{=} \frac{\partial^2 A}{\partial x^2}\frac{\partial^2 B}{\partial y^2} - 2\frac{\partial^2 A}{\partial x \partial y}\frac{\partial^2 B}{\partial x \partial y} + \frac{\partial A^2}{\partial y^2}\frac{\partial^2 B}{\partial x^2}. \tag{1.12}$$

1.3. The case of a constant cross section. In this case $h = \text{const.}$, $D = \text{const.}$ The equation (1.10) assumes the simplified form

$$\nabla^4 w = \frac{q}{D}. \tag{1.13}$$

1.4. The basic dynamic equations of small deflection theory. Since the nature of the load q has not been specified, we may assume that q is in part the inertia load, opposing the acceleration of the plate; that is, the load q consists partially of an outside load $q_0(t)$ and partially of an inertia load $-\rho(x, y)\partial^2 w/\partial t^2$, where $\rho(x, y)$ is the mass density of the plate.

The dynamic load $q(x, y, t)$ is the *control function*, or simply control.

The corresponding homogeneous equation is

$$L(w) = 0. \tag{1.14}$$

1.5. The boundary conditions.

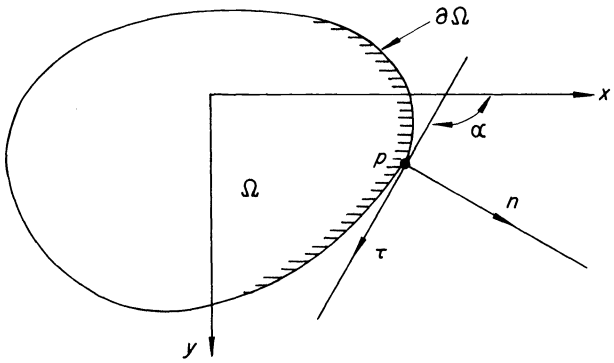


FIG. 1

Assuming that except for finitely many points of $\partial\Omega$ the outward normal unit vector n and the tangential unit vector τ are defined, then selecting a point $p \in \partial\Omega$ at which these directions exist, we define $\partial/\partial n$, $\partial/\partial\tau$ to be the directional derivatives:

$$\begin{aligned} \frac{\partial f}{\partial n} &= \frac{\partial f}{\partial x}\cos(x, n) + \frac{\partial f}{\partial y}\cos(y, n), \\ \frac{\partial f}{\partial \tau} &= \frac{\partial f}{\partial x}\cos(x, \tau) + \frac{\partial f}{\partial y}\cos(y, \tau). \end{aligned} \tag{1.15}$$

We have

$$\begin{aligned} \frac{\partial^2 w}{\partial \tau^2} &= \frac{\partial^2 w}{\partial s^2} + \kappa\frac{\partial w}{\partial n}, \\ \frac{\partial^2 w}{\partial \tau \partial n} &= \frac{\partial^2 w}{\partial s \partial n} - \kappa\frac{\partial w}{\partial s}, \end{aligned} \tag{1.16}$$

where $\partial/\partial s$ denotes the differentiation along the coordinate following the boundary $\partial\Omega$, and κ is the curvature of the boundary.

We assume here that the second derivatives in (1.16) are defined on $\partial\Omega$ possibly with the exception of finitely many points. (It is clear that $\partial w/\partial\tau \equiv dw/ds$ whenever this derivate is defined on $\partial\Omega$.)

The boundary conditions will be of either of the three types:

The clamped edge:

$$w = 0 \quad \text{and} \quad \frac{\partial w}{\partial n} = 0. \tag{B1}$$

The simply supported edge:

$$w = 0, \quad \frac{\partial^2 w}{\partial n^2} + \nu \frac{\partial^2 w}{\partial \tau^2} = 0. \tag{B2}$$

The second condition in (B2) may be replaced by

$$M_{nn} = \frac{\partial^2 w}{\partial n^2} + \nu\left(\frac{\partial^2 w}{\partial s^2} + \kappa \frac{\partial w}{\partial n}\right) = 0. \tag{B2a}$$

The free edge: The absence of external moments or forces on the free edge would require the simultaneous vanishing of M_{nn}, M_{ns}, and Q_n on the subarc of $\partial\Omega$ which is the free edge. It has been shown by Kirchhoff [22] that these three conditions are equivalent to the pair of conditions

$$\begin{aligned} M_{nn} &= 0, \\ Q_n + \frac{\partial M_{ns}}{\partial s} &= 0. \end{aligned} \tag{B3}$$

Substituting the expressions (1.5a), (1.5b), and (1.6a), we can rewrite (B3) in the form

$$\begin{aligned} &\frac{\partial^2 w}{\partial n^2} + \nu\left(\frac{\partial^2 w}{\partial s^2} + \kappa \frac{\partial w}{\partial n}\right) = 0, \\ &D\left[\frac{\partial}{\partial n}\nabla^2 w + (1-\nu)\frac{\partial}{\partial s}\left(\frac{\partial^2 w}{\partial n \partial s} - \kappa \frac{\partial w}{\partial s}\right)\right] + \frac{\partial D}{\partial n}\left[\frac{\partial^2 w}{\partial n^2} + \nu\left(\frac{\partial^2 w}{\partial s^2} + \kappa \frac{\partial w}{\partial n}\right)\right] \\ &\quad + 2(1-\nu)\frac{\partial D}{\partial s}\left(\frac{\partial^2 w}{\partial n^2} - \kappa \frac{\partial w}{\partial s}\right) = 0. \end{aligned} \tag{B3a}$$

Obvious simplifications occur in the formulas (B1), (B2a), (B3a) in the cases when either the edge is a straight line (that is $\kappa = 0$), or when the flexural rigidity of the plate $D(x, y)$ is constant in Ω. We recall that in the case of $D = \text{const.}$, the second formula (B3a) can be rewritten in the form

$$\frac{\partial}{\partial n}(\nabla^2 w) + (1-\nu)\frac{\partial}{\partial s}\left(\frac{\partial^2 w}{\partial n \partial s} - \kappa \frac{\partial w}{\partial s}\right) = 0.$$

1.6. The initial conditions. We shall consider the solutions $w(x, y, t)$ of (1.10) in the region $\Omega \subset E^2$, for values of $t \geqq 0$, where w obeys the conditions (B1), (B2), and (B3) on subarcs Γ_1, Γ_2, and Γ_3 of $\partial\Omega$, such that $\bar{\Gamma}_1 \cup \bar{\Gamma}_2 \cup \bar{\Gamma}_3 = \partial\Omega$. (Bars over the symbols indicate the closure operation.) $w(x, y, t)$ also obeys the initial conditions

$$w(x, y, 0) = \psi(x, y), \tag{C1}$$

$$\frac{\partial w(x, y, 0)}{\partial t} = \eta(x, y). \tag{C2}$$

1.7. The energy terms. If only the bending action of the plate is considered, and no energy is supplied or absorbed on the boundary, then the strain energy of

the plate is identified with the complementary energy and is given by

$$U(t) = \frac{1}{2}\int\int_{\Omega} D(x,y)\left[\left(\frac{\partial^2 w}{\partial x^2}\right)^2 + \left(\frac{\partial^2 w}{\partial y^2}\right)^2 + 2\nu\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2} + 2(1-\nu)\left(\frac{\partial^2 w}{\partial x \partial y}\right)^2\right] dx\,dy$$

$$= \frac{1}{2}\int\int_{\Omega} D[(\nabla^2 w)^2 - (1-\nu)\diamondsuit^4(w,w)]\,dx\,dy.$$

The kinetic energy of the plate is

$$T = \frac{1}{2}\int\int_{\Omega} \rho\left(\frac{\partial w}{\partial t}\right)^2 dx\,dy \tag{1.17}$$

and the total energy is

$$\mathscr{E}(t) = U(t) + T(t). \tag{1.18}$$

Since no provision was made in the assumptions for internal dissipation of energy, the total energy is assumed to be constant in the case of a free vibration.

1.8. A general discussion of the applied loads. Before proceeding with the theoretical discussion of the control principles it will be necessary to define exactly the class of functions to be considered as the admissible controls.

The most common assumptions found in the mathematical papers concerning the optimum controls of partial differential equations of the hyperbolic type induce the square integrability of the inhomogeneous term. (See, for example, [17] or [2].) In our case this would imply the square integrability of the function $q(x, y, t)$ in (1.11) as a function of x and y (t is regarded as fixed) in the region Ω:

$$\int\int_{\Omega} q(x,y,t)^2\,dx\,dy < \infty \qquad \text{for all } t \in [0,\infty). \tag{1.19}$$

An additional assumption of integrability and uniform boundedness in the L_1-norm is also included in the usual hypothesis:

$$\int\int_{\Omega} |q(x,y,t)|\,dx\,dy \leqq C \qquad \text{for all } t \in [0,\infty), \tag{1.20}$$

where C is some a priori given constant.

Without any loss of generality the inequality (1.20) may be replaced by

$$\|q\|_{\Omega} = \int\int_{\Omega} |q(x,y,t)|\,dx\,dy \leqq 1 \qquad \text{for all } t \in [0,\infty). \tag{1.20a}$$

Physically this implies that the load has been distributed over the surface of the plate in a manner determined by a bounded, measurable function of x, y: $q(x, y, t)$, and the total force exerted does not exceed unity.

A dynamic load $q(x, y, t)$ satisfying the inequalities (1.19) and (1.20a) will be called an *admissible distributed load control.* The distributed load controls may fail to contain suitable functions, required for a formulation of certain extremal problems. In fact any engineer would feel most unhappy if simple point loads were excluded from his consideration. The necessity of considering point loads or point moments also was implied by our assumptions concerning the class of solutions

$w(x, y, t)$ of the equation (1.11) which would satisfy the "elastica" hypothesis. On physical grounds we excluded solutions $w(x, y, t)$ which would be discontinuous functions of either x or y, for a fixed t (a broken plate), or would have discontinuous first derivatives $\partial w/\partial x$, $\partial w/\partial y$ for a fixed t, or for a fixed $x_0, y_0 \in \Omega$ would have a discontinuity in the derivative $\partial w/\partial t$ for some $t \in [0, \infty)$ on a set of positive measure in Ω (an infinite acceleration of some neighborhood of a point (x_0, y_0), which would imply an infinite force acting on that neighborhood). We have *not* however ruled out the possibility of discontinuities of the higher order derivatives of $w(x, y, t)$, and we have not implied that the expressions $\rho\partial^2 w/\partial t^2$, $\nabla^2(D\nabla^2 w)$ or $\diamondsuit^4(D, w)$ are defined in the usual sense in $\Omega \times [0, \infty)$, i.e., that they are measurable, locally integrable functions of x, y, t. Instead the above expressions and consequently also their linear combination $q(x, y, t)$ will be regarded as generalized derivatives of $w(x, y, t,)$ in the sense of Sobolev (see [19, § 5, pp. 39–41] for the definition).

The point loads will be represented by the Dirac delta function

$$(\delta(x - x_0, y - y_0), \psi(x, y))_\Omega \stackrel{\text{def}}{=} \psi(x_0, y_0). \tag{1.21}$$

We shall denote the product $(\delta(x - x_0, y - y_0), \psi(x, y))_\Omega$ by the symbol $\iint_\Omega \delta(x - x_0, y - y_0) \cdot \psi(x, y)\, dx\, dy$. This symbolism is commonly accepted in physics and engineering and results in an economy of notation.

Similarly we shall denote the derivatives $(\partial\delta(x - x_0, y - y_0)/\partial x, \psi(x, y))$ by $\int\int_\Omega (\partial\delta(x - x_0, y - y_0)/\partial x \cdot \psi(x, y)\, dx\, dy$, etc.

Let N_ε denote an ε-neighborhood of $\partial\Omega$ for some $\varepsilon > 0$. In analogy with the admissible distributed load control we define the *admissible concentrated load* (or point load) control to satisfy the inequality

$$\|q\|_\Omega = \limsup \iint_{\Omega - N_\varepsilon} |q(x, y, t)|\, dx\, dy \equiv \limsup_{\varepsilon > 0} (|q(x, y, t)|, \mathbf{1})_{\Omega - N_\tau} \leqq 1. \tag{1.22}$$

Here $\mathbf{1}$ denotes the test function $\psi(x, y, t) \equiv 1$ for all $x, y \in \Omega - N_\varepsilon(\partial\Omega)$ and any fixed $t \in [0, \infty)$. Point load controls ($\psi(x, y, t) \equiv 0$ outside of Ω) will be considered of the form $\sum_{i=1}^N \delta(x - \xi_i(t), y - \eta_i(t))\phi_i(t)$, where $\xi_i(t)$, $\eta_i(t)$ are functions whose domain is $[0, \infty)$ and whose range lies in Ω; i.e., for any $t \in [0, \infty)$ the pair $(\xi_i(t), \eta_i(t))$ can be identified with the x, y-coordinates of a point in Ω. In addition the functions $\phi_i(t)$ are bounded and measurable on $[0, \infty)$ obeying

$$\sum_{i=1}^N |\phi_i(t)| \leqq 1.$$

Hence our controls are allowed to be distributions in the sense of Schwarz in $\Omega(x, y)$ but are assumed to be measurable functions of t.

To summarize our discussion we offer this definition: A bounded linear functional $f(x, y, t)$ over the space of test functions obeying the elastica hypothesis will be called an *admissible control*, if $\|f(x, y, t)\|_\Omega \leqq 1$; that is, the usual norm of the functional f over the domain Ω is bounded above by 1, and $F(t) = \|f(x, y, t)\|_\Omega$ is a measurable function of the time variable t. Since $F(t)$ is measurable and

uniformly bounded on the finite interval $[0, \tau]$ it is a square integrable and also an absolutely integrable function of t. This immediately implies that the total energy of the plate is also uniformly bounded on $[0, \tau]$.

1.9. The statement of Duhamel's principle. In what follows we shall assume without proof the existence and uniqueness of the solutions of the mixed boundary value and initial value problem (MBVP), i.e., of the problem posed by (1.11) with suitable boundary conditions of the form (B1), (B2), (B3) prescribed on $\partial\Omega$, and with initial conditions (C1) and (C2) and with an admissible control $q(x, y, t)$ given a priori. Such a system will be denoted in the future by MBVP. Let $w_H(x, y, t)$ denote the solution of the homogeneous MBVP (i.e., it corresponds to the case when $q(x, y, t) \equiv 0$), with the appropriate boundary value conditions of the form (B1), (B2), (B3), and the initial conditions of the form (C1), (C2). Let $q(x, y, t)$ be a distributed load control function. Then Duhamel's principle asserts that the solution of MBVP, $w(x, y, t)$, will be given by the formula

$$\begin{aligned} w(x, y, t) &= w_H(x, y, t) + \int_0^t \int_\Omega G((x - \xi), (y - \eta), (t - \tau)) q(\xi, \eta, \tau)\, d\xi\, d\eta\, d\tau \\ &= w_H + G * q, \end{aligned} \tag{1.23}$$

where $G(x - \xi, y - \eta, t - \tau)$ depends only on (1.11) and on the boundary conditions, but does not depend on the initial conditions or on the function $q(x, y, t)$. The symbol $*$ denotes the convolution operation.

The proofs of Duhamel's principle for linear systems of differential equations can be found in textbooks. The usual proof (see, for example, Yosida [24, pp. 76–80]) utilizes the Ascoli–Arzela theorem, and clearly is not applicable in the case when $q(x, y, t)$ is a point load; however, it is not hard to extend the proof to cover this case as well, by considering the point load to be the limit of a δ-convergent sequence.

1.10. Some identities arising from (1.11). Following our initial assumptions, the total energy is given by the sum of strain energy, kinetic energy, and the potential energy due to the effects of the boundary forces, where

$$U_B = \int_{\partial\Omega} \left(Q_n w - M_{nn} \frac{\partial w}{\partial n} - M_{ns} \frac{\partial w}{\partial s} \right) ds, \tag{1.24}$$

will denote the potential energy due to the constraint forces applied to the boundary $\partial\Omega$. The total energy is given by

$$\begin{aligned} \mathscr{E}(t) = \frac{1}{2} \iint_\Omega & D(\nabla^2 w)^2 - D(1 - \nu) \diamondsuit^4(w, w)\, dx\, dy \\ & + \int_{\partial\Omega} \left(Q_n w - M_{nn} \frac{\partial w}{\partial n} - M_{ns} \frac{\partial w}{\partial} \right) ds + \frac{1}{2} \iint_\Omega \rho \left(\frac{\partial w}{\partial t} \right)^2 dx\, dy. \end{aligned} \tag{1.25}$$

Differentiating both sides of (1.25) with respect to the time variable, we obtain

$$\frac{d\mathscr{E}(t)}{dt} = \iint_{\Omega} \left\{ D[(\nabla^2 w)(\nabla^2 v) - (1 - \nu)\diamond^4(w, v)] + \rho v \frac{\partial^2 w}{\partial t^2} \right\} dx\, dy$$

(1.26)

$$+ \frac{\partial}{\partial t} \int_{\partial\Omega} \left(Q_n w - M_{nn} \frac{\partial w}{\partial n} - M_{ns} \frac{\partial w}{\partial s} \right) ds$$

(as before, $v = \partial w(x, y, t)/\partial t$).

With regard to the last term of the formula (1.26), we note that it vanishes identically in the case of a free edge, since then $Q_n = M_{nn} = M_{ns} \equiv 0$ on $\partial\Omega$, and that it also vanishes in the case of a simply supported edge, since then $w = \partial w/\partial s \equiv 0$, and $M_{nn} \equiv 0$ on $\partial\Omega$. In the only remaining case, i.e., the case of a fixed edge, we have $w = \partial w/\partial s \equiv 0$ and $\partial w/\partial n = 0$, and again the entire expression is identically equal to zero on the boundary $\partial\Omega$.

We rewrite (1.26) accordingly:

(1.26a)
$$\frac{d\mathscr{E}}{dt} = \iint_{\Omega} \left\{ D[(\nabla^2 w)(\nabla^2 v) - (1 - \nu)\diamond^4(w, v)] + \rho v \frac{\partial^2 w}{\partial t^2} \right\} dx\, dy.$$

In the entire discussion, which follows, we shall consider only the cases when the boundary conditions are of the (B1), (B2), or (B3) type and therefore when $Q_n w = M_{nn}\partial w/\partial n = M_{ns}\partial w/\partial s \equiv 0$ on $\partial\Omega$. We manipulate the formulas (1.25) and (1.26) by using the Green's identity:

$$\iint_{\Omega} D(\nabla^2 w)(\nabla^2 w)\, dx\, dy = \iint_{\Omega} \nabla^2(D\nabla^2 w)\, dx\, dy + \int_{\partial\Omega} \left\{ D\nabla^2 w \frac{\partial w}{\partial n} \right\} ds$$

$$- \int_{\partial\Omega} \left\{ w \frac{\partial}{\partial n}(D\nabla^2 w) \right\} ds.$$

We recall also the formulas for the moments and shears:

$$M_{nn} + M_{ss} = -D(1 + \nu)\left(\frac{\partial^2 w}{\partial n^2} + \frac{\partial^2 w}{\partial \tau^2} \right) = -D(1 + \nu)\nabla^2 w.$$

Hence, $D\nabla^2 w = -(M_{nn} + M_{ss})/(1 + \nu)$.

We shall need the following formula:

$$\iint_{\Omega} D(\nabla^2 w)(\nabla^2 w)\, dx\, dy = \iint_{\Omega} w\nabla^2(D\nabla^2 w)\, dx\, dy - \frac{1}{1 + \nu} \int_{\partial\Omega} \left\{ (M_{nn} + M_{ss}) \frac{\partial w}{\partial n} \right\} ds$$

$$+ \frac{1}{1 + \nu} \int_{\partial\Omega} \left\{ w \frac{\partial}{\partial n}(M_{nn} + M_{ss}) \right\} ds.$$

In the case of the free edge, $M_{nn} \equiv 0$ on $\partial\Omega$. In the case of a simply supported edge, $w \equiv 0$ on $\partial\Omega$ and the second boundary integral vanishes, and finally in the case

of a fixed edge, i.e., $\partial w/\partial n \equiv 0$ and $w \equiv 0$ on $\partial\Omega$, we have

$$\iint_\Omega D(\nabla^2 w)(\nabla^2 w)\,dx\,dy \equiv \iint_\Omega w\nabla^2(D\nabla^2 w)\,dx\,dy.$$

Following these remarks we rewrite the formula (1.26):

$$\begin{aligned}\frac{d\mathscr{E}}{dt} &= \frac{1}{2}\left[\frac{\partial}{\partial t}\iint_\Omega \{w\nabla^2(D\nabla^2 w) - D(1-\nu)\diamondsuit^4(w,w)\}\,dx\,dy\right] \\ &\quad - \iint_\Omega \left(qv + w\frac{\partial q}{\partial t}\right)dx\,dy + \frac{1}{2}\iint_\Omega v\cdot\rho\frac{\partial^2 w}{\partial t^2}\,dx\,dy \\ &\quad - \frac{1}{1+\nu}\int_{\partial\Omega}\left(\chi\frac{\partial w}{\partial n} + w\frac{\partial\chi}{\partial n}\right)ds,\end{aligned} \tag{1.27a}$$

where $v \equiv dw/dt$ and χ denotes the invariant quantity $M_{yy} + M_{xx}$. (Note: The proof that $M_{xx} + M_{yy} = M_{nn} + M_{ss}$ for any orthogonal coordinates following the directions n, s comes directly from the well-known fact that the Laplacian $\nabla^2 w$ is invariant under any orthogonal transformation of coordinates.) We use the fact that $w(x, y, t)$ obeys the equation

$$v\left(\rho\cdot\frac{\partial^2 w}{\partial t^2}\right) = v[q - \nabla^2(D\nabla^2 w) + (1-\nu)\diamondsuit^4(D, w)]$$

and obtain

$$\begin{aligned}\frac{d\mathscr{E}}{dt} &= \frac{1}{2}\iint_\Omega w\cdot\nabla^2(D\nabla^2 v) - 2D(1-\nu)\diamondsuit^4(w,v)\,dx\,dy - \iint_\Omega (qv)\,dx\,dy \\ &\quad + \frac{1-\nu}{2}\iint_\Omega \diamondsuit^4(D,v)\,dx\,dy - \frac{1}{1+\nu}\frac{\partial}{\partial t}\int_{\partial\Omega}\left(\chi\frac{\partial w}{\partial n} + w\frac{\partial\chi}{\partial n}\right)ds.\end{aligned} \tag{1.27b}$$

We shall be interested in various modes of optimal control, or of excitation, and one of our main problems is to find q, such that either $-d\mathscr{E}/dt$ is maximized in some interval $[0, 1]$, that $\mathscr{E}$ assumes some value in the shortest possible time, or that $\mathscr{E}$ assumes an extreme value at some given time $t = \tau$. In all cases the formula (1.27b) is crucial.

We introduce the following inner product of two admissible transverse displacement functions:

$$\begin{aligned}\langle w_1, w_2\rangle &= \frac{1}{2}\iint_\Omega [D\nabla^2 w_1\cdot\nabla^2 w_2 - (1-\nu)D\diamondsuit^4(w_1, w_2)]\,dx\,dy \\ &\quad + \frac{1}{2}\iint_\Omega \rho\frac{\partial w_1}{\partial t}\frac{\partial w_2}{\partial t}\,dx\,dy.\end{aligned} \tag{1.28}$$

We see in particular that $\langle w_1, w_1\rangle = U_1(t) + T_1(t) = \mathscr{E}_1(t)$, and $d\langle w, w\rangle/dt = d\mathscr{E}/dt$.

We need to check the fact that $\langle w_1, w_2\rangle$ does indeed satisfy all axiomatic requirements of an inner product. That is a routine exercise, and it will be omitted. From the fact that $\langle w_1, w_2\rangle$ is an inner product, follows immediately the Cauchy–Schwarz inequality

$$\langle w_1, w_2\rangle^2 \leqq \langle w_1, w_1\rangle \cdot \langle w_2, w_2\rangle = \mathscr{E}_1\mathscr{E}_2,$$

valid for all $t \in [0, T]$. We note that the exact equality $\langle w_1, w_2\rangle^2 = \mathscr{E}_1\mathscr{E}_2$ is true only if there exists a constant C, such that $\nabla^2 w_1 = C\nabla^2 w_2$, $\partial w_1/\partial t = C\partial w_2/\partial t$, and $\diamondsuit^4(w_1, w_1) = C\diamondsuit^4(w_2, w_2)$, valid for all $t \in [0, T]$. We compute the time rate change of this product:

$$\begin{aligned}\frac{d}{dt}\langle w_1, w_2\rangle = \frac{1}{2}\iint_\Omega \Big\{ & D[\nabla^2 w_1 \cdot \nabla^2 v_2 + \nabla^2 w_2 \nabla^2 v_1] \\ & - D(1-\nu)[\diamondsuit^4(v_1, w_2) + \diamondsuit^4(v_2, w_1)] \\ & + \rho\left[v_1 \frac{\partial^2 w_2}{\partial t^2} + v_2 \frac{\partial^2 w_1}{\partial t^2}\right]\Big\}\, dx\, dy.\end{aligned} \tag{1.29}$$

We use Green's identity

$$\begin{aligned}\iint_\Omega D\nabla^2 w_1 \cdot \nabla^2 v_2\, dx\, dy = & \iint_\Omega v_2 \cdot (\nabla^2 D \nabla^2 w_1)\, dx\, dy \\ & + \int_{\partial\Omega} \left\{ D\nabla^2 w_1 \cdot \frac{\partial v_2}{\partial n} - v_2 \cdot \frac{\partial}{\partial n}(D\nabla^2 w_1)\right\} ds\end{aligned}$$

and obtain

$$\begin{aligned}\frac{d}{dt}\langle w_1, w_2\rangle = \frac{1}{2}\iint_\Omega \Big\{ & v_2\left[\nabla^2(D\nabla^2 w_1) + \rho\frac{\partial w_1}{\partial t^2} - (1-\nu)\diamondsuit^4(D, w_1)\right] \\ & + v_1\left[\nabla^2(D\nabla^2 w_2) + \rho\frac{\partial^2 w_2}{\partial t^2} - (1-\nu)\diamondsuit^4(D, w_2)\right]\Big\}\, dx\, dy \\ & + \frac{1}{2}\iint_\Omega (1-\nu)[v_2\diamondsuit^4(D, w_1) - D\diamondsuit^4(v_2, w_1) \\ & \qquad + v_1\diamondsuit^4(D, w_2) - D\diamondsuit^4(v_1, w_2)]\, dx\, dy \\ & + \frac{1}{2}\int_{\partial\Omega}\Big\{ D\nabla^2 w_1 \frac{\partial v_2}{\partial n} - v_2\frac{\partial}{\partial n}(D\nabla^2 w_1) \\ & \qquad + D\nabla^2 w_2 \frac{\partial v_1}{\partial n} - v_1\frac{\partial}{\partial n}(D\nabla^2 w_2)\Big\}\, ds.\end{aligned} \tag{1.29a}$$

In the case when $D = \text{const.}$ the above formula reduces to

$$\begin{aligned}\frac{d}{dt}\langle w_1, w_2\rangle &= \frac{1}{2}\iint_\Omega \Big\{ v_2\Big[\nabla^2(D\nabla^2 w_1) + \rho\frac{\partial^2 w_1}{\partial t^2}\Big] \\ &\quad + v_1\Big[\nabla^2(D\nabla^2 w_2) + \rho\frac{\partial^2 w_2}{\partial t^2}\Big] \\ &\quad - D(1-\nu)[\diamond^4(v_2, w_1) + \diamond^4(v_1, w_2)]\Big\}\, dx\, dy \\ &\quad + \frac{1}{2}\int_{\partial\Omega}\Big\{ D(\nabla^2 w_1)\frac{\partial v_2}{\partial n} + D(\nabla^2 w_2)\frac{\partial v_1}{\partial n} + v_2 Q_{n_1} + v_1 Q_{n_2}\Big\}\, ds \qquad \text{(1.29b)} \\ &= \frac{1}{2}\iint_\Omega \Big\{(v_2 q_1 + v_1 q_2) - D(1-\nu)\frac{d}{dt}[\diamond^4(w_1, w_2)]\Big\}\, dx\, dy \\ &\quad + \frac{1}{2}\int_{\partial\Omega}\Big\{ D\Big[(\nabla^2 w_1)\frac{\partial v_2}{\partial n} + (\nabla^2 w_2)\frac{\partial v_1}{\partial n}\Big] + v_2 Q_{n_1} \\ &\qquad\qquad + v_1 Q_{n_2}\Big\}\, ds.\end{aligned}$$

However, we note that if $D = \text{const.}$, then $\iint D\diamond^4(w_1, w_2)\, dx\, dy = 0$ for any displacement functions w_1, w_2 which solve the basic equation (1.11). We use the well-known fact that $\iint_\Omega D\diamond^4(w, w)\, dx\, dy = 0$ if $\diamond^4(D, w) = 0$. (See, for example, [9, (6.5), p. 80 and the discussion of § 1.4].)

To prove our claim we first observe that when $D = \text{const.}$, we have

$$\begin{aligned}\iint_\Omega \diamond^4(w_1, w_2)\, dx\, dy &= \iint_\Omega \{\diamond^4(w_1, w_2) + \tfrac{1}{2}\diamond^4(w_1, w_1) + \tfrac{1}{2}\diamond^4(w_2, w_2)\}\, dx\, dy \\ &= \iint_\Omega \Big\{\frac{\partial^2 w_1}{\partial x^2}\frac{\partial^2 w_2}{\partial y^2} + \frac{\partial^2 w_2}{\partial x^2}\frac{\partial^2 w_1}{\partial y^2} + 2\frac{\partial^2 w_1}{\partial x\,\partial y}\frac{\partial^2 w_2}{\partial x\,\partial y} \\ &\quad + \frac{\partial^2 w_1}{\partial x^2}\frac{\partial^2 w_1}{\partial y^2} - \Big(\frac{\partial^2 w_1}{\partial x\,\partial y}\Big)^2 + \frac{\partial^2 w_2}{\partial x^2}\frac{\partial^2 w_2}{\partial y^2} - \Big(\frac{\partial^2 w_2}{\partial x\,\partial y}\Big)^2\Big\}\, dx\, dy \\ &= \iint_\Omega \Big\{\Big(\frac{\partial^2 w_1}{\partial x^2} + \frac{\partial^2 w_2}{\partial x^2}\Big)\Big(\frac{\partial^2 w_1}{\partial y^2} + \frac{\partial^2 w_2}{\partial y^2}\Big) \\ &\qquad - \Big[\frac{\partial^2(w_1 + w_2)}{\partial x\,\partial y}\Big]^2\Big\}\, dx\, dy \\ &= \iint_\Omega \det\begin{bmatrix} \dfrac{\partial^2(w_1+w_2)}{\partial x^2} & \dfrac{\partial^2(w_1+w_2)}{\partial x\,\partial y} \\ \dfrac{\partial^2(w_1+w_2)}{\partial x\,\partial y} & \dfrac{\partial^2(w_1+w_2)}{\partial y^2}\end{bmatrix} dx\, dy.\end{aligned}$$

Hence if we denote by w_3 the displacement $w_3 = (w_1 + w_2)/2$, we can easily obtain

$$\iint_\Omega \det(A(w_1 + w_2))\,dx\,dy = \iint_\Omega \det(A(w_3 + w_3))\,dx\,dy$$
$$= \iint_\Omega \diamond^4(w_3, w_3)\,dx\,dy = 0,$$

and consequently

$$\iint_\Omega \diamond^4(w_1, w_2)\,dx\,dy = 0$$

for any w_1, w_2 which solve (1.11).

The matrix operator A in the above manipulation stood for

$$A = \begin{bmatrix} \dfrac{\partial^2}{\partial x^2} & \dfrac{\partial^2}{\partial x \partial y} \\ \dfrac{\partial^2}{\partial x \partial y} & \dfrac{\partial^2}{\partial y^2} \end{bmatrix}.$$

This result enables us to simplify our formula (1.29b) in the case $D = \text{const.}$, whereupon we obtain:

$$\frac{d}{dt}\langle w_1, w_2\rangle = \frac{1}{2}\iint_\Omega (v_2 q_1 + v_1 q_2)\,dx\,dy$$
$$+\frac{1}{2}\int_{\partial\Omega}\left\{D\left[(\nabla^2 w_1)\frac{\partial v_2}{\partial n} + (\nabla^2 w_2)\frac{\partial v_1}{\partial n}\right] + v_2 Q_{n_1} + v_1 Q_{n_2}\right\} ds.$$

An identical result is obtained if we assume that $\diamond^4(D, w) = 0$ for any $w(x, y, t)$ which is a solution of (1.11).

1.11. The case of $\diamond^4(D, w) \equiv 0$. The above expressions and formulas can be greatly simplified if $\diamond^4(D, w) \equiv 0$, which is true in the physically important cases when $D \equiv \text{const.}$ in Ω, or when D depends linearly on x and y. The second case occurs in the optimum weight design of plates. The equation (1.11) becomes

$$\nabla^2(D\nabla^2 w) + \rho\frac{\partial^2 w}{\partial t^2} = q, \tag{1.11a}$$

and if $D = \text{const.}$, this becomes

$$\nabla^4 w + \frac{\rho}{D}\frac{\partial^2 w}{\partial t^2} = \frac{q}{D}. \tag{1.11b}$$

The expression for the strain energy is

$$U = \frac{1}{2}\iint_\Omega D(\nabla^2 w)^2\,dx\,dy. \tag{1.30}$$

(A variational argument for this statement also follows easily. See for example [9, pp. 79–82], or [12].)

A similar conclusion is reached in the case when $\diamond^4(D, w) \equiv 0$ even if $D \not\equiv$ const. in Ω. The product $\langle w_1, w_2 \rangle$ assumes the form

$$\langle w_1, w_2 \rangle = \frac{1}{2}\iint_\Omega D(\nabla^2 w_1)(\nabla^2 w_2)\,dx\,dy + \frac{1}{2}\iint_\Omega \rho \frac{\partial w_1}{\partial t}\frac{\partial w_2}{\partial t}\,dx\,dy. \tag{1.28a}$$

The rate of change of this product is given below:

$$\frac{d}{dt}\langle w_1, w_2 \rangle = \iint_\Omega (v_1 q_2 + v_2 q_1)\,dx\,dy - \frac{1}{2}\int_{\partial\Omega} \left(v_1 \frac{\partial}{\partial n}(D\nabla^2 w_2) + v_2 \frac{\partial}{\partial n}(D\nabla^2 w_1) - D\nabla^2 w_2 \frac{\partial v_1}{\partial n} - D\nabla^2 w_1 \frac{\partial v_2}{\partial n} \right) ds. \tag{1.29c}$$

We substitute $\chi_1 = M_{xx_1} + M_{yy_1} = -D(1 + \nu)\nabla^2 w_1$ and $\chi_2 = -D(1 + \nu)\nabla^2 w_2$ to obtain

$$\frac{d}{dt}\langle w_1, w_2 \rangle = \frac{1}{2}\iint_\Omega (v_1 q_2 + v_2 q_1)\,dx\,dy - \frac{1}{2(1 + \nu)}\int_{\partial\Omega} \left(\chi_2 \frac{\partial v_1}{\partial n} + \chi_1 \frac{\partial v_2}{\partial n} - v_1 \frac{\partial \chi_2}{\partial n} - v_2 \frac{\partial \chi_1}{\partial n} \right) ds. \tag{1.29c'}$$

If D = const., we can rewrite (1.29c') in the form

$$\frac{d}{dt}\langle w_1, w_2 \rangle = \frac{1}{2}\iint_\Omega (v_1 f_2 + v_2 f_1)\,dx\,dy + \frac{1}{2}\int_{\partial\Omega} \left[v_1 Q_{n_2} + v_2 Q_{n_1} - \frac{1}{1 + \nu}\left(\chi_1 \frac{\partial v_2}{\partial n} + \chi_2 \frac{\partial v_1}{\partial n} \right) \right] ds, \tag{1.29d}$$

where Q_{n_i} are the shear forces which are related to the moments by (1.7).

We note that in the case of the clamped edge (condition (B1)) (1.29d) reduces to:

$$\frac{d}{dt}\langle w_1, w_2 \rangle = \frac{1}{2}\iint_\Omega (v_1 q_2 + v_2 q_1)\,dx\,dy, \tag{1.29e}$$

since in this case $v_1 = v_2 = \partial v_1/\partial n = \partial v_2/\partial n \equiv 0$ on $\partial\Omega$, and the contour integral vanishes.

The next result will be used in proving the basic theorem, Theorem 2.2. We shall state it as a lemma.

LEMMA 1.1. *Let $f(x, y, t)$ be an admissible control and $w(x, y, t)$ be the corresponding deflection of a plate, whose flexural rigidity D and density ρ are constant. Let w_H represent the solution of the homogeneous equation* (1.11). *Let both $w(x, y, t)$ and $w_H(x, y, t)$ satisfy the condition $w = 0$ on $\partial\Omega$ and $w_H = 0$ on $\partial\Omega$ (the boundary*

of the plate). *Then*

$$\frac{d}{dt}\langle w, w_H\rangle = \frac{1}{2}\int_\Omega (w_H f)\,dx\,dy + \frac{D}{2}\int_{\partial\Omega}\left(\nabla^2 w\frac{\partial v_H}{\partial n} + \nabla^2 w_H\frac{\partial v}{\partial n}\right)ds,$$

where, as before,

$$v = \frac{\partial w}{\partial t}, \qquad v_H = \frac{\partial w_H}{\partial t}.$$

The proof follows from the formula (1.29b) upon substituting $q_1 = q$, $q_2 \equiv 0$, $w_1 = w$, $w_2 = w_H$ and from the observation that

$$\iint_\Omega \diamond^4(w, w_H)\,dx\,dy = 0.$$

2. Optimal control principles for small deflection theory in mixed boundary and initial value problems for thin plates.

2.1. Statement of the control problems.

(a) *The minimal time control problem.* We consider (1.11) and the corresponding MBVP. The initial conditions (C1) and (C2) determine the initial value of the total energy $\mathscr{E}_0 = \mathscr{E}(0)$. Given a real number $0 \leqq \hat{\mathscr{E}} < \mathscr{E}_0$ find an admissible control $q(x, y, t)$ such that the total energy of the vibrating plate $\mathscr{E}(t)$ is reduced to the value $\hat{\mathscr{E}}$ in the shortest possible time.

(b) *A similar problem is defined below as the minimal time excitation problem.* Again the initial conditions determine the initial value of the total energy $\mathscr{E}_0$. We allow $\mathscr{E}_0 \geqq 0$. Given a real number $\hat{\mathscr{E}} > \mathscr{E}_0$ find an admissible control $q(x, y, t)$ such that the total energy of the vibrating plate $\mathscr{E}(t)$ is raised to the value $\hat{\mathscr{E}}$ in the shortest possible time.

(c) *The fixed time interval optimal control* (*or the optimal interval control*). Given suitable boundary and initial conditions of the MBVP, and given a time interval $[0, T]$, find an admissible control $\tilde{q}(x, y, t)$, such that the total energy of the plate is reduced (raised) to the lowest (highest) possible level at the time $t = T$, i.e., $\mathscr{E}(\tilde{q}(x, y, t), T) \leqq \mathscr{E}(q(x, y, t), T)$ for any admissible control $q(x, y, t)$ (or in the excitation problem $\mathscr{E}(\tilde{q}(x, y, t), T) \geqq \mathscr{E}(q(x, y, t), T)$ for any admissible control $q(x, y, t)$).

Remark. Optimal controls are generally not unique. However, some form of nonuniqueness turns out to be acceptable, but other forms will defeat the whole idea of a meaningful control. (See [7] for an analogous discussion.)

We consider the MBVP with suitable initial and boundary conditions.

LEMMA 2.1. *Let $\tilde{q}(x, y, t)$ be an optimal time control reducing the total energy from the initial value $E(0)$ to a given value $0 < \hat{\mathscr{E}} < E(0)$ in the shortest possible time $T > 0$. Then $\tilde{q}(x, y, t)$ is also an optimal interval control for the fixed interval $[0, T]$.*

(See [17] for the proof.)

Lemma 2.1, above, implies that it will suffice to develop the control principles for the fixed interval case, since any optimal time control will also be a fixed interval optimal control and the validity of our results will be preserved.

2.2. Some mathematical preliminaries.

LEMMA 2.2. *Let $\mathscr{R}$ denote the class of functions $u(x, y, t)$ and of the time derivatives $\partial u/\partial t$ of such functions obeying the "elastica" hypothesis in Ω, and satisfying the condition that either the displacement $w(x, y, t)$ or the moment $M_{nn}(w(x, y, t))$ vanishes on $\partial\Omega$ for all $t \in [0, \infty)$, where M_{nn} is defined by* (B2). *Let $\mathscr{R}^*$ denote the space of all continuous linear functionals mapping elements of $\mathscr{R}$ into the real line. We shall consider only a subset $\hat{\mathscr{R}}^* \subset \mathscr{R}^*$ of all such functionals f obeying $\|f\|_\Omega \leqq 1$, where $\|f\|_\Omega = \sup_{\|u\|=1} |(f, u)_\Omega|$. (If $\sup_{\|u\|_\Omega=1} |(f, u)|$ does not exist we assign $\|f\|_\Omega = \infty$.) Then we assert that the space $\mathscr{R}^*$ is complete.*

The proof of this lemma is more elementary but similar to the lemma stated in [5, Appendix A, pp. 368–369].

2.3. The basic convexity lemma.

LEMMA 2.3. *The set of admissible controls is convex (i.e., if f_1, f_2 are admissible controls then $\lambda f_1 + (1 - \lambda) f_2$ is also an admissible control for any $0 \leqq \lambda \leqq 1$).*

Combining the result of this lemma with Duhamel's principle we obtain an important corollary.

COROLLARY. *The set of all admissible transverse displacements is convex (i.e., if $w_1(x, y, t)$ is an admissible transverse displacement corresponding to an admissible control f_1, and if $w_2(x, y, t)$ is an admissible displacement corresponding to an admissible control f_2, then $w = \lambda w_1 + (1 - \lambda) w_2$ is also an admissible displacement for any $0 \leqq \lambda \leqq 1$. Of course, w_1, w_2 are assumed to be solutions obeying the stated boundary and initial conditions of the MBVP).*

2.4. The uniqueness of the finite state.

LEMMA 2.4. *Let us assume that no energy is transmitted at the boundary. Let $f_1(x, y, t)$, $f_2(x, y, t)$ be two admissible controls, which are optimal controls for the $[0, \tau]$ fixed time interval. Then the corresponding shapes of the plate, and the corresponding velocities coincide at the time $t = \tau$, i.e.,*

$$w_1(f_1, x, y, \tau) = w_2(f_2, x, y, \tau)$$

and

$$\frac{\partial w_1(f_1, x, y, \tau)}{\partial t} = \frac{\partial w_2(f_2, x, y, \tau)}{\partial t}.$$

Proof. Let us assume to the contrary that $\underset{\sim}{w}_1 \neq \underset{\sim}{w}_2$ at the time $t = \tau$, where

$$\underset{\sim}{w} \overset{\text{def}}{=} \begin{cases} w(x, y, t), \\ \dfrac{\partial w}{\partial t}(x, y, t). \end{cases}$$

Let us denote by $\tilde{\mathscr{E}}(\tau)$ the lowest value of total energy attainable at the time τ. By the convexity of the set of admissible displacements we conclude that $\underset{\sim}{w}(x, y, t) = \frac{1}{2}(w_1 + w_2)$ is also an admissible displacement.

The corresponding total energy at the time $t = \tau$ is given by (1.25):

$$\mathscr{E}(w, \tau) = U(\tau) + T(\tau) = \frac{1}{2} \iint_\Omega \{\tfrac{1}{4} D(\nabla^2 w_1 + \nabla^2 w_2)^2 - \tfrac{1}{4} D(1 - \nu) \diamondsuit^4 (w_1 + w_2, w_1 + w_2)\} \, dx \, dy$$

$$+\frac{1}{2}\iint_{\Omega}\frac{1}{4}\rho\left(\frac{\partial(w_1+w_2)}{\partial t}\right)^2 dx\,dy$$

$$=\frac{1}{8}\left\{\iint_{\Omega}\left[D(\nabla^2 w_1)^2+D(\nabla^2 w_2)^2-D(1-\nu)\diamondsuit^4(w_1,w_1)\right.\right.$$

$$-D(1-\nu)\diamondsuit^4(w_2,w_2)+\rho\left(\frac{\partial w_1}{\partial t}\right)^2+\rho\left(\frac{\partial w_2}{\partial t}\right)^2\Big]\,dx\,dy$$

$$\left.-2\iint_{\Omega}\left[D(\nabla^2 w_1)(\nabla^2 w_2)-D(1-\nu)\diamondsuit^4(w_1,w_2)+\rho\frac{\partial w_1}{\partial t}\frac{\partial w_2}{\partial t}\right]dx\,dy\right\}$$

$$=\tfrac{1}{8}\{4\tilde{\mathscr{E}}(\tau)+4\langle w_1,w_2\rangle_{t=\tau}\}=\tfrac{1}{2}\tilde{\mathscr{E}}(\tau)+\tfrac{1}{2}\langle w_1,w_2\rangle_{t=\tau}.$$

By the Cauchy–Schwarz inequality, $\langle w_1,w_2\rangle \leqq \sqrt{\langle w_1,w_1\rangle\langle w_2,w_2\rangle}$. Since $\langle w_1,w_1\rangle_{t=\tau}=\langle w_2,w_2\rangle_{t=\tau}=\tilde{\mathscr{E}}(\tau)$ we obtain

$$\mathscr{E}(w,\tau)=\tfrac{1}{2}\tilde{\mathscr{E}}(\tau)+\tfrac{1}{2}\langle w_1,w_2\rangle_{t=\tau}\leqq\tfrac{1}{2}\tilde{\mathscr{E}}(\tau)+\tfrac{1}{2}\tilde{\mathscr{E}}(\tau)=\tilde{\mathscr{E}}(\tau).$$

However $\tilde{\mathscr{E}}(\tau)$ was the lowest possible level of total energy attainable at the time $t=\tau$. Hence we must have the strict equality

$$\mathscr{E}(w,\tau)=\tilde{\mathscr{E}}(\tau).$$

This equality means that

$$\langle w_1,w_2\rangle^2=\langle w_1,w_1\rangle\langle w_2,w_2\rangle$$

which implies that $\underset{\sim}{w}_1=\alpha\underset{\sim}{w}_2$ where α is some constant; but the only suitable constant turns out to be $\alpha=1$, and the uniqueness of the final condition is established.

2.4.1. Discussion of the uniqueness of the optimal excitation. We observe that the above arguments fail in the case of the optimum excitation of a plate. Again let us denote by $\tilde{\mathscr{E}}(\tau)$ the greatest level of energy attainable at the time $t=\tau$ following an admissible excitation of the plate. Let $\tilde{f}(x,y,t)$ be an optimal excitation (assuming that it exists) and $\tilde{w}$ be the corresponding transverse displacement, $\tilde{w}=\tilde{w}(\tilde{f})$. Then if two such optimal admissible excitation functions exist, say $\tilde{f}_1,\tilde{f}_2$, so that $\mathscr{E}(\tilde{f}_1,\tau)=\mathscr{E}(\tilde{f}_2,\tau)=\tilde{\mathscr{E}}$, then $\Lambda f_1+(1-\Lambda)f_2$ is again an admissible excitation, with $\underset{\sim}{w}=\Lambda\underset{\sim}{w}_1+(1-\Lambda)\underset{\sim}{w}_2$ being the corresponding displacement.

Then $\mathscr{E}(w,\tau)=\frac{1}{2}\mathscr{E}(\tau)+\frac{1}{2}\langle w_1,w_2\rangle_{t=\tau}\leqq\mathscr{E}(\tau)$, if we substitute $\Lambda=\frac{1}{2}$. And the strict inequality must be true if $w_1\neq\underset{\sim}{w}_2$ at the time τ, since then $\langle w_1,w_2\rangle<\tilde{\mathscr{E}}$. Hence if f_1 and f_2 are optimal excitations, $\frac{1}{2}(f_1+f_2)$ cannot be optimal if $\underset{\sim}{w}_1(\tau)\neq\underset{\sim}{w}_2(\tau)$. In fact, $\Lambda f_1+(1-\Lambda)f_2$ cannot be optimal if $w_1(\tau)\neq w_2(\tau)$ for any $0<\Lambda<1$. This lack of convexity of the set of optimal excitations prevents us from following an identical argument, and reaching a conclusion analogous to the optimal control case.

2.5. Pontryagin's principle. The formulation of Pontryagin's principle for thin plates with uniform rigidity and density, as given in § 2.5.1 of this paper, is in complete analogy with the corresponding formulation for the vibrating beam

as given by the author in [7]. The more complex formulation given in § 2.5.4 and § 2.5.5 results from the presence of the terms of the form $\diamond^4(\cdot, \cdot)$ which do not occur in the beam theory, and from our inability to integrate by parts on the boundary of the region Ω, and then omit the troublesome boundary terms.

Even more complex formulation of Pontryagin's principle would result if we dropped the assumption that no energy is transmitted at the boundary. Since the complexity of the problem increases immensely with the removal of each assumption, for the sake of clarity we shall first discuss the maximal principle in the simplest possible case, then formulate the increasingly more complex cases, rather than attempting to formulate it in the most general case, and derive the simpler cases by ignoring appropriate terms of the general expression. In all cases discussed below, we shall assume no energy transfer at the boundary of the plate.

2.5.1. Pontryagin's principle for the case D = const., ρ = const. with boundary conditions of the type (B1). The rate of change of the total energy is given by the formula (1.26a) with the term $\diamond^4(w, v)$ vanishing. Since the energy is conserved, we obtain

$$0 = \frac{d\mathscr{E}}{dt} = D \iint_{\Omega} (\nabla^2 w)(\nabla^2 v)\, dx\, dy + \rho \iint_{\Omega} \left(v \frac{\partial v}{\partial t} \right) dx\, dy. \tag{2.1}$$

The formula (1.29c) is applicable in this case:

$$\frac{d}{dt}\langle w_1, w_2 \rangle = -\frac{1}{2(1+\nu)} \int_{\partial\Omega} \left\{ \chi_2 \frac{\partial v_1}{\partial n} + \chi_1 \frac{\partial v_2}{\partial n} + v_1 \frac{\partial \chi_2}{\partial n} + v_2 \frac{\partial \chi_1}{\partial n} \right\} ds$$

$$+ \frac{1}{2} \iint_{\Omega} (v_1 q_2 + v_2 q_1)\, dx\, dy.$$

It is clear that if the condition (B1) is valid on $\partial\Omega$ ($v \equiv 0 \equiv \partial v/\partial n$ on $\partial\Omega$), then

$$\frac{d}{dt}\langle w_1, w_2 \rangle = \frac{1}{2} \iint_{\Omega} (v_1 q_2 + v_2 q_1)\, dx\, dy. \tag{2.2}$$

In particular, if $q_2 \equiv 0$, that is, if $w_H = w_2$ is a solution of the homogeneous equation (1.14), we have

$$\langle w_1, w_H \rangle = \frac{1}{2} \iint_{\Omega} (v_H q_1)\, dx\, dy. \tag{2.2a}$$

We are ready to state the simplest version of Pontryagin's principle for thin plates.

THEOREM 2.1. *Let us assume that $\hat{\phi}(x, y, t)$ is an optimal control on the fixed time interval $[0, t]$ for a thin homogeneous plate, whose flexural rigidity and density are constant in the domain Ω of the plate. Let the plate's edge be clamped along the entire boundary $\partial\Omega$. (That is, on $\partial\Omega$ the displacement function $w(x, y, t)$ satisfies the condition* (B1): *$w \equiv 0$, $\partial w/\partial n \equiv 0$.) Let $w_H(x, y, t)$ denote the displacement of this*

plate vibrating freely, so that the final conditions at the time T are

$$w_H(x, y, T) = w(\hat{\phi}(x, y, t), T),$$
$$\frac{\partial w_H(x, y, T)}{\partial t} = \frac{\partial w(\hat{\phi}(x, y, t), T)}{\partial t}.$$

Then the following inequality is true:

$$\iint_\Omega \left[-\hat{\phi}(x, y, t) \frac{\partial w_H(x, y, t)}{\partial t} \right] dx\, dy \geqq \iint_\Omega \left[-f(x, y, t) \frac{\partial w_H(x, y, t)}{\partial t} \right] dx\, dy \tag{2.3}$$

for all $t \in [0, T]$, *where* $f(x, y, t)$ *is any admissible control.*

We note that this statement is completely analogous to Theorem 3 of [7]. The proof turns out to be a repetition of the proof given in [7] and for that reason will be omitted. As in [7], (2.2) is crucial in the proof of (2.3).

Let us now observe that Pontryagin's principle as given by the inequality (2.3) is inapplicable, if $\mathscr{E}(\hat{\phi}(x, y, t), T) = 0$. If the total energy of the plate can be reduced to zero at the time T, then $\hat{w}_H(x, y, t) \equiv 0$, $t \in [0, T]$, and clearly the inequality (2.3) is meaningless. However, if $\mathscr{E}(\hat{\phi}(x, y, t), T) = 0$ but $\mathscr{E}(\hat{\phi}(x, y, t), \tau) > 0$ for any $0 < \tau < T$, it is possible to introduce a sequence of optimal controls $\{\phi_i\}$ converging to $\hat{\phi}(x, y, t)$ with the inequality (2.3) applicable to each element ϕ_i of that sequence. A detailed description of this limiting process will not be given here.

We observe also the usual shortcomings of Pontryagin's principle. To effect a comparison of an arbitrary control with supposedly an optimal control we need to know the final state of the vibrating plate obtained after the application of an optimal control. Again, however, this principle may be useful in a negative way. That is, we can use the inequality (2.3) to demonstrate that some control $\phi(x, y, t)$ is *not* an optimal control.

Example 1. Let us consider a homogeneous circular plate subjected to a uniformly distributed load of intensity p_0. The edge is clamped. At the time $t = 0$ the load is suddenly removed. The initial deflection is then given by

$$w(r, 0) = \frac{p_0 R^4}{64D} \left[1 - \frac{r^2}{R^2} \right]^2, \qquad r = \sqrt{x^2 + y^2}, \quad r \leqq R. \tag{2.4}$$

It is clear that $w(0, 0) = \max w(r, 0)$, $0 \leqq r \leqq R$.

A control consisting of a constant load $\phi(x, y, t) = Cp_0$ is suggested for the fixed time interval $[0, T]$, $T = \frac{1}{4}n_1$, with the constant C chosen to be $C = 1/(\pi p_0 R^2)$, to assure $\iint_\Omega |\phi|\, dx\, dy = 1$.

The time interval n_1 selected above corresponds to one free vibration cycle of the plate. We observe that the freely vibrating plate will vibrate with the angular velocity $\omega = 4T/(2\pi)$; we also note that the average velocity will be distributed in the same manner as $w(r, 0)$, and that

$$w_H(r, T) = \frac{R^2}{64\pi D} \left[1 - \frac{r^2}{R^2} \right]^2$$

so that

$$\frac{\partial w_H(0, t)}{\partial t} > \frac{\partial w_H(r, t)}{\partial t}, \qquad r \neq 0, \quad t \in [0, T].$$

Hence the sign of our control load was correct, but its distribution certainly was not optimal. Choosing, for example, a time independent admissible load $f_1(x, y, t) = 4Cp_0$ when $0 \leqq r \leqq R/2$, and $f_1(x, y, t) = 0$ when $R/2 \leqq r \leqq R$ (C is given as before by $C = 1/(\pi p_0 R^2)$), we obtain

$$\iint_{\Omega} - f_1(x, y, t)\frac{\partial w_H(x, y, t)}{\partial t}\, dx\, dy > \iint_{\Omega}\left(-\phi\frac{\partial w_H}{\partial t}\right) dx\, dy,$$

showing that ϕ was not an optimal control.

A gradual improvement technique using the standard form of Pontryagin's principle was discussed by the author in [7]. Clearly the same technique, which is also suggested by the above example, can be used to improve a control in a number of iterative steps.

The technique is obviously tedious, and in addition we should point out that while such iteration results in improvements of some of the arbitrarily selected controls, we offer *no* assurance that this iterative process will result in the controls $f_i(x, y, t)$ converging to the optimal control.

We remark that the choice of "optimal" control was made in this example in a deliberately clumsy manner.

2.5.2. Pontryagin's principle for the homogeneous plate (D = const., ρ = const.) with a simply supported part of the boundary consisting of straight lines. The expression (1.29c) for the product $d\langle w_1, w_2\rangle/dt$ becomes

$$\frac{d}{dt}\langle w_1, w_2\rangle = \frac{1}{2}\iint_{\Omega}(v_1 q_2 + v_2 q_1)\, dx\, dy - \frac{1}{2(1+\nu)}\int_{\partial\Omega}\left(\chi_1\frac{\partial v_2}{\partial n} + \chi_2\frac{\partial v_1}{\partial n}\right) ds.$$

Assuming that $q_2 \equiv 0$ ($w_2 = w_H$), we have

$$\frac{d}{dt}\langle w_1, w_H\rangle = \frac{1}{2}\iint_{\Omega}(v_H q_1) - \frac{1}{2(1+\nu)}\int_{\partial\Omega}\left(\chi_1\frac{\partial v_H}{\partial n} + \chi_H\frac{\partial v_1}{\partial n}\right) ds. \tag{2.5}$$

As we remarked following the development of (1.29c), the contour integral in (2.5) does not have to vanish if the boundary of the plate is only simply supported. In an exceptional case when a part of the simply supported boundary (say Γ_1) is a straight line and $D \neq 0$ on Γ_1, we have

$$\int_{\Gamma_1}\left(\chi_1\frac{\partial v_2}{\partial n} + \chi_2\frac{\partial v_1}{\partial n}\right) dx = 0, \tag{2.6}$$

because $\chi_1 = \chi_2 \equiv 0$ on Γ_1 independently of the controls q_1, q_2.

It follows easily now that the inequality (2.3) is applicable to the case when $\partial\Omega$ consists of subarcs Γ_1 and Γ_2 such that $\overline{\Gamma_1 \cup \Gamma_2} = \partial\Omega$, and Γ_1 is the simply supported part of the boundary (condition (B2)) consisting of a straight line,

while Γ_2 is the part of the boundary (not necessarily straight) on which the edge is clamped (condition (B1)).

We intend to show that the inequality (2.3) is also applicable to the physically important case when the simply supported part of the boundary meets the clamped part of the boundary at a corner point. Let us now state the following theorem.

THEOREM 2.2. *Let us assume that the boundary of* Ω *consists of a finite collection of smooth arcs* Γ_1, *such that condition* (B1) *is satisfied on* Γ_1 (*i.e.*, $w \equiv 0$, $\partial w/\partial n = 0$ *on* Γ_1) *and of a finite number of line segments* Γ_2, *such that the plate is freely supported on* Γ_2 (*the condition* (B2)). *Let us assume that all corners are internal corners, i.e., the angle contained in* Ω *between* Γ_2 *and the tangent to* Γ_1 *at the corner point does not exceed* π. *Let* $\hat{\phi}(x, y, t)$ *be an optimal* (*admissible*) *control on the fixed time interval* $[0, T]$ *and let* $f(x, y, t)$ *be any admissible control. Then the inequality* (2.3) *holds, i.e.,*

$$\iint_{\Omega}\left[-\hat{\phi}(x, y, t)\frac{\partial w_H(x, y, t)}{\partial t}\right] dx\, dy \geqq \iint_{\Omega}\left[-f(x, y, t)\frac{\partial w_H(x, y, t)}{\partial t}\right] dx\, dy,$$

where $w_H(x, y, t)$ *has the same meaning as in the statement of the Theorem* 2.1.

Proof. Let us replace the corner points by circle segments of radius $\varepsilon_i = 1/2^i$ with i chosen sufficiently large to permit such change. The segment of the circle drawn with the radius ε_i is contained in Ω, and is tangential to the arc of Γ_1 at a point p_2, and to line Γ_2 at a point p_1, as shown on Fig. 2. The rounded part of the boundary, that is, the circular arc p_1p_2, will be denoted by C_i. The modified region now occupied by the plate (with all corners rounded off) will be denoted by Ω_i.

We assume that conditions (B1) will be satisfied on C_i and on the unchanged part of Γ_1, while condition (B2) is satisfied on the unchanged part of Γ_2.

The conditions are now satisfied for the correctness of the inequality

$$\iint_{\Omega_i}\left[\hat{\phi}_i\frac{\partial w_{H_i}}{\partial t}\right] dx\, dy \geqq \iint_{\Omega_i}\left(-f_i\frac{\partial w_{H_i}}{\partial t}\right) dx\, dy,$$

where $\hat{\phi}_i$ is an optimum control for the region Ω_i, f_i is an arbitrary admissible control for Ω_i, w_{H_i} is the solution of the homogeneous equation of MBVP,

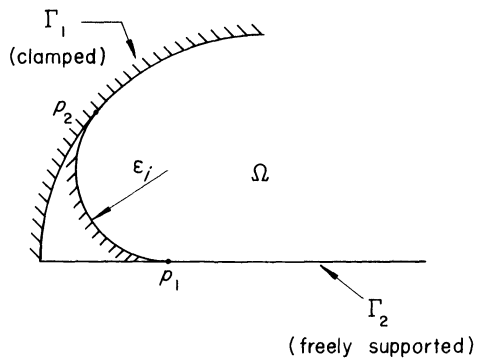


FIG. 2

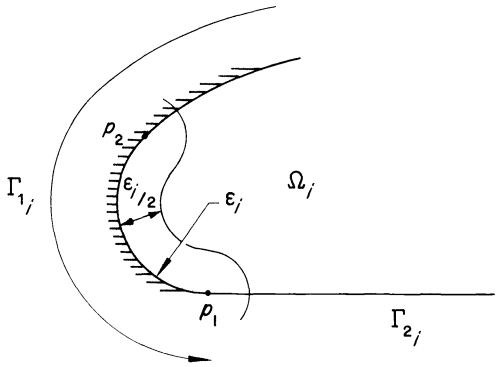


FIG. 3

satisfying the same final condition as $w(\phi_i(x, y, t))$. In the region Ω_i the boundary conditions are posed as stated above and as illustrated on Fig. 3. Let N_{ε_i} denote the $\varepsilon_{i/2}$ neighborhood of the rounded corner. In the region $\Omega_i - (\Omega_i \cap N_{\varepsilon_i})$ we have

$$w(x, y, 0) = \psi(x, y), \qquad \frac{\partial w(x, y, 0)}{\partial t} = \eta(x, y),$$

which are the specified initial conditions for $w(x, y, t)$ in Ω as given in the initial conditions (C1), (C2). In $\Omega_i \cap N_{\varepsilon_i}$ we apply a mollifier function (of class C^∞) which meets both the conditions (C1), (C2) on the boundary of N_{ε_i} and the condition

$$w(x, y, 0) = 0, \qquad \frac{\partial w}{\partial n}(x, y, 0) = 0 \quad \text{on } C_i.$$

(We recall that C_i is the rounded part of $\partial\Omega$ drawn with the radius ε_i.) For example, the well-known type of function

$$\phi(r, \varepsilon) = \begin{cases} \exp\{(-\varepsilon/2)/(\varepsilon_{i/4}^2 - r^2)\}, & r < \varepsilon_{i/2}, \\ 0, & r \geqq \varepsilon_{i/2} \end{cases}$$

(where r is the distance from $\partial N_{\varepsilon_i} \cap \Omega_i$), could serve as the mollifier function. If we consider a sequence of numbers $\varepsilon_k = 1/2^k$, $k = i, i+1, i+2, \cdots$, and the corresponding sequence of optimal controls $\{\hat{\phi}_k\}$, we are assured (Lemma 1 of Appendix 1 of [7]) that there exists a control $\hat{\phi}$, such that some subsequence of controls $\{\phi_k\}$ (say $\{\phi_{km}\}$) converges to $\hat{\phi}$ as $k \to \infty$. Moreover,

$$\iint_\Omega \left(-\hat{\phi}\,\frac{\partial w_H}{\partial t}\right) \geqq \iint_\Omega \left(-f_i \frac{\partial w_H}{\partial t}\right) \tag{2.7}$$

is valid for any admissible control f_i acting on Ω_i. A serious problem arises at this point of the proof. Namely, we observe that while f_i is an admissible control on Ω_i, i.e., $\|f_i\|_{\Omega_i} \leqq 1$, it may be an inadmissible control on Ω, or even on Ω_{i+1}.

We recollect however that either the applied loads and moments are given by a finite number of Dirac delta functions and their derivatives applied at a

finite number of points in Ω, in which case there exist $\varepsilon_1 > 0$ and a neighborhood N_{ε_1} of $\partial\Omega$ which is free from such loads, or that they are square integrable functions in Ω, bounded in the maximum norm, so that we can choose $\varepsilon_2 > 0$ and a neighborhood N_{ε_2} of $\partial\Omega$ such that given $0 < \varepsilon \ll 1$, $\iint_{N_{\varepsilon_2}\cap\Omega} \left| \sum_{j=1}^{n} f_j \right| dx\, dy < \varepsilon/2^i$ (for a fixed i). Hence $\iint_{N_{\varepsilon_3}\cap\Omega} |f_i|\, dx\, dy < \varepsilon/2^i$ if $\varepsilon_3 = \min(\varepsilon_1, \varepsilon_2)$.

Since w_H is continuously differentiable and $\partial w_H/\partial t \equiv 0$ on $\partial\Omega$ we can choose $\varepsilon_4 > 0$ and a neighborhood N_{ε_4} of $\partial\Omega$ such that

$$\iint_{N_{\varepsilon_4}\cap\Omega} \left| \frac{\partial w_H}{\partial t} \right| dx\, dy < \varepsilon/2^i,$$

and finally denoting $\hat{\varepsilon} = \min(\varepsilon_3, \varepsilon_4)$ we have also

$$\iint_{N_{\hat{\varepsilon}}\cap\Omega} \left| f_i \frac{\partial w_{H_i}}{\partial t} \right| dx\, dy < \varepsilon^2/2^{2i}. \tag{2.8}$$

All we need to do now is redefine the class of admissible controls to satisfy the inequality

$$\|f_i\|_{\Omega_i} \leqq 1 - \varepsilon/2^i.$$

Now if f_i is an admissible control on Ω_i, that is, if $\|f_i\|_{\Omega_i} < 1 - \varepsilon/2^i$, then after reducing the radius of the corner to $1/2^{i+1}$, we have $\|f_i\|_{\Omega_{i+1}} \leqq \|f_i\|_{\Omega_i} + \|\sum f_j\|_{N_\varepsilon} < 1 - \varepsilon/2^i \leqq 1 - \varepsilon/2^{i+1}$, and f_{i+1} is again an admissible control. We conclude that given an optimal control $\hat{\phi}(x, y, t)$ acting on Ω, and a sequence of regions Ω_i, we can select a sequence of optimal controls $\{\phi_i\}$, each ϕ_i acting in Ω_i, such that $\hat{\phi}(x, y, t) = \lim_{i\to\infty} \phi_i(x, y, t)$.

If it were not so, then we could find $\varepsilon > 0$, such that $\|\phi_i - \hat{\phi}\|_\Omega > \varepsilon$, for all optimal controls ϕ_i (acting on Ω_i) for sufficiently large indices i. Regarding ϕ_i as controls on Ω (i.e., $\phi_{i(\Omega)} = \phi_i$ on Ω_i and $\phi_{i(\Omega)} = 0$ on $\Omega - \Omega_i$), and remembering that each ϕ_i is an optimal control on Ω_i, we obtain an easy contradiction to the statement (2.8). This shows that such an $\varepsilon > 0$ cannot be found, and that indeed such a sequence $\{\phi_i\}$ can be selected.

The corresponding sequence w_{H_i} of displacement functions satisfying the homogeneous equation must also converge to the function $w_H(\hat{\phi}(x, y, t), x, y, t)$. To prove this statement we use the Arzela–Ascoli theorem, since $\{w_{H_i}\}$ forms an equicontinuous family of functions. Hence, some subsequence of $\{w_{H_i}\}$ must converge to a function $\tilde{w}_H(x, y, t)$. Because of the elastica hypothesis concerning each function w_{H_i}, $\tilde{w}_H(x, y, t)$ is also a differentiable function. Using Duhamel's principle we have

$$\tilde{w}_H(x, y, t) - w_H(x, y, t) = \int_0^t \iint_\Omega G((x - \xi), (y - \eta), (t - \tau)) \cdot [\hat{\phi}(x, y, t) - \lim_{i\to\infty} \phi_i(x, y, t)]\, dx\, dy\, dt = 0$$

so that $\tilde{w}_H(x, y, t) = w_H(x, y, t)$, $t \in [0, T]$.

Now given an arbitrary admissible control f acting on Ω, we can select a sequence of admissible controls $\{f_i\}$ on Ω_i such that $\lim_{i\to\infty} f_i = f$. (The argument is identical with the preceding one.) Since for each f_i the inequality (2.7) is valid, we have in the limit

$$\lim_{i\to\infty} \iint_{\Omega} \left(-\phi_i \frac{\partial w_{H_i}}{\partial t} \right) dx\, dy \geqq \lim_{i\to\infty} \iint_{\Omega} \left(-f_i \frac{\partial w_{H_i}}{\partial t} \right) dx\, dy,$$

and finally

$$\iint_{\Omega} \left(-\hat{\phi} \frac{\partial w_H}{\partial t} \right) \geqq \iint_{\Omega} \left(-f \frac{\partial w_H}{\partial t} \right) dx\, dy,$$

which was to be proved.

Example 2. Consider a semicircular plate occupying the region Ω as shown on Fig. 4. The plate is simply supported along the diameter Γ_2 and clamped along the entire arc Γ_1. A uniform load p_0 is applied to this plate, then suddenly removed at the time t_0. Show that an admissible load uniformly distributed on some circular disc contained in Ω applied the instant t_0 and maintained for some time interval $[t_0, t_1]$ cannot be altered at t_1, so that the resulting control $\hat{\phi}(x, y, t)$, $t \in [t_0, T] (t_1 < T)$, will be optimal on the fixed interval $[t_0, T]$, where T is chosen for convenience to be $\frac{1}{4}$ of the time necessary to complete one cycle of vibration with the lowest natural frequency. The argument supporting this claim is similar to that of Example 1. We use the fact that the velocity of a freely vibrating plate does assume a maximum value at an isolated point in Ω. (A superposition of two uniformly distributed loads will have that effect.) Considering a Dirac delta function applied to the point of maximum velocity during a sufficiently short subinterval of $[t_0, t_1]$, we can show that the inequality (2.3) must be incorrect in that subinterval. However, the boundary conditions stated in this example imply the validity of (2.3) if our assumed control load was optimal. Hence it cannot be optimal, as we intended to show.

In what follows Γ_1 will denote the part of $\partial\Omega$ obeying the fixed edge condition, Γ_2 will be the simply supported part of $\partial\Omega$, and Γ_3 will denote the part of $\partial\Omega$ obeying the free edge condition.

2.5.3. Proof of Pontryagin's principle for the case when Γ_2 and Γ_3 consist of straight line intervals.

THEOREM 2.3. *Let us assume that $\hat{\phi}(x, y, t)$ is an optimal control for the fixed time interval $[0, T]$ for a thin homogeneous plate, whose flexural rigidity and density*

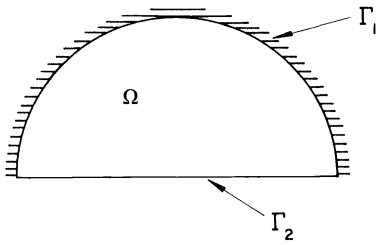


FIG. 4

are constant. Let the boundary consist of parts $\Gamma_1, \Gamma_2, \Gamma_3 (\overline{\Gamma_1 \cup \Gamma_2 \cup \Gamma_3} = \partial\Omega)$, *where on* Γ_1 *the plate obeys the condition* (B1) (*i.e., clamped edge condition*), *on* Γ_2 *it obeys the condition* (B2), *and on* Γ_3 *the condition* (B3). *We assume that* Γ_2 *and* Γ_3 *are a union of a finite number of straight lines and that* Γ_1 *is a union of a finite number of piecewise smooth arcs* (*consequently that we have only a finite number of corner points*). *We assume that all corners are internal corners, and all corner points are endpoints of an arc of* Γ_1. *Then the inequality* (2.3) *holds*:

$$\iint_{\Omega} -\hat{\phi}(x, y, t)\frac{\partial w_H(x, y, t)}{\partial t}\,dx\,dy \geqq \iint_{\Omega} -f(x, y, t)\frac{\partial w_H(x, y, t)}{\partial t}\,dx\,dy,$$

where $w_H(x, y, t)$ *has the same meaning as in Theorems* 2.1 *and* 2.2.

Proof. We can prove this theorem under the assumption that all corner points have been replaced by circular arcs of Γ_1 of sufficiently small radius, that is, the assumption that $\partial\Omega$ is smooth, and then deal with the corner points in exactly the same manner as in Theorem 2.2. The rest of the proof is omitted since it repeats the previous arguments and follows the standard technique of [14].

We emphasize the fact that the corner points were restricted to be either the points which lie on Γ_1, or the points which are the endpoints of an arc of Γ_1. The limiting process which was used in the proof of Theorem 2.2 succeeded because each contour integral vanished independently of the curvature κ. This would not be true in the case of the free edge, or if the simple support conditions were present, since in that case the value of each integral would depend on κ, and as the sequence of smooth boundaries approximated $\partial\Omega$, it would increase without bounds, and we would have to consider in the limit a singular contour integral.

2.5.4. The case when $\partial\Omega = \Gamma_2$ and is composed of a finite number of smooth arcs with no corner points. D = const., ρ = const. The absence of corner points allows us to avoid singular contour integrals. However, the crucial relationship

$$d\langle w_1, w_2\rangle\,dt = \frac{1}{2}\iint_{\Omega} (v_2 q_1 + v_1 q_2)\,dx\,dy$$

is no longer true. Instead we must consider the formula (1.29c), modified by putting $v_1 = v_2 \equiv 0$ on $\partial\Omega$:

$$\frac{d}{dt}\langle w_1, w_2\rangle = \frac{1}{2}\iint_{\Omega} (v_2 q_1 + v_1 q_2)\,dx\,dy + \frac{1}{2}\int_{\partial\Omega} D\left[\nabla^2 w_1 \frac{\partial v_2}{\partial n} + \nabla^2 w_2 \frac{\partial v_1}{\partial n}\right] ds.$$

If w_1, w_2 are solutions of the homogeneous equation (1.11) in some interval I, then the sign of $d\langle w_1, w_2\rangle/dt$ is the same as the sign of

$$\int_{\partial\Omega} D\left\{\nabla^2 w_1 \frac{\partial v_2}{\partial n} + \nabla^2 w_2 \frac{\partial v_1}{\partial n}\right\} ds. \tag{2.9}$$

Since the plate is simply supported on $\partial\Omega$, we can effect some simplifications of the formula (2.9). We have

$$\chi = M_{nn} + M_{\tau\tau} = -D(1 + \nu)\nabla^2 w = -D(1 + \nu)\left(\frac{\partial^2 w}{\partial n^2} + \frac{\partial^2 w}{\partial \tau^2}\right).$$

However on $\partial\Omega$ we have

$$w = \frac{\partial w}{\partial s} = \frac{\partial^2 w}{\partial s^2} \equiv 0 \quad \text{and} \quad M_{nn} \equiv 0,$$

because of the simple support condition (B2). We use the relationship $\partial^2 w/\partial\tau^2 = \partial^2 w/\partial s^2 + \kappa(\partial w/\partial n)$, where as before κ is the curvature of the boundary.

$$\chi = M_{\tau\tau} = -D(1+\nu)\left(\frac{\partial^2 w}{\partial n^2} + \frac{\partial^2 w}{\partial s^2} + \kappa\frac{\partial w}{\partial n}\right) = -D(1+\nu)\left(\frac{\partial^2 w}{\partial n^2} + \kappa\frac{\partial w}{\partial n}\right) \quad \text{on } \partial\Omega.$$

$$\frac{\partial^2 w}{\partial n^2} = -\nu\frac{\partial^2 w}{\partial \tau^2} = -\nu\left(\frac{\partial^2 w}{\partial s^2} + \kappa\frac{\partial w}{\partial n}\right) = -\nu\kappa\frac{\partial w}{\partial n}. \tag{2.10}$$

Hence on $\partial\Omega$ we obtain

$$\nabla^2 w = \frac{\partial^2 w}{\partial n^2} + \frac{\partial^2 w}{\partial \tau^2} = \frac{\partial^2 w}{\partial n^2} + \kappa\frac{\partial w}{\partial n} = -\nu\kappa\frac{\partial w}{\partial n} + \kappa\frac{\partial w}{\partial n} = (1-\nu)\kappa\frac{\partial w}{\partial n}. \tag{2.11}$$

The equation (2.10) can be rewritten

$$\begin{aligned}\frac{d}{dt}\langle w_1, w_2\rangle &= \int_{\partial\Omega} D(1-\nu)\kappa\left[\frac{\partial w_1}{\partial n}\frac{\partial v_2}{\partial n} + \frac{\partial v_1}{\partial n}\frac{\partial w_2}{\partial n}\right] ds \\ &= D(1-\nu)\int_{\partial\Omega} \kappa\frac{\partial}{\partial t}\left(\frac{\partial w_1}{\partial n}\cdot\frac{\partial w_2}{\partial n}\right) ds.\end{aligned} \tag{2.12}$$

We are now ready to repeat the arguments of [14]. Let $w_1 = \hat{w}$ be the optimal displacement corresponding to the optimal control $\hat{\phi}(x, y, t)$. Let w_H be the solution of the homogeneous equation (1.14) with the property

$$w_H(x, y, T) = \hat{w}(x, y, T),$$

and as before we denote $w'(x, y, t)$ by the function $w'(x, y, t) = \hat{w} + \varepsilon w_\delta$ where w_δ is a function whose support is the time interval I_δ, with the properties identical to the function w_δ described in Theorem 3 of [7]. We have the equality

$$\mathscr{E}(w') = \mathscr{E}(\hat{w}) + 2\varepsilon\langle\hat{w}, w_\delta\rangle + f(\varepsilon^2), \qquad |\varepsilon| < \varepsilon_0.$$

Hence if w_H is a solution of the homogeneous equation (1.14) satisfying

$$w_H(T) = \hat{w}(T), \qquad v_H(T) = \hat{v}(T),$$

then

$$\langle w_H, w_\delta\rangle_{t=T} \geqq 0.$$

But

$$\begin{aligned}\langle w_H, w_\delta\rangle_{\tau=T} &= \langle w_H, w_\delta\rangle_{t=t_0+\delta} + \int_{t_0+\delta}^{T}\frac{d}{dt}\langle w_H, w_\delta\rangle \\ &= \langle w_H, w_\delta\rangle_{t=t_0+\delta} + \tfrac{1}{2}D(1-\nu)\int_{t_0+\delta}^{T}\left\{\int_{\partial\Omega}\frac{\partial}{\partial t}\left(\kappa\frac{\partial w_H}{\partial n}\frac{\partial w_\delta}{\partial n}\right) ds\right\} dt\end{aligned}$$

$$= \frac{1}{2}\int_{t_0-\delta}^{t_0+\delta}\left[\iint_\Omega (\phi_\delta v_H)\,dx\,dy\right]dt$$

$$+\tfrac{1}{2}D(1-\nu)\int_{t_0-\delta}^{T}\left[\int_{\partial\Omega}\frac{\partial}{\partial t}\left(\kappa\frac{\partial w_H}{\partial n}\frac{\partial w_\delta}{\partial n}\right)ds\right]dt.$$

Hence we must have

$$\int_{t-\delta}^{t+\delta}\left[\iint_\Omega (-\phi_\delta v_H)\,dx\,dy\right]dt \leqq \int_{t-\delta}^{T}\left[\int_{\partial\Omega}\frac{\partial}{dt}\left(\kappa\frac{\partial w_\delta}{\partial n}\frac{\partial w_H}{\partial n}\right)ds\right]dt.$$

By an argument analogous to [14], we finally obtain for an arbitrary control $\tilde{\phi} = \hat{\phi} + \phi_\delta$ the result

$$\int_{t_0-\delta}^{t_0+\delta}\left[\iint_\Omega (-\tilde{\phi}v_H)\,dx\,dy\right]dt \leqq \int_{t_0-\delta}^{t_0+\delta}\left[\iint_\Omega (-\hat{\phi}v_H)\,dx\,dy\right]dt$$

$$+\int_{t_0-\delta}^{T}\left[\int_{\partial\Omega}\frac{\partial}{\partial t}\left(\kappa\frac{\partial(\tilde{w}\cdot\hat{w})}{\partial n}\cdot\frac{\partial w_H}{\partial n}\right)ds\right]dt.$$

Since δ was arbitrary, and since $\lim_{\delta\to 0}\delta^{-1}\int_{t-\delta}^{t+\delta}(\tilde{\phi}v_H) = 0$ uniformly, for any admissible control ϕ, we have for any $t\in[0,T]$,

$$\iint_\Omega (-\phi v_H)\,dx\,dy - \int_t^T\left[\int_{\partial\Omega}\frac{\partial}{\partial t}\left(\kappa\frac{\partial\tilde{w}}{\partial n}\frac{\partial w_H}{\partial n}\right)ds\right]dt$$

(2.13)

$$\leqq \iint_\Omega (-\hat{\phi}v_H)\,dx\,dy - \int_t^T\left[\int_{\partial\Omega}\frac{\partial}{\partial t}\left(\kappa\frac{\partial\hat{w}}{\partial n}\frac{\partial w_H}{\partial n}\right)ds\right]dt$$

which is a form of the maximum principle of Pontryagin. It reduces to the formula (2.3) if we either change the boundary conditions, or if we put $\kappa \equiv 0$ on $\partial\Omega$, or if we demand that for some reason

$$\int_{\partial\Omega}\kappa\left(\frac{\partial\tilde{w}}{\partial n}\frac{\partial w_H}{\partial n}\right)ds = \text{const.} \tag{2.14}$$

for all $t\in[0,T)$, and for any admissible displacement $\tilde{w}(x,y,t)$.

In its present form the inequality (2.13) appears to be quite useless. Analogous formulas can be easily developed for a boundary consisting of the arcs $\Gamma_1, \Gamma_2, \Gamma_3$ obeying the boundary conditions (B1), (B2), and (B3) respectively. These formulas will not be reproduced here, since their usefulness is also questionable.

2.5.5. The case when $\partial\Omega = \Gamma_1 \cup \Gamma_2$ and Γ_2 is composed of straight-line segments. $\partial\Omega$ may contain internal corners which are situated on $\partial\Omega$. (As before we assume that $\partial\Omega$ is a union of a finite number of smooth arcs.) A special case when the corner points occur either on Γ_1 or at a point where an arc of Γ_1 joins an arc of Γ_2 has already been covered. We need only to consider the behavior of the line integral along some subset γ of Γ_2, which contains an interior corner. As

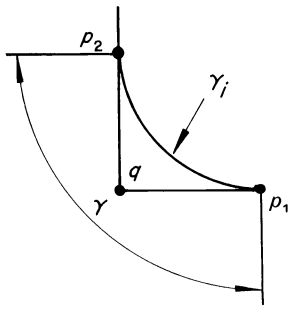


FIG. 5

in Theorem 2.2 we can approximate each corner by a sequence of circular arcs γ_i of radius $\varepsilon_i = 1/2^i$, $i \geqq N$, where N is chosen so that the circular arc ε_N lies entirely in Ω.

The contour integral (from p_1 to p_2 along γ)

$$\int_\gamma \kappa \left(\frac{\partial w}{\partial n} \frac{\partial w_H}{\partial n} \right) ds$$

cannot be evaluated directly since neither κ nor $\partial w/\partial n$ is defined at the corner point q. However along each circular arc γ_i we have

$$\int_{\gamma_i} \left(\kappa_i \frac{\partial w}{\partial n} \frac{\partial w_H}{\partial n} \right) ds = \frac{1}{\nu^2} \int_{\gamma_i} \frac{1}{\kappa_i} \left(\frac{\partial^2 w}{\partial n^2} \cdot \frac{\partial^2 w_H}{\partial n^2} \right) ds$$

because of formula (2.10); or using the formula (2.11), we have

$$\kappa_i \frac{\partial w}{\partial n} = - \frac{M_{\tau\tau}}{D(1 - \nu^2)},$$

and therefore

$$\begin{aligned} \int_{\gamma_i} \left(\kappa_i \frac{\partial w}{\partial n} \frac{\partial w_H}{\partial n} \right) ds &= \frac{1}{D^2(1 - \nu^2)^2} \int_{\gamma_i} \left(\frac{1}{\kappa_i} M_{\tau\tau} \cdot M_{\tau\tau_H} \right) ds \\ &\leqq \int \left| \frac{1}{\kappa_i} \right| ds \cdot \int |M_{\tau\tau}| \, ds \cdot \int |M_{\tau\tau_H}| \, ds. \end{aligned}$$

Since $\lim_{i \to \infty} 1/\kappa_i = 0$, and by assumption $\int |M_{\tau\tau}| \, ds$ and $\int |M_{\tau\tau_H}| \, ds$ are bounded, we obtain the desired result

$$\lim_{i \to \infty} \int_{\gamma_i} \left(\kappa_i \frac{\partial w}{\partial n} \cdot \frac{\partial w_H}{\partial n} \right) ds = 0.$$

The following result is an easy consequence: *In Theorem* 2.3 *the last sentence*, namely: "all corner points are the endpoints of an arc of Γ_1," *can be omitted.*

3. Instantly optimal controls of thin vibrating plates. The definition of an instantly optimal control was given by the author in [7].

We can prove (see [7]) that if the initial fixed interval $[0, T]$ optimal control $\hat{\phi}(x, y, t)$ satisfies the maximum principle of Theorems 2.1, 2.2, 2.3, i.e., if

$$\iint_{\Omega} -\left[\hat{\phi}(x, y, t)\frac{\partial w_H(x, y, t)}{\partial t}\right] dx\, dy \geqq \iint_{\Omega}\left[-f(x, y, t)\frac{\partial w_H(x, y, t)}{\partial t}\right] dx\, dy$$

for any admissible control $f(x, y, t)$, then the instantly optimal control $\tilde{\phi}$ will satisfy the maximum principle

$$\iint_{\Omega} -\left[\tilde{\phi}(x, y, t)\frac{\partial w(\tilde{\phi}; x, y, t)}{\partial t}\right] dx\, dy \geqq \iint_{\Omega}\left[-f(x, y, t)\frac{\partial w(f; x, y, t)}{\partial t}\right] dx\, dy$$

for any admissible control f.

The usefulness of this maximum principle greatly depends on the following lemma.

LEMMA 3.1. *The instantly optimal control $\tilde{\phi}$ is unique (that is, independent of either the manner in which we subdivided the energy, or of our choice of the intermediate optimal controls $\hat{\phi}_{i,j}(x, y, t)$).*

4. Some comments on the optimum excitation problem. We consider the following problem. (a) Let the boundary conditions of the types (B1), (B2), (B3) and the initial conditions (C1), (C2) be given for the MBVP. Find an admissible control $\tilde{\phi}(x, y, t)$ for the fixed interval $[0, T]$ such that the total energy of the plate $\tilde{\mathscr{E}}(T) = \mathscr{E}(\tilde{\phi}(x, y, t), T)$ at the time T attains the maximum possible value, i.e., $\mathscr{E}(\tilde{\phi}(x, y, t), T) \geqq \mathscr{E}(f(x, y, t), T)$ for any admissible control $f(x, y, t)$.

This problem is closely related to the resonance problem, and the corresponding maximum principle reveals a physical interpretation of one possible kind of resonance. In fact any control $\phi(x, y, t)$ such that $\lim_{t\to\infty} \mathscr{E}(\phi(x, y, t), t) = \infty$ can be designated as a control of the resonance type.

A different optimal excitation is obtained by requiring a control of the MBVP to obey one of the following two conditions:

(b) The rate of increase of total energy is maximized, i.e., $d\tilde{\mathscr{E}}(\tilde{\phi}(x, y, t), t)/dt \geqq d\mathscr{E}(f(x, y, t), t)/dt$ for any admissible control $f(x, y, t)$, $t > 0$.

(c) Given any $\tilde{\mathscr{E}} > \mathscr{E}(t = 0)$ find a control $\tilde{\phi}(x, y, t)$ such that the plate attains the total energy level $\tilde{\mathscr{E}}$ in the shortest possible time.

A control satisfying (a) will be called an optimal excitation for a fixed time interval. Condition (b) will be called an excitation with the steepest rate of energy increase. Condition (c) will be called the time optimal excitation. Other definitions of optimality can be readily proposed.

To see the basic relationship between controls of the types (a) and (c) we need the following lemma.

LEMMA 4.1. *Let the boundary conditions and the initial conditions (at $t = 0$) be given. Let us assume no energy transfer at the boundary $\partial\Omega$. Then given $t_1 > 0$, there exists a control $\tilde{\phi}(x, y, t)$ such that $\mathscr{E}(\phi(x, y, t), t) > \mathscr{E}(t = 0)$.*

Proof. If the initial conditions are $w(x, y, 0) \equiv 0$ in Ω, then any control function $\phi(x, y, t)$, such that $\phi(x, y, t) > 0$ in Ω and in a sufficiently small sub-

interval of $[0, t_1]$ and $\phi(x, y, t) \equiv 0$ in the remainder of $[0, t_1]$, will serve our purpose. If the initial conditions are different from $w(x, y, 0) = 0$ in Ω, then there must be some subinterval $[\tau_1, \tau_2]$ of $[0, t_1]$ such that in some open neighborhood $N_{(\xi,\eta)}$ of a point $(x = \xi, y = n) \in \Omega$ the velocity $dw(x, y, t)/dt$ retains a constant sign. Then we apply the control

$$\phi(x, y, t) \begin{cases} \equiv 0 & \text{if } t \notin [\tau_1, \tau_2], \\ = \delta(x - \xi, y - \eta) \cdot \operatorname{sgn} \dfrac{dw(\xi, \eta, t)}{dt} & \text{if } t \in [\tau_1, \tau_2]\,; \end{cases}$$

ϕ is easily shown to increase the energy of the plate.

LEMMA 4.2. *Every optimal excitation for a fixed time interval is also a time optimal excitation.*

Proof. We assume that there can be found $\hat{\phi}(x, y, t)$ which is an optimal excitation for the fixed time interval $[0, T]$, but fails to be a time optimal excitation, and we shall show that this assumption leads to a contradiction. Since $\hat{\phi}(x, y, t)$ was not a time optimal excitation, there must exist a control $\phi_1(x, y, t)$ such that the energy level $\hat{\mathscr{E}}(\hat{\phi}(x, y, t), T)$ can be reached in time $t_1 < T$, i.e., $\mathscr{E}(\phi_1(x, y, t), t_1) = \hat{\mathscr{E}}(\hat{\phi}(x, y, t), T)$. By Lemma 4.1 there exists some admissible control $\hat{\phi}_2(x, y, t)$ on the time interval $[t_1, T]$ such that

$$\mathscr{E}(\tilde{\phi}_2(x, y, t), T) > \hat{\mathscr{E}}.$$

The control

$$\tilde{\phi} = \begin{cases} \phi_1(x, y, t), & 0 < t \leqq t_1, \\ \phi_2(x, y, t), & t_1 < t \leqq T, \end{cases}$$

is an admissible control, and we have

$$\mathscr{E}(\tilde{\phi}(x, y, t), T) > \hat{\mathscr{E}}(\hat{\phi}(x, y, t), T)$$

which contradicts the fact that $\hat{\phi}(x, y, t)$ was optimal for the fixed time interval $[0, T]$.

4.1. Pontryagin's principle for the optimal excitation of a plate for a fixed time interval.

Let the boundary conditions be those of either Theorem 2.1, *or* 2.2, *or* 2.3. *Let* $\tilde{f}(x, y, t)$ *be an optimal excitation of the plate for a fixed time interval* $[0, T]$. *Let* $f(x, y, t)$ *be any admissible control. Then the inequality*

$$\iint_{\Omega} \left(\tilde{f}(x, y, t) \frac{\partial w_H(x, y, t)}{\partial t} \right) dx\, dy \geqq \iint_{\Omega} \left(f(x, y, t) \frac{\partial w_H(x, y, t)}{\partial t} \right) dx\, dy \tag{4.1}$$

holds (w_H *has the same meaning as before*).

The proof repeats the one given for the optimum control with all inequalities reversed.

Some remarks concerning (4.1). Despite the fact that this formula is identical, except for the reversal of the inequality sign, with the optimal control formula, it is less useful because of the shortcomings discussed in § 2.4. In particular, the absence of a convexity lemma is a critical defect, preventing a parallel development.

5. Anisotropic plates and reinforced plates. Assuming a generalized Hooke's law of the form

$$\varepsilon_{ij} = c_{ijkl}\tau_{kl} \tag{5.1}$$

and assuming in addition the hypothesis (E1)–(E4) we can derive the equation of equilibrium

$$\frac{\partial^2 M_{xx}}{\partial x^2} - \frac{\partial^2 M_{xy}}{\partial x \partial y} + \frac{\partial^2 M_{yx}}{\partial x \partial y} + \frac{\partial^2 M_{yy}}{\partial y^2} = -q_0 - \rho \frac{\partial^2 w}{\partial t^2}$$

by substituting the appropriate form of M_{xx}, M_{yy} (see for example [25, pp. 53–54]).

While the resulting equations look complex, it can be seen that there is no essential change in the argument regarding the optimal control, and consequently the basic form of Pontryagin's principle (inequality (2.3)) will be applicable with analogous boundary conditions.

Concluding remarks. Some comment should be made regarding the specific choice of admissible controls, which was made in this paper in § 1.8; in particular, we should explain why only the Dirac delta function and its first derivative were considered as the admissible loads $\phi(x, y)$ which were not bounded measurable functions in Ω. It can be shown in fact that among all distributions which are not regular functions, these two are the only admissible distributions in the following sense. The corresponding solutions of the basic plate equations satisfy all hypotheses of linear plate theory and the "elastica" hypothesis (E1)–(E4) of this article. The proof of this statement follows directly from the theorem of Lidskii, which states that every distribution over the space K is a derivative (of some order) of a completely continuous function, and from the well-known lemma of Sobolev. This proof will be published elsewhere. Of course, this statement may be regarded as trivial by structural engineers who for centuries have admitted point loads and point moments in their analysis, but excluded as possible loads the more complicated distributions which do occur in the problems of theoretical physics.

REFERENCES

[1] N. I. AKHIEZER AND I. M. GLAZMAN, *Theory of Linear Operators in Hilbert Space*, vol. 1, Frederick Ungar, New York, 1961.

[2] A. G. BUTKOVSKII AND A. Y. LERNER, *Optimal control systems with distributed parameters*, Dokl. Akad. Nauk SSSR, 134 (1960), no. 4, pp. 778–781.

[3] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints. I, II*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412, 413–430.

[4] A. I. EGOROV, *Optimal processes and invariance theory*, this Journal, 4 (1966), pp. 601–661.

[5] I. M. GELFAND AND G. E. SHILOV, *Generalized Functions*, vol. 1, Academic Press, New York, 1964.

[6] V. KOMKOV, *A note on the vibration of thin inhomogeneous plates*, Z. Angew. Math. Mech., 48 (1968), pp. 11–16.

[7] ———, *The optimal control of a transverse vibration of a beam*, this Journal, 6 (1968), pp. 401–421.

[8] J. P. LA SALLE, *The time optimal control problem*, Contributions to Non-linear Oscillations, vol. V, Ann. Math. Studies No. 45, Princeton University Press, Princeton, 1960, pp. 1–24.

[9] E. H. MANSFIELD, *The Bending and Stretching of Plates*, Macmillan, New York, 1964.

[10] ———, *On the analysis of elastic plates of variable thickness*, Quart. J. Mech. Appl. Math., 15 (1962), pp. 167–181.

[11] L. S. D. MORLEY, *Skew Plates and Structures*, Macmillan, New York, 1963.

[12] ———, *Variational reduction of the clamped plate to two successive membrane problems with an application to a uniformly loaded section*, Quart. J. Mech. Appl. Math., 16 (1963), pp. 451–471.
[13] N. I. MUSKHELISHVILI, *Some Basic Problems of the Mathematical Theory of Elasticity*, P. Noordhoff, Groningen, The Netherlands, 1953.
[14] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
[15] E. REISSNER, *On bending of elastic plates*, Quart. Appl. Math., 5 (1947), pp. 55–68.
[16] D. L. RUSSELL, *Linear symmetric hyperbolic systems*, this Journal, 4 (1966), pp. 276–294.
[17] ———, *Optimal regulation of linear symmetric hyperbolic systems with finite dimensional controls*, MRC Tech. Rep. 566, Mathematics Research Center, Madison, Wisconsin, 1965.
[18] D. I. SHERMAN, *On the solution of a plane static problem of the theory of elasticity for given external forces*, Dokl. Akad. Nauk SSSR, 27 (1940), no. 9, pp. 907–910.
[19] S. L. SOBOLEV, *Applications of Functional Analysis in Mathematical Physics*, Amer. Math. Soc. Transl., Providence, 1963.
[20] I. S. SOKOLNIKOFF, *Mathematical Theory of Elasticity*, McGraw-Hill, New York, 1956.
[21] S. P. TIMOSHENKO AND S. WOINOWSKI-KRIEGER, *The Theory of Plates and Shells*, McGraw-Hill, New York, 1960.
[22] S. P. TIMOSHENKO, *History of Strength of Materials*, McGraw-Hill, New York, 1954.
[23] S. P. TIMOSHENKO AND J. M. GERE, *Theory of Elastic Stability*, McGraw-Hill, New York, 1961.
[24] K. YOSIDA, *Lectures on Differential and Integral Equations*, Interscience, New York, 1960.
[25] S. A. AMBARTSUMIAN, *Theory of Anisotropic Plates*, Nauka, Moscow, 1967.

SIAM J. CONTROL
Vol. 8, No. 3, August 1970

AN EXAMPLE OF A MAX-MIN PROBLEM IN PARTIAL DIFFERENTIAL EQUATIONS*

JEAN CEA† AND KAZIMIERZ MALANOWSKI‡

Introduction. We study in this paper a "max-min" problem. By solving explicitly the "min problem," the original problem can be stated as a control problem. The state equation is a partial differential equation, and the control appears only in the coefficients of the state equation. We prove existence and uniqueness (in a suitable sense) of "an optimal problem" and we give a convergent iterative method to compute the solution.

For variational problems and Sobolev spaces, we refer to J. L. Lions and E. Magenes [6].

Some problems of the same nature can be found in J. L. Lions [5, pp. 92–96] and R. J. Duffin and D. K. McLain [2].

1. The problem. We denote by Ω an open bounded regular set of $\mathfrak{R}^n$ and by Γ its boundary. Recall that $y \in H_0^1(\Omega)$ means

$$y \in L^2(\Omega), \quad D_i y = \frac{\partial}{\partial x_i} y \in L^2(\Omega) \quad \text{for } i = 1, \cdots, n,$$

$$y = 0 \quad \text{on } \Gamma.$$

It is known that $H_0^1(\Omega)$ is a Hilbert space, with the inner product

$$((y, z)) = \sum_{i=1}^{n} (D_i y, D_i z)_{L^2(\Omega)}$$

or

$$((y, z)) = \int_\Omega \operatorname{grad} y(x) \cdot \operatorname{grad} z(x)\, dx,$$

where grad y is the vector of components $D_i y$, $i = 1, \cdots, n$.

We introduce now a bounded, closed, convex subset U of $L^\infty(\Omega)$: $u \in U$ if $u \in L^\infty(\Omega)$ and if

$$\alpha \leqq u(x) \leqq \beta \quad \text{a.e. in } \Omega, \tag{1.1}$$

$$\int_\Omega u(x)\, dx = \gamma, \tag{1.2}$$

* Received by the editors April 10, 1969, and in revised form October 20, 1969. This research was supported in part by the Office of Naval Research under Contract NONR 233(76).

† U. E. R. Mathématiques et Informatique, 35-Rennes, France, and University of California, Los Angeles, California.

‡ Institute Automatyki, Polska Akademia Nauk, Warsaw, Poland, and University of California, Los Angeles, California.

where the numbers α, β, γ satisfy

$$0 < \alpha \leqq \beta < +\infty,$$
$$\alpha \operatorname{meas}(\Omega) < \gamma < \beta \operatorname{meas}(\Omega). \tag{1.3}$$

We denote by f a given element of $L^2(\Omega)$. Our aim is to study the following problem: Find

$$\max_{u \in U} \min_{y \in V} \left[\frac{1}{2} \int_\Omega u \operatorname{grad}^2 y \, dx - \int_\Omega fy \, dx \right], \tag{1.4}$$

where $V = H_0^1(\Omega)$.

We can transform this problem in the following way. If we fix $u \in U$, we know that there exists a unique $y \in V$ such that

$$\frac{1}{2} \int_\Omega u \operatorname{grad}^2 y \, dx - \int_\Omega fy \, dx \leqq \frac{1}{2} \int_\Omega u \operatorname{grad}^2 z \, dx - \int_\Omega fz \, dx \qquad \text{for all } z \in V;$$

furthermore, this y is the unique solution of

$$\int_\Omega u \operatorname{grad} y \cdot \operatorname{grad} \varphi \, dx = \int_\Omega f\varphi \, dx \quad \text{for all } \varphi \in V. \tag{1.5}$$

If we put $\varphi = y$ in (1.5) we obtain

$$\frac{1}{2} \int_\Omega u \operatorname{grad}^2 y \, dx - \int_\Omega fy \, dx = -\frac{1}{2} \int_\Omega fy \, dx.$$

Hence the problem (1.4) can be written in the following form: Find

$$\min_{u,y} \int_\Omega fy \, dx, \tag{1.6}$$

where u and y satisfy the "state equation"

$$\int_\Omega u \operatorname{grad} y \cdot \operatorname{grad} \varphi \, dx = \int_\Omega f\varphi \, dx \quad \text{for all } \varphi \in V$$

and $u \in U$.

In what follows we shall study problem (1.6).

DEFINITIONS. Problem (1.6) will be called Problem P. u is *admissible* means $u \in U$. u, y are *admissible* means $u \in U$, $y \in V$ and u, y are related by the state equation (1.5). y is *admissible* means there exists u such that u, y are admissible.

The cost function is

$$J(u) = \int_\Omega f(x)y(x) \, dx = \int_\Omega u(x) \operatorname{grad}^2 y(x) \, dx.$$

We shall say u, y is a *solution of Problem* P if u, y are admissible and

$$\int_\Omega fy \, dx \leqq \int_\Omega fz \, dx$$

for any admissible pair v, z.

2. Existence of a solution.

THEOREM 2.1. *Problem* P *has at least one solution.*[1]

Proof. First let us note that $J(u)$ is lower semicontinuous in $L^\infty(\Omega)$ with the weak* topology.

Indeed let

$$v_n \to v, \tag{2.1}$$

and y_n (respectively y) correspond to v_n (respectively v) in (1.5). Then we have

$$J(v) = \int_\Omega fy\,dx = \int_\Omega v \operatorname{grad}^2 y\,dx = \int_\Omega v_n \operatorname{grad} y_n \cdot \operatorname{grad} y\,dx$$

$$= 2\int_\Omega v_n \operatorname{grad} y_n \cdot \operatorname{grad} y\,dx - \int_\Omega v \operatorname{grad}^2 y\,dx,$$

$$J(v_n) = \int_\Omega v_n \operatorname{grad}^2 y_n\,dx.$$

Combining these two equations and adding and subtracting the term $\int_\Omega u_n \operatorname{grad}^2 y\,dx$ we obtain

$$J(v_n) - J(v) = \int_\Omega v_n \operatorname{grad}^2 y_n\,dx - 2\int_\Omega v_n \operatorname{grad} y_n \cdot \operatorname{grad} y\,dx$$

$$+ \int_\Omega v \operatorname{grad}^2 y\,dx + \int_\Omega v_n \operatorname{grad}^2 y\,dx - \int_\Omega v_n \operatorname{grad}^2 y\,dx$$

$$= \int_\Omega v_n \operatorname{grad}^2 (y_n - y)\,dx - \int_\Omega (v_n - v) \operatorname{grad}^2 y\,dx.$$

By (1.1) and (2.1) we have

$$\liminf J(v_n) - J(v) \geqq 0. \tag{2.2}$$

Hence $J(u)$ is weakly* lower semicontinuous. Now let us come back to the proof of Theorem 2.1.

Let us denote

$$j = \inf J(u), \qquad u \in U, \tag{2.3}$$

and let $\{u_n\} \subset U$ be a sequence minimizing $J(u)$, i.e.,

$$\lim J(u_n) = j.$$

Since the set U is bounded and closed in $L^\infty(\Omega)$, it is compact in the weak* topology of this space.

Therefore from $\{u_n\}$ we can extract a subsequence $\{u_{n'}\}$ convergent in the weak* topology to, say, $u \in U$:

$$u_{n'} \to u.$$

[1] It was pointed out to us by the referee that the theorem follows directly from a general abstract minimax theorem (cf. [4], [7]).

By (2.1) we have

$$J(u) \leqq \liminf J(u_{n'}) = \lim J(u_n) = j.$$

Hence by (2.3), u is a solution of Problem P.

3. The maximum principle.

THEOREM 3.1. *If the pair u, y is a solution of Problem* P, *then u, y satisfy the following maximum principle*:

$$\int_\Omega v \operatorname{grad}^2 y \, dx \leqq \int_\Omega u \operatorname{grad}^2 y \, dx \quad \text{for all } v \in U. \tag{3.1}$$

Proof. We use here a classical method. Let $v \in U$; denote by $y + \Delta y$ the function related to $u + \rho(v - u)$ by the state equation. We have

$$J(u + \rho(v - u)) = \int_\Omega f(y + \Delta y)\, dx = \int_\Omega u \operatorname{grad} y \cdot \operatorname{grad} (y + \Delta y)\, dx,$$

$$J(u + \rho(v - u)) = J(u) + \int_\Omega u \operatorname{grad} y \cdot \operatorname{grad} \Delta y \, dx. \tag{3.2}$$

But since

$$\int_\Omega (u + \rho(v - u)) \operatorname{grad} (y + \Delta y) \cdot \operatorname{grad} \varphi \, dx = \int_\Omega f\varphi \, dx \quad \text{for all } \varphi \in V,$$

we have

$$\begin{aligned} \rho \int_\Omega (v - u) \operatorname{grad} y \cdot \operatorname{grad} \varphi \, dx + \rho \int_\Omega (v - u) \operatorname{grad} \Delta y \cdot \operatorname{grad} \varphi \, dx \\ + \int_\Omega u \operatorname{grad} \Delta y \cdot \operatorname{grad} \varphi \, dx = 0; \end{aligned} \tag{3.3}$$

and if we put $\varphi = y$, it follows that

$$\int_\Omega u \operatorname{grad} y \cdot \operatorname{grad} \Delta y \, dx + \rho \int_\Omega (v - u) \operatorname{grad} (y + \Delta y) \cdot \operatorname{grad} y \, dx = 0.$$

Taking into consideration (3.2) we obtain

$$J(u + \rho(v - u)) = J(u) - \rho \int_\Omega (v - u) \operatorname{grad} (y + \Delta y) \cdot \operatorname{grad} y \, dx. \tag{3.4}$$

But the pair u, y is a solution of Problem P, i.e.,

$$J(u + \rho(v - u)) \geqq J(u) \quad \text{for all } \rho \in [0, 1];$$

hence

$$\int_\Omega (v - u) \operatorname{grad} (y + \Delta y) \cdot \operatorname{grad} y \, dx \leqq 0.$$

Now, if ρ tends to zero, Δy tends to 0 and we get (3.1).

Now we shall study the local consequences of the maximum principle (3.1).

THEOREM 3.2. *If the maximum of* $\int_\Omega v \operatorname{grad}^2 y\, dx$ *in* U (y *is fixed*) *is attained at* $v = u$, *then there exist a Lagrange multiplier* $\lambda \in \mathscr{R}$ *such that*

$$\int_\Omega u \operatorname{grad}^2 y\, dx - \lambda \int_\Omega u\, dx \geqq \int_\Omega v \operatorname{grad}^2 y\, dx - \lambda \int_\Omega v\, dx \tag{3.5}$$

for all $v \in L^\infty(\Omega)$ *such that* $\alpha \leqq v(x) \leqq \beta$.

Proof. Let us consider the linear continuous mapping $\phi(u)$ from $L^\infty(\Omega)$ into $\mathfrak{R}^2$ given by

$$\phi(v) = \left(\int_\Omega v \operatorname{grad}^2 y\, dx, \int_\Omega v\, dx \right) = (\xi_1, \xi_2). \tag{3.6}$$

Let $G \in \mathfrak{R}^2$ denote the image of the set $\{v \in L^\infty(\Omega) | \alpha \leqq v(x) \leqq \beta\}$ under the mapping (3.6). It is obvious that G is closed, convex and bounded in $\mathfrak{R}^2$.

To u corresponds the point $(\xi_1^0, \xi_2^0) \in G$ such that ξ_1^0 is maximal in G for $\xi_2^0 = \gamma$.

Hence (ξ_1^0, ξ_2^0) belongs to the boundary of G and there is a hyperplane supporting G at (ξ_1^0, ξ_2^0).

Let this hyperplane be defined by the vector $(1, -\lambda)$; then we have

$$\xi_1^0 - \lambda \xi_2^0 \geqq \xi_1 - \lambda \xi_2 \quad \text{for all } (\xi_1, \xi_2) \in G.$$

Using the definition (3.6) of the mapping $\phi(v)$, we obtain (3.5). Relation (3.5) can be written in the form

$$\int_\Omega u (\operatorname{grad}^2 y - \lambda)\, dx \geqq \int_\Omega v (\operatorname{grad}^2 y - \lambda)\, dx \tag{3.5'}$$

for all $v \in L^\infty(\Omega)$ such that $\alpha \leqq v(x) \leqq \beta$. This shows that

$$\begin{aligned} &\operatorname{grad}^2 y(x) < \lambda \quad \text{implies} \quad u(x) = \alpha, \\ &\operatorname{grad}^2 y(x) > \lambda \quad \text{implies} \quad u(x) = \beta. \end{aligned} \tag{3.7}$$

When $\operatorname{grad}^2 y(x) = \lambda$, we cannot say anything about u.

Remark 3.1. We have proved that (3.1) implies (3.7); however, it is not true in general that (3.7) implies (3.1). In practice, at least when $n = 1$, the set $\Delta = \{x | \operatorname{grad}^2 y(x) = \lambda, x \in \Omega\}$ has nonvoid interior. We can prove, however, Theorem 3.3.

LOCAL MAXIMUM PRINCIPLE. There exists $\lambda \in \mathfrak{R}$ such that

$$\begin{aligned} &\operatorname{grad}^2 y(x) < \lambda \quad \text{implies} \quad u(x) = \alpha, \\ &\operatorname{grad}^2 y(x) > \lambda \quad \text{implies} \quad u(x) = \beta. \end{aligned} \tag{3.8}$$

THEOREM 3.3. *If the admissible pair* u, y *satisfies the local maximum principle, then* u, y *satisfies the maximum principle* (3.1).

Proof. By virtue of (3.8) we have

$$\int_\Omega u (\operatorname{grad}^2 y - \lambda)\, dx \geqq \int_\Omega v (\operatorname{grad}^2 y - \lambda)\, dx$$

for all $v \in L^\infty(\Omega)$ such that $\alpha \leqq v(x) \leqq \beta$ a.e. in Ω.

but u is admissible; thus

$$\int_\Omega u \operatorname{grad}^2 y \, dx - \lambda\gamma \geqq \int_\Omega v \operatorname{grad}^2 y \, dx - \lambda \int_\Omega v(x) \, dx$$

for all $v \in L^\infty(\Omega)$ such that $\alpha \leqq v(x) \leqq \beta$ a.e. in Ω. And finally,

$$\int_\Omega u \operatorname{grad}^2 y \, dx \geqq \int_\Omega v \operatorname{grad}^2 y \, dx$$

for all $v \in U$; that is, the pair u, y satisfies the maximum principle.

4. Uniqueness of the solution y.

THEOREM 4.1. *If u, y is a solution of Problem* P *and if v, z is an admissible pair which satisfies the maximum principle, then*

$$z = y. \tag{4.1}$$

Proof. By hypothesis we have

$$\begin{aligned} &\int_\Omega u \operatorname{grad} y \cdot \operatorname{grad} \varphi \, dx = \int_\Omega f\varphi \, dx \quad \text{for all } \varphi \in V, \\ &\int_\Omega v \operatorname{grad} z \cdot \operatorname{grad} \varphi \, dx = \int_\Omega f\varphi \, dx \quad \text{for all } \varphi \in V; \end{aligned} \tag{4.2}$$

and the maximum principle (applied to the pair v, z) implies in particular that

$$\int_\Omega v \operatorname{grad}^2 z \, dx \geqq \int_\Omega u \operatorname{grad}^2 z \, dx. \tag{4.3}$$

Now, we have

$$\begin{aligned} &\int_\Omega u \operatorname{grad}^2 (y - z) \, dx \\ &\quad = \int_\Omega u \operatorname{grad}^2 y \, dx - 2 \int_\Omega u \operatorname{grad} y \cdot \operatorname{grad} z \, dx + \int_\Omega u \operatorname{grad}^2 z \, dx. \end{aligned}$$

Using (4.2) and (4.3) we obtain

$$\begin{aligned} &\int_\Omega u \operatorname{grad}^2 (y - z) \, dx \leqq \int_\Omega fy \, dx - 2 \int_\Omega fz \, dx + \int_\Omega v \operatorname{grad}^2 z \, dx, \\ &\int_\Omega u \operatorname{grad}^2 (y - z) \, dx \leqq \int_\Omega fy \, dx - \int_\Omega fz \, dx. \end{aligned}$$

But u, y is a solution of the Problem P; hence

$$\int_\Omega u \operatorname{grad}^2 (y - z) \, dx \leqq \int_\Omega fy \, dx - \int_\Omega fz \, dx \leqq 0$$

and finally

$$y = z.$$

COROLLARY 4.1. *If u, y and v, z are two pairs of solutions of Problem* P, *then*

$$y = z.$$

COROLLARY 4.2. *Any admissible pair which satisfies the maximum principle (local or not) is a solution of Problem* P.

Proof. Use Theorems 3.1 and 4.1

5. Approximation of the solution.

5.1. The algorithm. We shall build a sequence u_r, y_r using a gradient method. The choice of the descent is similar to the choice of Frank and Wolfe [3] (cf. Cea [1]) in the case of infinite-dimensional spaces; the choice of the point in the descent is particular to this problem.

Let u_0, y_0 be an admissible pair. Suppose u_r, y_r are computed and we propose to compute u_{r+1}, y_{r+1} as follows:

From (3.4) we have

$$\begin{aligned} \Delta J_r &= J(u_r) - J(u_r + \rho(v - u_r)) \\ &= \rho \int_\Omega (v - u_r) \operatorname{grad}^2 y_r \, dx + \rho \int_\Omega (v - u_r) \operatorname{grad} y_r \cdot \operatorname{grad} \Delta y_r \, dx, \end{aligned} \tag{5.1}$$

the largest part of ΔJ_r being

$$\rho \int_\Omega (v - u_r) \operatorname{grad}^2 y_r \, dx.$$

Then it is natural to select an element $v_r \in U$ such that

$$\int_\Omega v_r \operatorname{grad}^2 y_r \, dx \geqq \int_\Omega v \operatorname{grad}^2 y_r \, dx \quad \text{for all } v \in U. \tag{5.2}$$

In particular, this implies

$$\int_\Omega v_r \operatorname{grad}^2 y_r \, dx \geqq \int_\Omega u_r \operatorname{grad}^2 y_r \, dx. \tag{5.3}$$

If $v_r = u_r$, the pair u_r, y_r is admissible and, by (5.2), satisfies the maximum principle; hence u_r, y_r is a solution. *If* $v_r(x) = u_r(x)$ for all x such that $\operatorname{grad}^2 y_r(x) > 0$, we obtain the same conclusion as before.

Thus we have only to study the case

$$\int_\Omega (v_r - u_r) \operatorname{grad}^2 y_r \, dx > 0;$$

that is, the case where there exists a descent direction $u_r + \rho(v_r - u_r), \rho > 0$. If we choose $\varphi = \Delta y_r$ in (3.3) we obtain

$$\rho \int_\Omega (v_r - u_r) \operatorname{grad} y_r \cdot \operatorname{grad} \Delta y_r \, dx + \int_\Omega (u_r + \rho(v_r - u_r)) \operatorname{grad}^2 \Delta y_r \, dx = 0; \tag{5.4}$$

hence

$$\alpha \| \Delta y_r \|^2_{H_0^1(\Omega)} \leqq \rho \| v_r - u_r \|_{L^\infty(\Omega)} \cdot \| \Delta y_r \|_{H_0^1(\Omega)} \cdot \| y_r \|_{H_0^1(\Omega)}$$

and

$$\|\Delta y_r\|_{H_0^1(\Omega)} \leqq \frac{\rho}{\alpha}\|y_r\| \cdot \|v_r - u_r\|_{L^\infty(\Omega)}. \tag{5.5}$$

Using (5.1) and (5.4), we obtain

$$\Delta J_r = \rho \int_\Omega (v_r - u_r) \operatorname{grad}^2 y_r \, dx - \int_\Omega (u_r + \rho(v_r - u_r)) \operatorname{grad}^2 \Delta y_r \, dx;$$

but

$$\left| \int_\Omega (u_r + \rho(v_r - u_r)) \operatorname{grad}^2 \Delta y_r \, dx \right| \leqq \beta \|\Delta y_r\|^2_{H_0^1(\Omega)}$$

$$\leqq \rho^2 \frac{\beta}{\alpha^2} \|y_r\|^2_{H_0^1(\Omega)} \cdot \|v_r - u_r\|^2_{L^\infty(\Omega)}.$$

Hence

$$\Delta J_r \geqq \rho \int_\Omega (v_r - u_r) \operatorname{grad}^2 y_r \, dx - \rho^2 \frac{\beta}{\alpha^2} \|y_r\|^2_{H_0^1(\Omega)} \cdot \|v_r - u_r\|^2_{L^\infty(\Omega)}. \tag{5.6}$$

Let

$$\hat{\rho}_r = \frac{1}{2} \frac{\alpha^2}{\beta \|y_r\|^2_{H_0^1(\Omega)} \cdot \|v_r - u_r\|^2} \cdot \int_\Omega (v_r - u_r) \operatorname{grad}^2 y_r \, dx. \tag{5.7}$$

We choose now ρ_r as follows:

$$\rho_r = \min(\hat{\rho}_r, 1); \tag{5.8}$$

and we put

$$u_{r+1} = u_r + \rho_r(v_r - u_r). \tag{5.9}$$

y_{r+1} is related to u_{r+1} by the state equation.

ALGORITHM.

(i) Select an element $v_r \in U$, which verifies (5.2).

(ii) Compute $\hat{\rho}_r$ with (5.7) and ρ_r with (5.8).

(iii) Define u_{r+1} by (5.9) and compute y_{r+1} using the state equation.

Remark 5.1. We can seek v_r using a Lagrange multiplier λ_r (see Theorem 3.2).

Remark 5.2. We could choose $\rho_{r+1} = \tilde{\rho}_{r+1}$ where the latter is defined in the following way: $\tilde{\rho}_{r+1}$ satisfies

$$\tilde{u}_{r+1} = u_r + \tilde{\rho}_{r+1}(v_r - u_r) \in U,$$

$$J(\tilde{u}_{r+1}) \leqq J(u_r + \rho(v_r - u_r))$$

for all ρ such that $u_r + \rho(v_r - u_r) \in U$. The difficulty lies in finding this $\tilde{\rho}_{r+1}$: the cost of the search would be very high.

5.2. Convergence. We have the following inequalities.

Case 1. $\rho_r = 1$. Then

$$J(u_r) - J(u_{r+1}) \geqq \int_\Omega (v_r - u_r) \operatorname{grad}^2 y_r \, dx - \frac{\beta}{\alpha^2} \|y_r\|^2_{H_0^1(\Omega)} \cdot \|v_r - u_r\|^2_{L^\infty(\Omega)},$$

(5.10) $\hat{\rho}_r \geqq 1$, or

$$\alpha^2 \int_\Omega (v_r - u_r) \operatorname{grad}^2 y_r \, dx \geqq 2\beta \|y_r\|^2_{H_0^1(\Omega)} \cdot \|v_r - u_r\|_{L^\infty(\Omega)}.$$

Case 2. $\rho_r = \hat{\rho}_r$. Then

$$(5.10') \qquad J(u_r) - J(u_{r+1}) \geqq \frac{\alpha^2 \left[\int_\Omega (v_r - u_r) \operatorname{grad}^2 y_r \, dx \right]^2}{4\beta \|y_r\|^2_{H_0^1(\Omega)} \cdot \|v_r - u_r\|^2_{L^\infty(\Omega)}}.$$

We know that the sequence $J(u_r)$ is decreasing. Since $J(u_r) \geqq 0$ for all r, it follows that

$$\lim_{r \to \infty} [J(u_r) - J(u_{r+1})] = 0;$$

and then, using (5.10) and (5.10'), we have the next case.

Case 3. $\rho_r = 1$. In this case,

$$(5.11) \qquad \lim_{r \to \infty} \|v_r - u_r\|_{L^\infty(\Omega)} = 0.$$

Case 4. $\rho_r = \hat{\rho}_r$. Then

$$(5.11') \qquad \lim_{r \to \infty} \int_\Omega (v_r - u_r) \operatorname{grad}^2 y_r \, dx = 0$$

(we disregard the trivial case where $\lim y_r = 0$). But we know that

$$\int_\Omega v_r \operatorname{grad}^2 y_r \, dx \geqq \int_\Omega v \operatorname{grad}^2 y_r \, dx \quad \text{for all } v \in U;$$

hence

$$\int_\Omega u_r \operatorname{grad}^2 y_r \, dx + \int_\Omega (v_r - u_r) \operatorname{grad}^2 y_r \, dx \geqq \int_\Omega v \operatorname{grad}^2 y_r \, dx$$

or

$$(5.12) \qquad \int_\Omega f y_r \, dx + \int_\Omega (v_r - u_r) \operatorname{grad}^2 y_r \, dx \geqq \int_\Omega v \operatorname{grad}^2 y_r \, dx \quad \text{for all } v \in U.$$

Since the sequences u_r and y_r are bounded, we can extract some subsequences $u_{r'}, y_{r'}$ such that:

$$(5.13) \qquad \begin{aligned} &u_{r'} \to u \quad \text{in } L^\infty(\Omega) \text{ with the weak* topology, } u \in U, \\ &y_{r'} \to y \quad \text{in } H_0^1(\Omega) \text{ with the weak topology.} \end{aligned}$$

From (5.11), (5.12) and (5.13) it follows that

$$\int_\Omega fy\,dx \geqq \int_\Omega v\,\mathrm{grad}^2 y\,dx \quad \text{for all } v \in U. \tag{5.14}$$

But

$$\begin{aligned}\int_\Omega u_{r'}\,\mathrm{grad}^2(y_{r'} - y)\,dx &= \int_\Omega u_{r'}\,\mathrm{grad}^2 y_{r'}\,dx \\ &\quad + \int_\Omega u_{r'}\,\mathrm{grad}^2 y\,dx - 2\int_\Omega u_{r'}\,\mathrm{grad}\,y_{r'}\cdot\mathrm{grad}\,y\,dx \\ &= \int_\Omega fy_{r'}\,dx - 2\int_\Omega fy\,dx + \int_\Omega u_{r'}\,\mathrm{grad}^2 y\,dx.\end{aligned}$$

When $r' \to +\infty$ we have

$$\lim_{r'\to\infty}\int_\Omega u_{r'}\,\mathrm{grad}^2(y_{r'} - y)\,dx = -\int_\Omega fy\,dx + \int_\Omega u\,\mathrm{grad}^2 y\,dx;$$

this together with (5.14) implies

$$\lim_{r'\to\infty}\int_\Omega u_{r'}\,\mathrm{grad}^2(y_{r'} - y)\,dx = 0.$$

However $u_{r'}(x) \geqq \alpha$ a.e. in Ω; therefore

$$\lim_{r'\to\infty}\|y_{r'} - y\|_{H_0^1(\Omega)} = 0. \tag{5.15}$$

Let us prove that u, y is admissible. We have

$$\int_\Omega u_{r'}\,\mathrm{grad}\,y_{r'}\cdot\mathrm{grad}\,\varphi\,dx = \int_\Omega f\varphi\,dx \quad \text{for all } \varphi \in H_0^1(\Omega)$$

or

$$\int_\Omega u_{r'}\,\mathrm{grad}\,(y_{r'} - y)\cdot\mathrm{grad}\,\varphi\,dx + \int_\Omega u_{r'}\,\mathrm{grad}\,y\cdot\mathrm{grad}\,\varphi\,dx = \int_\Omega f\varphi\,dx.$$

Using (5.15), (5.13) and the inequality $u_{r'}(x) \leqq \beta$, we have

$$\int_\Omega u\,\mathrm{grad}\,y\cdot\mathrm{grad}\,\varphi\,dx = \int_\Omega f\varphi\,dx \quad \text{for all } \varphi \in H_0^1(\Omega). \tag{5.16}$$

Thus u, y is admissible. But u, y satisfies the maximum principle. By Corollary 4.2 we know that u, y is a solution of the Problem P. Because we have proved the uniqueness of y, by a classical argument it follows that the entire sequence y_r converges to y. Thus we have proved the following theorem.

THEOREM 5.1. *The iterative method of* § 5.1 *is convergent*: *that is*,

$$\lim_{r\to\infty}\|y_r - y\|_{H_0^1(\Omega)} = 0.$$

Remark 5.3. We could simplify the computation of $\hat{\rho}_r$, introducing bounds independent of r.

6. An example. We give here an example for the case $n = 1$.

Let $V = H_0^1(\Omega)$ and $\Omega =]0, a[$, where

$$a = \frac{21 + \sqrt{278}}{10}.$$

We define the functions y, f and u as follows:

for $0 \leqq x \leqq 1$,

$$y(x) = -x, \qquad f(x) = 0, \qquad u(x) = 1;$$

for $1 < x \leqq 31/10$,

$$y(x) = \frac{1}{2}x^2 - 2x + \frac{1}{2}, \qquad f(x) = -1, \qquad u(x) = 1;$$

for $31/10 < x \leqq 32/10$,

$$y(x) = \left(\frac{11}{10}\right)\left(x - \frac{31}{10}\right) - \frac{179}{200}, \quad f(x) = -220\left(x - \frac{31}{10}\right), \quad u(x) = 100\left(x - \frac{31}{10}\right)^2 + 1;$$

for $32/10 < x \leqq a$,

$$y(x) = \frac{1}{2}x^2 - \frac{21}{10}x + \frac{163}{200}, \qquad f(x) = -2, \qquad u(x) = 2.$$

We could verify that

$$y \in C^1(\bar{\Omega}), \qquad u \in C^0(\bar{\Omega}),$$

$$\int_0^a uy'\varphi'\,dx = \int_0^a f\varphi\,dx \quad \text{for all } \varphi \in V \tag{6.1}$$

and

$$1 \leqq u(x) \leqq 2,$$

$$\int_0^a \left(y'^2 - \left(\frac{11}{10}\right)^2\right)v\,dx \leqq \int_0^a \left(y'^2 - \left(\frac{11}{10}\right)^2\right)u\,dx \tag{6.2}$$

for all v such that $1 \leqq v(x) \leqq 2$. We define γ by

$$\gamma = \int_0^a u(x)\,dx. \tag{6.3}$$

Using (6.1), (6.2), (6.3), we can prove that u, y is a solution of the problem

$$\max_{v \in U} \min_{y \in H_0^1(\Omega)} \left[\frac{1}{2}\int_0^a vy'^2\,dx - \int_0^a fy\,dx\right],$$

where

$$v \in U \text{ if and only if } v \in L^\infty(\Omega), \quad 1 \leqq v(x) \leqq 2, \quad \int_0^a v(x)\,dx = \gamma.$$

Note that in the interval (31/10, 32/10), we have

$$y'^2 = \left(\frac{11}{10}\right)^2.$$

In this example the value of the Lagrange multiplier is $(11/10)^2$.

Acknowledgment. The authors wish to acknowledge Mrs. Barbara Klosowicz for introducing the problem considered in this paper to one of us. We are also grateful to the referee, whose comments helped us very much to improve the paper.

REFERENCES

[1] J. Dea, *Les methodes de descente dans la théorie de l'optimisation*, Rev. A.F.I.R.O., Série Rouge, no. 13, 1969.

[2] R. J. Duffin and D. K. McLain, *Optimum shape of a cooling fin on a convex cylinder*, J. Math. Mech., 17 (1968), pp. 769–784.

[3] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[4] Ky Fan, *Sur un théorème minimax*, C. R. Acad. Sci. Paris, 259 (1964), pp. 3925–3928.

[5] J. L. Lions, *Controle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Gauthier-Villars, Paris, 1968.

[6] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes*, Dunod, Paris, 1968.

[7] M. Sion, *On general minimax theorems*, Pacific J. Math., 8 (1958), pp. 171–176.

SIAM J. CONTROL
Vol. 8, No. 3, August 1970

DECOUPLING AND POLE ASSIGNMENT BY DYNAMIC COMPENSATION*

A. S. MORSE† AND W. M. WONHAM‡

Introduction. In a previous article [1] the authors defined in geometric terms the decoupling problem for a constant linear multivariable system: namely, the problem of achieving independent control of specified outputs by means of suitably combined inputs and of suitable linear state variable feedback. Necessary and sufficient conditions for decoupling to be possible were found in two important cases; but the general problem is unsolved. However, if in addition to state feedback, dynamic (integrator) compensation may be utilized, it becomes possible to state general necessary and sufficient conditions for decoupling in a simple and constructive way. Geometrically the decoupling synthesis amounts to extending the state space of the original system to a larger space, the increase in dimension being the number of integrators used in dynamic compensation. In addition, state space extension can be used to achieve a desired pole distribution for the closed loop system transfer matrix.

The possibility of exploiting dynamic compensation in the decoupling problem was illustrated by an example in [2].

In the present article we state and solve the extended decoupling problem (§ 1). Under certain restrictions, the problem of minimizing the order of dynamic compensation (i.e., the dimension of the extended state space) is solved in § 2. This solution is actually the best possible if the number of scalar inputs is equal to the number of output blocks to be decoupled (§ 3). In § 4 the role of state space extension in pole assignment is determined. It is shown that, with dynamic compensation of high enough order, any pole distribution can be synthesized for the decoupled system, whenever decoupling is possible at all. An example is given in § 5. In conclusion (§ 6) a more general view of decoupling is taken, with the restriction to linear compensation relaxed. The resulting open loop decoupling problem is shown to be equivalent, however, to the extended decoupling problem of § 1.

In the sequel, the material in [1] is assumed to be known.

Notation. Script letters $\mathscr{E}$, $\mathscr{E}'$, $\mathscr{R}$, $\mathscr{N}$, $\cdots$ denote vector spaces over the real numbers, with elements $x, y, \cdots$; $d(\mathscr{E})$ is the dimension of $\mathscr{E}$; $\mathscr{U} \approx \mathscr{V}$ means $\mathscr{U}$, $\mathscr{V}$ are isomorphic, i.e., $d(\mathscr{U}) = d(\mathscr{V})$. $A, B, C, \cdots$ are linear maps; $A|\mathscr{R}$ is the restriction of A to $\mathscr{R}$; $\mathscr{B}$ or $\{B\}$ is the range of B. *Spectrum* means complex spectrum. A *symmetric* set of complex numbers is one of the form

$$\{\alpha_1, \alpha_2, \cdots; \beta_1, \bar{\beta}_1; \beta_2, \bar{\beta}_2; \cdots\},$$

where the α_i are real and $\bar{\beta}_i$ is the complex conjugate of β_i. $\mathscr{N}(H)$ is the kernel (null space) of H.

* Received by the editors December 11, 1969.

† Electronics Research Center, National Aeronautics and Space Administration, Cambridge, Massachusetts 02139.

‡ Electronics Research Center, National Aeronautics and Space Administration, Cambridge, Massachusetts 02139. This article was written while this author held an NRC Postdoctoral Resident Research Associateship, supported by the National Aeronautics and Space Administration.

With A, B, E' fixed, $\mathbf{C}(\mathscr{V})$ is the set of maps C such that $(A + BC)\mathscr{V} \subset \mathscr{V}$, $\mathbf{C}'(\mathscr{V})$ the set of C such that $(A + (B + E')C)\mathscr{V} \subset \mathscr{V}$. $\mathscr{I}$ (respectively $\mathscr{I}'$) is the class of $\mathscr{V}$ such that $\mathbf{C}(\mathscr{V}) \neq \varnothing$ (respectively $\mathbf{C}'(\mathscr{V}) \neq \varnothing$). If $d(\mathscr{E}) = n$, $A : \mathscr{E} \to \mathscr{E}$ and $\mathscr{B} \subset \mathscr{E}$, then

$$\{A|\mathscr{B}\} \equiv \sum_{j=1}^{n} A^{j-1}\mathscr{B}.$$

$\mathscr{R} \subset \mathscr{E}$ is a *controllability subspace* (c.s.) for the pair (A, B), written $\mathscr{R} \in \mathscr{C}$, if $\mathbf{C}(\mathscr{R}) \neq \varnothing$ and if, for some $C \in \mathbf{C}(\mathscr{R})$,

$$\mathscr{R} = \{A + BC|\mathscr{B} \cap \mathscr{R}\};$$

$\mathscr{R}$ is determined uniquely, as written, by any $C \in \mathbf{C}(\mathscr{R})$. Similarly $\mathscr{S}$ is a c.s. for $(A, B + E')$, written $\mathscr{S} \in \mathscr{C}'$, if $\mathbf{C}'(\mathscr{S}) \neq \varnothing$ and

$$\mathscr{S} = \{A + (B + E')C|(\mathscr{B} + \mathscr{E}') \cap \mathscr{S}\}, \qquad C \in \mathbf{C}'(\mathscr{S}).$$

The maximal (i.e., largest) element of $\mathscr{I}$ (respectively $\mathscr{C}$) contained in a subspace $\mathscr{T}$ is denoted by $\max(\mathscr{I}, \mathscr{T})$ (respectively $\max(\mathscr{C}, \mathscr{T})$), and similarly for $\mathscr{I}', \mathscr{C}'$. It is known from [1] that these maximal elements exist and are unique for each fixed $\mathscr{T}$ and that, if $\mathscr{V} = \max(\mathscr{I}, \mathscr{T})$, then

$$\max(\mathscr{C}, \mathscr{T}) = \{A + BC|\mathscr{B} \cap \mathscr{V}\}, \qquad C \in \mathbf{C}(\mathscr{V}).$$

J is the set of integers $1, \cdots, k$. Unless otherwise noted, all summations and intersections are over J. If $\mathscr{R}_i$, $i \in J$, is a family of subspaces,

$$\mathscr{R}_i^* \equiv \sum_{j \neq i} \mathscr{R}_j, \qquad \mathscr{R}^* \equiv \bigcap_i \mathscr{R}_i^*,$$

$$\Delta[\mathscr{R}_i, J] \equiv \sum_i d(\mathscr{R}_i) - d(\sum_i \mathscr{R}_i).$$

Certain auxiliary results needed are collected in the Appendix.

1. Extended decoupling problem. As in [1] the control system is specified by the differential equation

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1.1}$$

and output relations

$$y_i(t) = H_i x(t), \qquad i \in J. \tag{1.2}$$

The state vector $x \in \mathscr{E}$, $d(\mathscr{E}) = n$; the control vector $u \in \mathscr{U}$, $d(\mathscr{U}) = m$; the output vector $y_i \in \mathscr{Y}_i$, $d(\mathscr{Y}_i) = q_i$. The maps A, B, H_i are independent of t; in fact, (1.1) qua differential equation plays no role until § 6, as our problem is purely algebraic.

Write $\mathscr{N}_i \equiv \mathscr{N}(H_i)$, $i \in J$; as in [1] we assume $\mathscr{N}_i \neq \mathscr{E}$, $i \in J$. In [1] we discussed the *restricted decoupling problem* (RDP): Find $\mathscr{R}_i \in \mathscr{C}$, $i \in J$, such that

$$\bigcap_i \mathbf{C}(\mathscr{R}_i) \neq \varnothing, \tag{1.3}$$

$$\mathscr{R}_i \subset \bigcap_{j \neq i} \mathscr{N}_j, \qquad i \in J, \tag{1.4}$$

$$\mathscr{R}_i + \mathscr{N}_i = \mathscr{E}, \qquad i \in J. \tag{1.5}$$

A family of c.s. $\mathscr{R}_i \in \mathscr{C}$, $i \in J$, which satisfies (1.4) (but not necessarily (1.3) or (1.5)) is *admissible*. Let $\mathscr{R}_i^M$ be the maximal admissible c.s. ($\mathscr{R}_i^M$ was denoted by $\overline{\mathscr{R}}_i$ in [1]). It is clear that RDP is solvable only if

$$\mathscr{R}_i^M + \mathscr{N}_i = \mathscr{E}, \qquad i \in J. \tag{1.6}$$

In general, (1.6) is not sufficient for solvability of RDP because (1.3) may fail for the $\mathscr{R}_i^M$, i.e., there may not exist any C such that $(A + BC)\mathscr{R}_i^M \subset \mathscr{R}_i^M$, $i \in J$. To surmount this difficulty we introduce an extended decoupling problem as follows.

Adjoin to (1.1) the equation of a new dynamic element:

$$\dot{x}'(t) = I'u'(t), \tag{1.7}$$

where $x' \in \mathscr{E}'$, $u' \in \mathscr{U}'$, $d(\mathscr{E}') = d(\mathscr{U}') = n'$ and $I' : \mathscr{U}' \approx \mathscr{E}'$; the input $u'(\cdot)$ can be freely chosen. For the system (1.1) extended by (1.7) define the state space

$$\mathscr{E}^e = \mathscr{E} \oplus \mathscr{E}'$$

and the extended input space

$$\mathscr{U}^e = \mathscr{U} \oplus \mathscr{U}'.$$

Define extensions A^e, B^e, E' of A, B, I' as follows:

$$\begin{aligned} &A^e : \mathscr{E}^e \to \mathscr{E}^e;\ A^e(x + x') \equiv Ax && (x \in \mathscr{E}, x' \in \mathscr{E}'), \\ &B^e : \mathscr{U}^e \to \mathscr{E}^e;\ B^e(u + u') \equiv Bu && (u \in \mathscr{U}, u' \in \mathscr{U}') \\ &E' : \mathscr{U}^e \to \mathscr{E}^e;\ E'(u + u') \equiv I'u' && (u \in \mathscr{U}, u' \in \mathscr{U}'). \end{aligned} \tag{1.8}$$

Below we write A, B for A^e, B^e; x for vectors in $\mathscr{E}^e$; and P for the projection $\mathscr{E} \oplus \mathscr{E}' \to \mathscr{E}$. Observe that $PA = AP = A, PB = B, P\mathscr{E}' = 0$. The combined system (1.1), (1.7) is now specified by the pair $(A, B + E')$.

The *extended decoupling problem* (EDP) is the following: *Given the original maps* $A : \mathscr{E} \to \mathscr{E}$, $B : \mathscr{U} \to \mathscr{E}$, *and subspaces* $\mathscr{N}_i \subset \mathscr{E}$, $i \in J$, *find*: (i) $\mathscr{E}'$ (*that is*, n'), (ii) *extensions* A, B, E', *as in* (1.8), (iii) $\mathscr{S}_i \in \mathscr{C}'$, $i \in J$, *with the properties*

$$\bigcap_i \mathbf{C}'(\mathscr{S}_i) \neq \varnothing, \tag{1.9}$$

$$\mathscr{S}_i \subset \bigcap_{j \neq i} (\mathscr{N}_j \oplus \mathscr{E}'), \qquad i \in J, \tag{1.10}$$

$$\mathscr{S}_i + (\mathscr{N}_i \oplus \mathscr{E}') = \mathscr{E} \oplus \mathscr{E}', \qquad i \in J. \tag{1.11}$$

It is clear that the choice of isomorphism I', and so of E' in (1.8), can be arbitrary after n' is fixed: for instance, $I' = n' \times n'$ identity matrix, in the coordinates selected.

EDP has the same structure in $\mathscr{E}^e$ as RDP has in $\mathscr{E}$, but flexibility is gained from the special form of the extended system map A and constraint spaces $\mathscr{N}_i \oplus \mathscr{E}'$. Justification of EDP as the correct description of decoupling by dynamic compensation is clear: the output relations (1.2) are preserved on replacing $\mathscr{N}_i$ by $\mathscr{N}_i \oplus \mathscr{E}'$ (equivalently by defining extensions H_i^e of H_i to be zero on $\mathscr{E}'$); no

additional control inputs (B) to the original system (1.1) are postulated; subject to the latter constraint, full linear coupling is allowed between (1.1) and (1.7).

Our main result (Theorem 1.1) states that decoupling by dynamic compensation is possible if and only if the maximal admissible c.s. $\mathscr{R}_i^M$ of RDP are sufficiently large.

THEOREM 1.1. *For the RDP of* (1.3)–(1.5), *let* $\mathscr{R}_i^M$ *be the maximal admissible c.s. in* $\mathscr{C}$. *The corresponding EDP of* (1.9)–(1.11) *is solvable if and only if*

$$\mathscr{R}_i^M + \mathscr{N}_i = \mathscr{E}, \qquad i \in J. \tag{1.6}$$

Proof. (i) (Only if) We show first that $\mathscr{S} \in \mathscr{C}'$ implies $\mathscr{R} \equiv P\mathscr{S} \in \mathscr{C}$. Since $\mathbf{C}'(\mathscr{S}) \neq \varnothing$, $A\mathscr{S} \subset \mathscr{S} + \mathscr{B} + \mathscr{E}'$, and $A\mathscr{R} = PA\mathscr{S} \subset \mathscr{R} + \mathscr{B}$, so that $\mathbf{C}(\mathscr{R}) \neq \varnothing$. Also, by Theorem 4.1 of [1], $\mathscr{S} = \lim \mathscr{S}^\mu$, $\mu = 0, 1, 2, \cdots$, where $\mathscr{S}^0 = 0$, $\mathscr{S}^{\mu+1} = \mathscr{S} \cap (A\mathscr{S}^\mu + \mathscr{B} + \mathscr{E}')$. Write $\mathscr{R}^\mu \equiv P\mathscr{S}^\mu$. Since $\mathscr{N}(P) = \mathscr{E}'$, Proposition A.5 implies

$$\mathscr{R}^{\mu+1} = P\mathscr{S}^{\mu+1} = \mathscr{R} \cap (A\mathscr{R}^\mu + \mathscr{B});$$

again by [1, Theorem 4.1], $\mathscr{R} = \lim \mathscr{R}^\mu \in \mathscr{C}$. Thus $\mathscr{S}_i \in \mathscr{C}'$, $i \in J$, implies $P\mathscr{S}_i \in \mathscr{C}$, $i \in J$; and (1.10), (1.11) yield

$$P\mathscr{S}_i \subset \bigcap_{j \neq i} \mathscr{N}_j, \qquad i \in J, \tag{1.12}$$

$$P\mathscr{S}_i + \mathscr{N}_i = \mathscr{E}, \qquad i \in J. \tag{1.13}$$

By (1.12) and maximality of the $\mathscr{R}_i^M$, $P\mathscr{S}_i \subset \mathscr{R}_i^M$: this and (1.13) imply (1.6).

(ii) (If) Assuming (1.6) holds, define $n' = \sum_i d(\mathscr{R}_i^M)$. With n' so large, there clearly exist maps $M_i : \mathscr{E}^e \to \mathscr{E}'$ with the properties:

$$\mathscr{R}_i^M \cap \mathscr{N}(M_i) = 0, \qquad \{M_i\} = M_i\mathscr{R}_i^M, \qquad i \in J,$$

and the ranges $\{M_i\}$, $i \in J$, independent. Define $\mathscr{S}_i = (P + M_i)\mathscr{R}_i^M$, $i \in J$. Then

$$A\mathscr{S}_i = A\mathscr{R}_i^M \subset \mathscr{R}_i^M + \mathscr{B} \subset \mathscr{S}_i + \mathscr{B} + \mathscr{E}', \qquad i \in J;$$

and since the $\mathscr{S}_i$ are clearly independent, there exists $C : \mathscr{E}^e \to \mathscr{U}^e$ with $C \in \bigcap_j \mathbf{C}'(\mathscr{S}_j)$. It will be shown that $\mathscr{S}_i \in \mathscr{C}'$. Dropping the subscript i, suppose $\mathscr{R} \in \mathscr{C}$, so that the relations

$$\mathscr{R}^0 \equiv 0, \qquad \mathscr{R}^{\mu+1} \equiv \mathscr{R} \cap (A\mathscr{R}^\mu + \mathscr{B}), \qquad \mu = 0, 1, \cdots,$$

imply $\mathscr{R}_\mu \uparrow \mathscr{R}$. Let $\{M\} \subset \mathscr{E}'$ and

$$\mathscr{S} \equiv (P + M)\mathscr{R}, \quad \mathscr{S}^0 \equiv 0, \quad \mathscr{S}^{\mu+1} \equiv \mathscr{S} \cap (A\mathscr{S}^\mu + \mathscr{B} + \mathscr{E}').$$

Then

$$\begin{aligned}
\mathscr{S}^0 &\supset (P + M)\mathscr{R}^0; \text{ and if } \mathscr{S}^\mu \supset (P + M)\mathscr{R}^\mu, \\
\mathscr{S}^{\mu+1} &\supset [(P + M)\mathscr{R}] \cap [A(P + M)\mathscr{R}^\mu + \mathscr{B} + \mathscr{E}'] \\
&= [(P + M)\mathscr{R}] \cap [A\mathscr{R}^\mu + \mathscr{B} + \mathscr{E}'] \\
&\supset (P + M)[\mathscr{R} \cap (A\mathscr{R}^\mu + \mathscr{B} + \mathscr{E}')] \\
&= (P + M)\mathscr{R}^{\mu+1}.
\end{aligned}$$

By induction, $\mathscr{S} \supset \mathscr{S}^{\mu} \supset (P + M)\mathscr{R}^{\mu} \uparrow (P + M)\mathscr{R} = \mathscr{S}$, i.e., $\mathscr{S}^{\mu} \uparrow \mathscr{S}$; so $\mathscr{S} \in \mathscr{C}'$. Application of this argument to the $\mathscr{R}_i^M$ and $\mathscr{S}_i$ yields the desired result.

The relation $P\mathscr{S}_i = \mathscr{R}_i^M$ implies

$$\mathscr{S}_i \subset \mathscr{R}_i^M \oplus \mathscr{E}' \subset (\bigcap_{j \neq i} \mathscr{N}_j) + \mathscr{E}' = \bigcap_{j \neq i} (\mathscr{N}_j \oplus \mathscr{E}'). \tag{1.10}$$

By (1.6),

$$\mathscr{S}_i + (P + M_i)\mathscr{N}_i = (P + M_i)\mathscr{E}$$

and addition of $\mathscr{E}'$ to both sides yields (1.11).

Remark 1. The proof reveals the symmetry between a c.s. and its extension: if $\mathscr{R} \in \mathscr{C}$ and $\mathscr{S} \equiv (P + M)\mathscr{R}$ with $\{M\} \subset \mathscr{E}'$, then $\mathscr{S} \in \mathscr{C}'$. Conversely, if $\mathscr{S} \in \mathscr{C}'$, then $\mathscr{R} \equiv P\mathscr{S} \in \mathscr{C}$.

Remark 2. In part 2 of the proof the $\mathscr{S}_i$ were constructed to be independent. By [1, Theorem 4.2], $C \in \bigcap_i \mathbf{C}'(\mathscr{S}_i)$ can be chosen such that, for each i, the spectrum of $(A + (B + E')C)|\mathscr{S}_i$ is any symmetric set of $d(\mathscr{S}_i)$ complex numbers.

Remark 3. Condition (1.6) is not implied by controllability of (A, B), i.e., by the condition $\{A|\mathscr{B}\} = \mathscr{E}$. For example, let

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$H_1 = (1, 0, 0), \qquad H_2 = (0, 1, 0).$$

By the methods of [1], one finds

$$\mathscr{R}_1^M = \mathscr{R}_2^M = \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}, \tag{1.14}$$

and (1.6) fails for $i = 1, 2$.

The following description of the structure of a decoupled system will be applied later to solutions of EDP. The result is stated for RDP for simplicity of notation. Let $\mathscr{R}_i$, $i \in J$, be any solution of RDP, write $\mathscr{R} \equiv \sum_i \mathscr{R}_i$, and write $A_c \equiv A + BC$ for $C \in \mathbf{C} \equiv \bigcap_i \mathbf{C}(\mathscr{R}_i)$. Let $\bar{x}$ be the coset of x in $\mathscr{E}/\mathscr{R}^*$. Noting that $A_c\mathscr{R}^* \subset \mathscr{R}^*$, we define the induced map $\bar{A}_c : \mathscr{E}/\mathscr{R}^* \to \mathscr{E}/\mathscr{R}^*$ by $\bar{A}_c\bar{x} \equiv \overline{A_c x}$.

THEOREM 1.2. *There exist $\hat{\mathscr{R}}_i \subset \mathscr{R}_i$, $i \in J$, independent of $C \in \mathbf{C}$ such that*

$$\mathscr{R} \equiv \mathscr{R}^* \oplus \hat{\mathscr{R}}_1 \oplus \cdots \oplus \hat{\mathscr{R}}_k \tag{1.15}$$

and

$$\hat{\mathscr{R}}_i \approx \bar{\mathscr{R}}_i \equiv (\mathscr{R}_i + \mathscr{R}^*)/\mathscr{R}^*. \tag{1.16}$$

The $\hat{\mathscr{R}}_i$ satisfy

$$\hat{\mathscr{R}}_i + \mathscr{N}_i = \mathscr{E}, \qquad i \in J, \tag{1.17}$$

$$A_c\hat{\mathscr{R}}_i \subset \hat{\mathscr{R}}_i \oplus \mathscr{R}^*, \qquad i \in J, \quad C \in \mathbf{C}. \tag{1.18}$$

The spectrum of $\bar{A}_c|\bar{\mathscr{R}}_i$, $i \in J$, can be assigned as any symmetric set of $d(\bar{\mathscr{R}}_i)$ complex numbers by suitable choice of $C \in \mathbf{C}$.

Proof. Let $\hat{\mathscr{R}}_i$ be any subspace such that $\mathscr{R}_i = \hat{\mathscr{R}}_i \oplus \mathscr{R}_i \cap \mathscr{R}^*$. Independence of $\mathscr{R}^*, \hat{\mathscr{R}}_i$, $i \in J$, follows by Proposition A.1; hence (1.15) is true, and (1.16) is clear. Since

$$\mathscr{R}^* \subset \bigcap_i \sum_{j \neq i} \bigcap_{\alpha \neq j} \mathscr{N}_a = \bigcap_i \mathscr{N}_i,$$

(1.17) follows from (1.4) and (1.5). Since $A_c\mathscr{R}_i \subset \mathscr{R}_i$, (1.18) is clear.

Let $C_0 \in \mathbf{C}$ be fixed; write $A_0 \equiv A_{c_0}$, $\bar{A}_0 \equiv \bar{A}_{c_0}$; and let Q be the projection: $\mathscr{E} \to \mathscr{E}/\mathscr{R}^*$; thus $\bar{A}_0 Q = QA_0$. Let $B_i : \mathscr{U} \to \mathscr{E}$ be any map with range $\mathscr{B} \cap \mathscr{R}_i$, and write $\bar{B}_i \equiv QB_i$, $\bar{\mathscr{B}}_i \equiv Q(\mathscr{B} \cap \mathscr{R}_i)$. It will be shown that $\bar{\mathscr{R}}_i$ is a c.s. for the pair $(\bar{A}_0, \bar{B}_i)$. In fact,

$$\begin{aligned} \bar{\mathscr{R}}_i = Q\mathscr{R}_i &= Q\{A_0|\mathscr{B} \cap \mathscr{R}_i\}, \\ &= \{\bar{A}_0|Q(\mathscr{B} \cap \mathscr{R}_i)\}, \\ &= \{\bar{A}_0|\bar{\mathscr{B}}_i\} \end{aligned}$$

and the assertion follows. By Proposition A.1, the $\bar{\mathscr{R}}_i$ are independent. Hence [1, §§ 4, 6] there exist $\bar{D}_i : \mathscr{E}/\mathscr{R}^* \to \mathscr{U}$, $i \in J$, such that $\bar{D}_i\bar{\mathscr{R}}_j = 0$, $i, j \in J$, $j \neq i$, $(\bar{A}_0 + \bar{B}_i\bar{D}_i)\bar{\mathscr{R}}_i \subset \bar{\mathscr{R}}_i$, $i \in J$, and $(\bar{A}_0 + \bar{B}_i\bar{D}_i)|\bar{\mathscr{R}}_i$, $i \in J$, has any preassigned spectrum. Define $D_i = \bar{D}_i Q$, $i \in J$. Then $D_i(\mathscr{R}_j + \mathscr{R}^*) = 0$, $i, j \in J$; $j \neq i$, and $B_iD_i\mathscr{R}_i \subset \mathscr{B}_i \subset \mathscr{R}_i$. Let $D : \mathscr{E} \to \mathscr{U}$ be any map such that $BD = \sum_i B_iD_i$; D exists since $\{B_i\} \subset \mathscr{B}$, $i \in J$. Then the map $C : \mathscr{E} \to \mathscr{U}$ defined by $C = C_0 + D$ has the properties required.

2. Minimal state space extension. Theorem 1.1 shows that if (1.6) holds, EDP can always be solved by dynamic compensation of order $n' \leqq \sum_i d(\mathscr{R}_i^M)$. There is then a least integer $n_0 \geqq 0$ for which EDP is solvable with $n' = n_0$; in case $n_0 = 0$, the corresponding EDP reduces to RDP. From a practical viewpoint it is of interest to find n_0: we call this the problem of *minimal* state space extension, or of minimal solution of EDP.

The general problem of minimal extension includes the general solvability problem for RDP, and is unsolved. However, suppose the additional constraint is imposed, that

$$P\mathscr{S}_i = \mathscr{R}_i^M, \qquad i \in J, \tag{2.1}$$

where the $\mathscr{R}_i^M$ are the maximal admissible c.s. in $\mathscr{C}$. In this case it will be shown how to compute the minimal n', say n_M. In general, $n_M > n_0$, because (2.1) rules out extension of any $\mathscr{R}_i \in \mathscr{C}$ which is properly contained in $\mathscr{R}_i^M$, but which still may be large enough to satisfy (1.5). However, if $d(\mathscr{B}) = k$, it will be shown in § 3 that (2.1) holds for any solution of EDP, hence $n_M = n_0$, and so this case will be solved completely.

It is convenient for later purposes to adjoin to (2.1) the additional constraint

$$P\mathscr{S}^* \subset \mathscr{V}, \tag{2.2}$$

where $\mathscr{V}$ is a subspace such that

$$\mathscr{V} \in \mathscr{J}, \qquad \mathscr{V} \subset (\mathscr{R}^M)^*. \tag{2.3}$$

In (2.3), $(\mathscr{R}^M)^*$ is the $*$ space (see Notation) of the family $\mathscr{R}_i^M$, $i \in J$. Relations of form (2.2) arise in the synthesis of pole distributions (§ 4). With $\mathscr{V}$ arbitrary, define

$$\mathscr{R}_0^M(\mathscr{V}) \equiv \bigcap_i \sum_{j \neq i} (\mathscr{R}_j^M \cap \mathscr{V}). \tag{2.4}$$

In the remainder of this section we write $\mathscr{R}_i \equiv \mathscr{R}_i^M$, $i \in \{0\} \cup J$.

THEOREM 2.1. *For the RDP of* (1.3)–(1.5) *let* $\mathscr{R}_i, i \in J$, *be the maximal admissible c.s. in* $\mathscr{C}$, *and assume* (1.6) *is true. If* $\mathscr{V}$ *satisfies* (2.3) *and if*

$$d(\mathscr{E}') \geqq n_M(\mathscr{V}) \equiv \Delta[(\mathscr{R}_i + \mathscr{R}_0(\mathscr{V}))/\mathscr{R}_0(\mathscr{V}), J], \tag{2.5}$$

then a solution $\mathscr{S}_i, i \in J$, *of EDP exists such that* $P\mathscr{S}_i = \mathscr{R}_i$, $i \in J$, *and* $\mathscr{S}^* \subset \mathscr{R}_0(\mathscr{V})$.

Conversely, if EDP has a solution $\mathscr{S}_i, i \in J$, *such that* $P\mathscr{S}_i = \mathscr{R}_i$, $i \in J$, *then*

$$P\mathscr{S}^* \in \mathscr{J}, \qquad P\mathscr{S}^* \subset \mathscr{R}^*. \tag{2.6}$$

If for some $\mathscr{V}$, $P\mathscr{S}^* \subset \mathscr{V}$, *then* $P\mathscr{S}^* \subset \mathscr{R}_0(\mathscr{V})$ *and* (2.5) *is true. If equality holds in* (2.5), *then* $P\mathscr{S}^* = \mathscr{R}_0(\mathscr{V})$ *and* $(\mathscr{S}_i + \mathscr{S}^*) \cap \mathscr{E}' = 0$, $i \in J$.

COROLLARY. *For the RDP of* (1.3)–(1.5), *suppose* (1.6) *is true and let* $\mathscr{V}^M = \max(\mathscr{J}, \mathscr{R}^*)$. *Under the constraint* (2.1) *there exists a solution* $\{\mathscr{E}', \mathscr{S}_i, i \in J\}$ *of EDP if and only if* $d(\mathscr{E}') \geqq n_M(\mathscr{V}^M)$.

Existence of $\mathscr{S}_i$ will be proved by a refinement of the construction used in the proof of Theorem 1.1. For this we need Lemmas 2.1–2.3. Of these the first two assert general properties of extensions.

LEMMA 2.1. *Let* $\mathscr{U}_i \subset \mathscr{E}$, $i \in J$. *If* $d(\mathscr{E}') \geqq \delta \equiv \Delta[\mathscr{U}_i, J]$, *there exist maps* $M_i : \mathscr{E}^e \to \mathscr{E}'$, $i \in J$, *such that the subspaces* $\mathscr{V}_i \equiv (P = M_i)\mathscr{U}_i$, $i \in J$, *are independent.*

Proof. Write $\mathscr{W}_1 = 0$, $\mathscr{W}_i = \mathscr{U}_i \cap \sum_{j=1}^{i-1} \mathscr{U}_j$, $i = 2, \cdots, k$. Then

$$\sum_{i=1}^{k} d(\mathscr{W}_i) = \sum_{i=2}^{k} \left[d(\mathscr{U}_i) + d\left(\sum_{j=1}^{i-1} \mathscr{U}_j \right) - d\left(\sum_{j=1}^{i} \mathscr{U}_j \right) \right] = \delta;$$

hence there exist M_i such that $\mathscr{N}(M_i) \cap \mathscr{W}_i = 0$, $\{M_i\} = M_i\mathscr{W}_i$ and the $\{M_i\}$, $i \in J$, are independent. Suppose the $\mathscr{V}_i$ are not independent, and let $i \geqq 2$ be the greatest integer such that $\mathscr{V}_i \cap \mathscr{V}_i^* \neq 0$. There is $x \neq 0$ such that

$$x = (P + M_i)u_i = \sum_{j=1}^{i-1} (P + M_j)u_j,$$

where $u_j \in \mathscr{U}_j$, $1 \leqq j \leqq i$, so that

$$Pu_i = u_i = \sum_{j=1}^{i-1} Pu_j = \sum_{j=1}^{i-1} u_j$$

and $u_i \in \mathscr{W}_i$. By independence of the $\{M_j\}$, $M_i u_i = 0$, hence $u_i = 0$ and $x = 0$, a contradiction.

LEMMA 2.2. *Let* $\mathscr{V}, \mathscr{R}_i \subset \mathscr{E}$, $i \in J$, *and define*

$$\mathscr{R}_0 \equiv \bigcap_i \sum_{j \neq i} (\mathscr{R}_j \cap \mathscr{V}),$$

$$\delta \equiv \Delta[(\mathscr{R}_i + \mathscr{R}_0)/\mathscr{R}_0, J].$$

If $d(\mathscr{E}') \geq \delta$, *there exist maps* $M_i : \mathscr{E}^e \to \mathscr{E}'$, $i \in J$, *such that, if* $\mathscr{V}_i \equiv (P + M_i)\mathscr{R}_i$, $i \in J$, *then* $\mathscr{V}^* = \mathscr{R}_0$.

Proof. Write $\bar{\mathscr{R}}_i \equiv (\mathscr{R}_i + \mathscr{R}_0)/\mathscr{R}_0$, $i \in J$, and let $\bar{P}$ be the projection: $\mathscr{E} \oplus \mathscr{E}' \to (\mathscr{E}/\mathscr{R}_0) \oplus \mathscr{E}'$. By Lemma 2.1, there exist $\bar{M}_i : (\mathscr{E}/\mathscr{R}_0) \oplus \mathscr{E}' \to \mathscr{E}'$ such that $\bar{\mathscr{V}}_i \equiv (\bar{P} + \bar{M}_i)\bar{\mathscr{R}}_i$, $i \in J$, are independent subspaces of $(\mathscr{E}/\mathscr{R}_0) \oplus \mathscr{E}'$. Let $M_i = \bar{M}_i \bar{P}$; then $\mathscr{V}_i$ is well-defined. Since $\bar{\mathscr{V}}_i = (\mathscr{V}_i + \mathscr{R}_0)/\mathscr{R}_0$, it follows by independence of the $\bar{\mathscr{V}}_i$, $i \in J$, and Proposition A.1 that $\mathscr{R}_0 \supset \mathscr{V}^*$. For the reverse inclusion observe that, by (A.1) and (A.2),

$$\mathscr{R}_0 = \sum_i (\mathscr{R}_i \cap \mathscr{R}_0 \cap \sum_{j \neq i} (\mathscr{R}_j \cap \mathscr{R}_0))$$

and that $M_i \mathscr{R}_0 = 0$, $i \in J$. Then $x \in \mathscr{R}_0$ implies $x = \sum_i x_i$, with

$$x_i = \sum_{j \neq i} x_{ij}; \quad x_i \in \mathscr{R}_i \cap \mathscr{R}_0; \qquad x_{ij} \in \mathscr{R}_j \cap \mathscr{R}_0$$

and

$$x_i = (P + M_i)x_i \in \mathscr{V}_i; \qquad x_{ij} = (P + M_j)x_{ij} \in \mathscr{V}_j.$$

Thus $x_i \in \mathscr{V}_i \cap \mathscr{V}_i^*$, so $x \in \mathscr{V}^*$.

LEMMA 2.3. *Let* $\mathscr{R}_i$, $i \in J$, *satisfy the hypotheses of Theorem* 2.1 *and let* $\mathscr{V}$ *satisfy* (2.3). *If* $\mathscr{R}_0$ *is defined by* (2.4), *then* $\mathscr{R}_0 \in \mathscr{I}$.

Proof. Since $\mathscr{R}^* \subset \bigcap_{j \neq i} \mathscr{N}_j$, $i \in J$, there follows $\mathscr{V} \subset \mathscr{V}_i$, $i \in J$, where $\mathscr{V}_i \equiv \max(\mathscr{I}, \bigcap_{j \neq i} \mathscr{N}_j)$. Hence for each $i \in J$ there exists $C_i : \mathscr{U} \to \mathscr{E}$ in $\mathbf{C}(\mathscr{V}_i) \cap \mathbf{C}(\mathscr{V})$. Since $\mathscr{R}_i = \max(\mathscr{C}, \mathscr{V}_i)$, there follows $\mathscr{R}_i = \{A + BC_i | \mathscr{B} \cap \mathscr{V}_i\}$, so that $C_i \in \mathbf{C}(\mathscr{R}_i) \cap \mathbf{C}(\mathscr{V}) \subset \mathbf{C}(\mathscr{R}_i \cap \mathscr{V})$. That is, $\mathscr{R}_i \cap \mathscr{V} \in \mathscr{I}$, $i \in J$, hence $\sum_{j \neq i} (\mathscr{R}_j \cap \mathscr{V}) \in \mathscr{I}$, $i \in J$. Now apply the same argument to the pair of subspaces $\mathscr{R}_i$, $\tilde{\mathscr{V}}_i \equiv \sum_{j \neq i} (\mathscr{R}_j \cap \mathscr{V})$ to get that $\mathscr{R}_i \cap \tilde{\mathscr{V}}_i \in \mathscr{I}$, $i \in J$. Finally, use (A.1), (A.2) to obtain $\mathscr{R}_0 = \sum_i \mathscr{R}_i \cap \tilde{\mathscr{V}}_i \in \mathscr{I}$.

Proof of Theorem 2.1 (direct statement). Lemmas 2.2 and 2.3 provide $\mathscr{E}'$ and $\mathscr{V}_i \subset \mathscr{E} \oplus \mathscr{E}'$, $i \in J$, with the properties: $d(\mathscr{E}') = n_M(\mathscr{V})$, and

$$P\mathscr{V}_i = \mathscr{R}_i, \qquad i \in J, \tag{2.7a}$$

$$\mathscr{V}^* = \mathscr{R}_0, \tag{2.7b}$$

$$\mathscr{V}^* \in \mathscr{I}. \tag{2.7c}$$

Since $\mathscr{I} \subset \mathscr{I}'$, $\mathscr{V}^* \in \mathscr{I}'$. Also, by (2.7a),

$$A\mathscr{V}_i \subset A(\mathscr{R}_i + \mathscr{E}') \subset \mathscr{R}_i + \mathscr{B} \subset \mathscr{V}_i + \mathscr{E}' + \mathscr{B}, \qquad i \in J;$$

hence $\mathscr{V}_i \in \mathscr{I}'$, and

$$\mathscr{V}_i + \mathscr{V}^* \in \mathscr{I}', \qquad i \in J. \tag{2.8}$$

Because the factor spaces $(\mathscr{V}_i + \mathscr{V}^*)/\mathscr{V}^*$ are independent, there exists $C \in \bigcap_i \mathbf{C}'(\mathscr{V}_i + \mathscr{V}^*)$. Define

$$\mathscr{S}_i = \{A + (B + E')C | (\mathscr{B} + \mathscr{E}') \cap (\mathscr{V}_i + \mathscr{V}^*)\}, \qquad i \in J. \tag{2.9}$$

It will be shown that $P\mathscr{S}_i = \mathscr{R}_i$ and $\mathscr{S}^* \subset \mathscr{R}_0$. By Remark 1 after Theorem 1.1, $P\mathscr{S}_i \in \mathscr{C}$; also

$$P\mathscr{S}_i \subset P(\mathscr{V}_i + \mathscr{V}^*) = \mathscr{R}_i + \mathscr{R}_0 \subset \mathscr{R}_i + \mathscr{R}^* \subset \bigcap_{j \neq i} \mathscr{N}_j.$$

Since $\mathscr{R}_i$ is maximal, $P\mathscr{S}_i \subset \mathscr{R}_i$. For the reverse inclusion, by Proposition A.5,

$$P\mathscr{S}_i \supset P[(\mathscr{B} + \mathscr{E}') \cap (\mathscr{V}_i + \mathscr{V}^*)] = \mathscr{B} \cap (\mathscr{R}_i + \mathscr{V}^*) \supset \mathscr{B} \cap \mathscr{R}_i.$$

Since $P\mathscr{S}_i \subset \mathscr{R}_i$, there exists $C_i \in \mathbf{C}(P\mathscr{S}_i) \cap \mathbf{C}(\mathscr{R}_i)$. Thus

$$P\mathscr{S}_i = \{A + BC_i | \mathscr{B} \cap P\mathscr{S}_i\} \supset \{A + BC_i | \mathscr{B} \cap \mathscr{R}_i\} = \mathscr{R}_i;$$

and so $P\mathscr{S}_i = \mathscr{R}_i$, $i \in J$. Finally, by ((ii), Appendix),

$$\mathscr{S}^* \subset \bigcap_i \sum_{j \neq i} (\mathscr{V}_j + \mathscr{V}^*) = \mathscr{V}^* = \mathscr{R}_0.$$

The idea of this proof was to use (2.9) to manufacture "compatible" c.s. contained in the $\mathscr{V}_i + \mathscr{V}^*$. The method works because the $\mathscr{V}_i + \mathscr{V}^*$ satisfy (2.8). For this one needs (2.7c), which is guaranteed (Lemma 2.3) by maximality of the $\mathscr{R}_i$, and also $\mathscr{V}_i \in \mathscr{I}'$, which follows by $\mathscr{R}_i \in \mathscr{I}$. Maximality ensures also that $R_i \supset P\mathscr{S}_i$.

Proof of Theorem 2.1 (converse statement). Since $\mathscr{S}_i$, $i \in J$, is a solution of EDP, $\mathscr{S}^* \in \mathscr{I}'$, hence $P\mathscr{S}^* \in \mathscr{I}$. Since $P\mathscr{S}_i = \mathscr{R}_i$, $i \in J$, clearly $P\mathscr{S}^* \subset \mathscr{R}^*$, so (2.6) is true. Let $P\mathscr{S}^* \subset \mathscr{V}$. Then $\mathscr{S}^* = \sum_{j \neq i} \mathscr{S}_j \cap \mathscr{S}^*$, $i \in J$, implies $P\mathscr{S}^* \subset \sum_{j \neq i} (\mathscr{R}_j \cap \mathscr{V})$, $i \in J$, and so $P\mathscr{S}^* \subset \mathscr{R}_0(\mathscr{V})$. By Proposition A.2 (where $\mathscr{S}_0$ is defined),

$$\begin{aligned} d(\mathscr{E}') \geqq \delta_1 &\equiv \Delta[(\mathscr{S}_i + \mathscr{S}_0 + \mathscr{E}')/(\mathscr{S}_0 + \mathscr{E}'), J] \\ &= \Delta[P(\mathscr{S}_i + \mathscr{S}_0)/P\mathscr{S}_0, J] \\ &= \Delta[(\mathscr{R}_i + \mathscr{R}_0)/\mathscr{R}_0, J]. \end{aligned}$$

Finally, if $d(\mathscr{E}') = \delta_1$, Proposition A.3 implies $\mathscr{S}^* + \mathscr{E}' = \mathscr{S}_0$, hence $P\mathscr{S}^* = \mathscr{R}_0$; and also $(\mathscr{S}_i + \mathscr{S}^*) \cap \mathscr{E}' = 0$.

Proof of Corollary. Any solution of EDP subject to (2.1) satisfies (2.6), hence $P\mathscr{S}^* \subset \mathscr{V}^M$, and by (2.5), $d(\mathscr{E}') \geqq n_M(\mathscr{V}^M)$. Thus $n_M(\mathscr{V}^M)$ is the least integer for which EDP is solvable subject to (2.1).

3. Minimal extension when $d(\mathscr{B}) = k$. Assume $d(\mathscr{B}) = k$ and let $\mathscr{S}_i$, $i \in J$, be any solution of EDP. It will be shown that

$$\mathscr{R}_i \equiv P\mathscr{S}_i = \mathscr{R}_i^M, \qquad i \in J. \tag{3.1}$$

By Remark 1 after Theorem 1.1, $\mathscr{R}_i \in \mathscr{C}$ and clearly the $\mathscr{R}_i$ satisfy (1.4), (1.5). It is enough to show that

$$d(\mathscr{B} \cap \mathscr{R}_i^M) = 1, \qquad i \in J. \tag{3.2}$$

In fact, since $\mathscr{N}_i \neq \mathscr{E}$, (1.5) implies $\mathscr{R}_i \neq 0$, hence (3.2) implies $\mathscr{B} \cap \mathscr{R}_i = \mathscr{B} \cap \mathscr{R}_i^M$. Since $\mathscr{R}_i \subset \mathscr{R}_i^M$, there exists $C_i \in \mathbf{C}(\mathscr{R}_i) \cap \mathbf{C}(\mathscr{R}_i^M)$. Thus

$$\mathscr{R}_i = \{A + BC_i | \mathscr{B} \cap \mathscr{R}_i\} = \{A + BC_i | \mathscr{B} \cap \mathscr{R}_i^M\} = \mathscr{R}_i^M.$$

To verify (3.2) start from

$$d\left(\mathscr{B} \cap \sum_{i=1}^{j} \mathscr{R}_i^M\right) \leqq d\left(\mathscr{B} \cap \sum_{i=1}^{j+1} \mathscr{R}_i^M\right), \qquad 1 \leqq j \leqq k-1. \tag{3.3}$$

If (3.3) holds with equality for $j = l$, then

$$\mathscr{B} \cap \sum_{i=1}^{l} \mathscr{R}_i^M = \mathscr{B} \cap \sum_{i=1}^{l+1} \mathscr{R}_i^M. \tag{3.4}$$

Write $\mathscr{P} \equiv \sum_{i=1}^{l} \mathscr{R}_i^M$. Then (3.4) implies

$$\mathscr{B} \cap (\mathscr{P} + \mathscr{R}_{l+1}^M) = \mathscr{B} \cap \mathscr{P} + \mathscr{B} \cap \mathscr{R}_{l+1}^M$$

so that (Proposition A.4)

$$\mathscr{P} \cap (\mathscr{B} + \mathscr{R}_{l+1}^M) = \mathscr{B} \cap \mathscr{P} + \mathscr{P} \cap \mathscr{R}_{l+1}^M.$$

Then

$$A(\mathscr{P} \cap \mathscr{R}_{l+1}^M) \subset (\mathscr{B} + \mathscr{P}) \cap (\mathscr{B} + \mathscr{R}_{l+1}^M) \subset \mathscr{B} + \mathscr{P} \cap \mathscr{R}_{l+1}^M.$$

By Lemma 7.1 of [1], there exists

$$C \in \mathbf{C}(P) \cap \mathbf{C}(\mathscr{R}_{l+1}^M)$$

so that

$$\mathscr{R}_{l+1}^M = \{A + BC | \mathscr{B} \cap \mathscr{R}_{l+1}^M\} \subset \{A + BC | \mathscr{P}\} \subset \mathscr{P} \subset \mathscr{N}_{l+1}$$

in contradiction to (1.5). Therefore (3.3) holds with inequality at each j. Since

$$d\left(\mathscr{B} \cap \sum_{i=1}^{k} \mathscr{R}_i^M\right) \leqq d(\mathscr{B}) = k$$

and $d(\mathscr{B} \cap \mathscr{R}_i^M) \geqq 1$, $i \in J$, the result (3.2) follows. Combining (3.1) with the corollary to Theorem 2.1, we obtain the following theorem.

THEOREM 3.1. *Let $d(\mathscr{B}) = k$. For the RDP of* (1.3)–(1.5) *suppose* (1.6) *is true, and let $\mathscr{V}^M = \max(\mathscr{I}, (\mathscr{R}^M)^*)$. There exists a solution $\{\mathscr{E}', \mathscr{S}_i, i \in J\}$ of EDP if and only if $d(\mathscr{E}') \geqq n_M(\mathscr{V}^M)$, where n_M is given by* (2.5).

4. State space extension and pole assignment. With the minimal extension of § 2 or § 3 it may happen that some poles of the closed loop transfer matrix are necessarily fixed at unstable, or otherwise "bad," locations. It is possible to shift the bad poles by additional dynamic compensation. This aim is achieved by choosing the extension such that all the fixed eigenvalues of $A + (B + E')C$ are "good."

To identify the fixed eigenvalues we need the following lemmas.

LEMMA 4.1. *Let $\mathscr{V} \in \mathscr{I}$, write $\mathbf{C} \equiv \mathbf{C}(\mathscr{V})$ and let $\mathscr{R} = \max(\mathscr{C}, \mathscr{V})$. Write $A_c \equiv A + BC$, $C \in \mathbf{C}$, and define $\bar{A}_c : \mathscr{V}/\mathscr{R} \to \mathscr{V}/\mathscr{R}$ as follows: if $\bar{x}$ is the coset of x*

in $\mathscr{V}/\mathscr{R}$, $\bar{A}_c\bar{x} \equiv \overline{A_c x}$. *Then* $\mathscr{R}$ *and* $\bar{A}_c$ *are constant with respect to* $C \in \mathscr{C}$. *In particular, the characteristic polynomial (ch. p.) of* $A_c|\mathscr{V}$ *has the form* $\pi(\lambda)\pi_c(\lambda)$, *where* π *is the ch. p. of* $\bar{A}_c$ *and is fixed for all* $C \in \mathbf{C}$; π_c *is the ch. p. of* $A_c|\mathscr{R}$, *and the roots of* π_c *can be assigned arbitrarily by suitable choice of* $C \in \mathbf{C}$.

Proof. By [1, Theorem 4.3],

$$\mathscr{R} = \{A + BC|\mathscr{B} \cap \mathscr{V}\}, \qquad C \in \mathbf{C},$$

and $\mathbf{C} \subset \mathbf{C}(\mathscr{R})$. If $C_1, C_2 \in \mathbf{C}$, and $x \in \mathscr{V}$, then $A_{c_i}x \in \mathscr{V}$, $i = 1, 2$, and

$$(A_{c_1} - A_{c_2})x = B(C_1 - C_2)x \in \mathscr{B} \cap \mathscr{V} \subset \mathscr{R},$$

hence $\bar{A}_{c_1} = \bar{A}_{c_2}$. Assignability of the roots of π_c follows by [1, Theorem 4.2].

LEMMA 4.2. *Under the conditions of Lemma* 4.1, let $\alpha(\lambda)$ *be the minimal polynomial of* $\bar{A}_c$, *and factor* $\alpha(\lambda) = \alpha_g(\lambda)\alpha_b(\lambda)$, *where the polynomials* α_g, α_b *are coprime. Then*

$$\mathscr{V} = \mathscr{R} \oplus \mathscr{R}_g \oplus \mathscr{R}_b, \tag{4.1}$$

where

$$\mathscr{R} \oplus \mathscr{R}_g = \{x : x \in \mathscr{V}, \alpha_g(\bar{A}_c)\bar{x} = \bar{0}\}, \tag{4.2}$$

and similarly for $\mathscr{R} \oplus \mathscr{R}_b$. *The subspaces* $\mathscr{R} \oplus \mathscr{R}_g$, $\mathscr{R} \oplus \mathscr{R}_b$ *are fixed with respect to* $C \in \mathbf{C}$.

Proof. Since α_g, α_b are coprime, $\mathscr{V}/\mathscr{R} = \bar{\mathscr{R}}_g \oplus \bar{\mathscr{R}}_b$, where

$$\bar{\mathscr{R}}_g = \{\bar{x} : \bar{x} \in \mathscr{V}/\mathscr{R}, \alpha_g(\bar{A}_c)\bar{x} = \bar{0}\},$$
$$\bar{\mathscr{R}}_b = \{\bar{x} : \bar{x} \in \mathscr{V}/\mathscr{R}, \alpha_b(\bar{A}_c)x = \bar{0}\}.$$

Since $\mathscr{R}$ and $\bar{A}_c$ are constant with respect to $C \in \mathbf{C}$, the result follows.

LEMMA 4.3. *Let* $\mathscr{W} \in \mathscr{I}'$ *and let* $\mathscr{S} = \max(\mathscr{C}', \mathscr{W})$. *Write* $\mathscr{V} \equiv P\mathscr{W}$ *and* $\mathscr{R} \equiv P\mathscr{S}$. *Then*

(i) $\mathscr{V} \in \mathscr{I}$, *and* $\mathscr{R} = \max(\mathscr{C}, \mathscr{V})$.
(ii) $\mathscr{V}/\mathscr{R} \approx \mathscr{W}/\mathscr{S}$.
(iii) *The fixed eigenvalues of* $(A + (B + E')C)|\mathscr{W}$, $C \in \mathbf{C}'(\mathscr{W})$, *coincide with the fixed eigenvalues of* $(A + BC_0)|\mathscr{V}$, $C_0 \in \mathbf{C}(\mathscr{V})$.

Proof.

(i) If $A\mathscr{W} \subset \mathscr{W} + \mathscr{B} + \mathscr{E}'$, $A\mathscr{V} \subset PA\mathscr{W} \subset \mathscr{V} + \mathscr{B}$, so $\mathscr{V} \in \mathscr{I}$. Let $\mathscr{R}^M \equiv \max(\mathscr{C}, \mathscr{V})$ and define

$$\mathscr{R}^0 \equiv 0, \mathscr{R}^{\mu+1} \equiv \mathscr{V} \cap (A\mathscr{R}^\mu + \mathscr{B}), \qquad \mu = 0, 1, \cdots. \tag{4.3}$$

It will be shown that $\mathscr{T} \equiv \lim \mathscr{R}^\mu = \mathscr{R}^M$. Since $\mathscr{R}^\mu \subset \mathscr{V}$ and $A\mathscr{V} \subset \mathscr{V} + \mathscr{B}$,

$$A\mathscr{R}^\mu \subset (\mathscr{V} + \mathscr{B}) \cap (A\mathscr{R}^\mu + \mathscr{B}) = \mathscr{R}^{\mu+1} + \mathscr{B},$$

so that $A\mathscr{T} \subset \mathscr{T} + \mathscr{B}$. Since $\mathscr{R}^\mu \subset \mathscr{T} \subset \mathscr{V}$, $\mu = 0, 1, \cdots$, (4.3) implies

$$\mathscr{R}^{\mu+1} = \mathscr{T} \cap (A\mathscr{R}^\mu + \mathscr{B}), \qquad \mu = 0, 1, \cdots.$$

By [1, Theorem 4.1], $\mathscr{T} \in \mathscr{C}$ and $\mathscr{T} \subset \mathscr{V}$, hence $\mathscr{T} \subset \mathscr{R}^M$. On the other hand, $\mathscr{R}^M = \lim \hat{\mathscr{R}}^\mu$, where $\hat{\mathscr{R}}^0 = 0$ and

$$\hat{\mathscr{R}}^{\mu+1} = \mathscr{R}^M \cap (A\hat{\mathscr{R}}^\mu + \mathscr{B}), \qquad \mu = 0, 1, \cdots.$$

Since $\mathscr{R}^M \subset \mathscr{V}$, by induction on μ we have $\hat{\mathscr{R}}^\mu \subset \mathscr{R}^\mu$, hence $\mathscr{R}^M \subset \mathscr{T}$. Thus the rule (4.3) computes $\max(\mathscr{C}, \mathscr{V})$.

Applying this result to the pair $\mathscr{S}, \mathscr{W}$ we have $\mathscr{S} = \lim \mathscr{S}^\mu$, where $\mathscr{S}^0 = 0$ and

$$\mathscr{S}^{\mu+1} = \mathscr{W} \cap (A\mathscr{S}^\mu + \mathscr{B} + \mathscr{E}'), \qquad \mu = 0, 1, \cdots.$$

Thus (Proposition A.5) $P\mathscr{S}^{\mu+1} = \mathscr{V} \cap (AP\mathscr{S}^\mu + \mathscr{B})$, and comparison with (4.3) yields

$$\mathscr{R}^M = \lim \mathscr{R}^\mu = \lim P\mathscr{S}^\mu = P\mathscr{S}.$$

(ii) In general,

$$\mathscr{W}/\mathscr{S} \approx (\mathscr{W} + \mathscr{E}')/(\mathscr{S} + \mathscr{E}') \oplus (\mathscr{W} \cap \mathscr{E}')/(\mathscr{S} \cap \mathscr{E}').$$

But

$$\frac{\mathscr{W} + \mathscr{E}'}{\mathscr{S} + \mathscr{E}'} \approx \frac{(\mathscr{W} + \mathscr{E}')/\mathscr{E}'}{(\mathscr{S} + \mathscr{E}')/\mathscr{E}'} \approx P\mathscr{W}/P\mathscr{S} = \mathscr{V}/\mathscr{R}. \tag{4.4}$$

Also, for $C \in \mathbf{C}'(\mathscr{W})$,

$$\mathscr{S} = \{A + (B + E')C | (\mathscr{B} + \mathscr{E}') \cap \mathscr{W}\}$$

so that $\mathscr{W} \cap \mathscr{E}' \subset \mathscr{S}$, hence

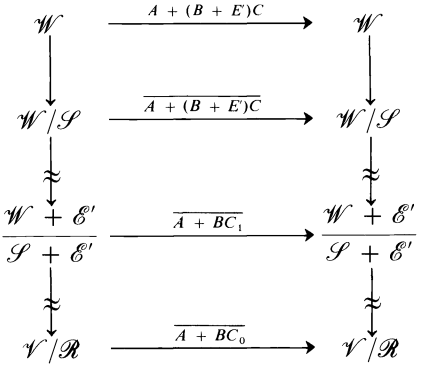
$$(\mathscr{W} \cap \mathscr{E}')/(\mathscr{S} \cap \mathscr{E}') \approx 0.$$

(iii) Let $C \in \mathbf{C}'(\mathscr{W})$. It will be shown that there exist

$$C_1 \in \mathbf{C}(\mathscr{S} + \mathscr{E}') \cap \mathbf{C}(\mathscr{V}), \tag{4.5a}$$

$$C_0 \in \mathbf{C}(\mathscr{V}) \tag{4.5b}$$

such that the diagram commutes (see Diag. 1). By the isomorphisms shown, the result will then follow from Lemma 4.1. In the diagram a bar denotes the



DIAG. 1

induced map in the indicated factor space. Turning to the proof, since

$$(A + (B + E')C)\mathscr{S} \subset \mathscr{S},$$

the top square commutes (by definition of bar). Recall that $\mathscr{E}' \cap \mathscr{W} \subset \mathscr{S}$ and write

$$\begin{aligned} \mathscr{W} &= ((\mathscr{S} + \mathscr{E}') \cap \mathscr{W}) \oplus \mathscr{Z}, \\ &= \mathscr{S} \oplus \mathscr{Z}. \end{aligned} \tag{4.6}$$

Since $A\mathscr{E}' = 0$ and $\mathscr{S} \in \mathscr{I}'$, there follows $\mathscr{S} + \mathscr{E}' \in \mathscr{I}$, and there exists $C_1 \in \mathbf{C}(\mathscr{S} + \mathscr{E}')$ such that $(C_1 - C)\mathscr{Z} = 0$. Then $A + BC_1$ is defined, and

$$[A + (B + E')C - (A + BC_1)]\mathscr{W} \subset \mathscr{S} + \mathscr{E}',$$

so the middle square commutes. Clearly $(A + BC_1)(\mathscr{W} + \mathscr{E}') \subset \mathscr{W} + \mathscr{E}'$; since $\mathscr{V} \subset \mathscr{W} + \mathscr{E}'$, $(A + BC_1)\mathscr{V} = P(A + BC_1)\mathscr{V} \subset P(\mathscr{W} + \mathscr{E}') = \mathscr{V}$, i.e., $C_1 \in \mathbf{C}(\mathscr{V})$ and (4.5a) is true. By (4.4) and (4.6), $P\mathscr{Z} \approx \mathscr{Z}$, i.e., $\mathscr{V} = \mathscr{R} \oplus P\mathscr{Z}$, and C_0 exists such that

$$(C_0 - C_1)\mathscr{R} = 0, \qquad (C_0P - C_1)\mathscr{Z} = 0. \tag{4.7}$$

Then

$$[(A + BC_0)P - P(A + BC_1)]\mathscr{Z} = 0. \tag{4.8}$$

Also, if $x \in \mathscr{S}$, then $Px \in \mathscr{R}$ and

$$\begin{aligned} (A + BC_0)Px &= (A + BC_1)Px \\ &= (A + BC_1)(x + e') \qquad (\text{for some } e' \in \mathscr{E}') \\ &= (A + BC_1)x + e'' \quad (\text{for some } e'' \in \mathscr{S} + \mathscr{E}') \end{aligned}$$

so that

$$\begin{aligned} (A + BC_0)Px &= P(A + BC_0)Px \\ &= P(A + BC_1)x + Pe'' \end{aligned} \tag{4.9}$$

and $Pe'' \in \mathscr{R}$. Then (4.6), (4.8) and (4.9) imply

$$[(A + BC_0)P - P(A + BC_1)]\mathscr{W} \subset \mathscr{R},$$

so that the bottom square of the diagram commutes. Finally it is clear from (4.5a) and (4.7) that (4.5b) is true.

We now state a procedure for minimal extension of c.s. $\mathscr{R}_i^M$ to achieve both decoupling and an assigned distribution of eigenvalues of $A + (B + E')C$. We write $\mathscr{R}_i \equiv \mathscr{R}_i^M$ and assume the hypotheses of Theorem 2.1.

Extension procedure (*EXT*). Here A will denote the original map in $\mathscr{E}$, not its extension, and similarly for B, C. Under the conditions of Theorem 2.1, let $\mathscr{V}^M \equiv \max(\mathscr{I}, \mathscr{R}^*)$, $\mathscr{R}^M \equiv \max(\mathscr{C}, \mathscr{V}^M)$. For $C \in \mathbf{C} \equiv \mathbf{C}(\mathscr{V}^M)$, write $A_c \equiv A + BC$, and let $\alpha(\lambda)$ be the minimal polynomial (mod $\mathscr{R}^M$) of $A_c|\mathscr{V}^M$. Factor $\alpha(\lambda) = \alpha_g(\lambda)\alpha_b(\lambda)$, where the roots of $\alpha_g(\alpha_b)$ are good (bad). For arbitrary $C \in \mathbf{C}$ determine

$$\mathscr{R}^M \oplus \mathscr{R}_g \equiv \{x : x \in \mathscr{V}^M, \alpha_g(A_c)x \in \mathscr{R}^M\}. \tag{4.10}$$

In (2.4) substitute $\mathscr{V} = \mathscr{R}^M \oplus \mathscr{R}_g$, compute $\mathscr{R}_0 \equiv \mathscr{R}_0(\mathscr{V})$, and construct a minimal solution of EDP as in the proof (direct half) of Theorem 2.1.

With EXT completed, a solution of EDP is now in hand: symbols A etc. will again denote the extended maps, defined by (1.8). Write $\mathbf{C}' \equiv \bigcap_i \mathbf{C}'(\mathscr{S}_i)$, $A_c \equiv A + (B + E')C$.

THEOREM 4.1. *Any solution $\mathscr{E}'$, $\mathscr{S}_i$, $i \in J$, of EDP determined by EXT has the following properties*:

(i) $\mathscr{R}^M \subset \mathscr{S}^* \subset \mathscr{R}^M \oplus \mathscr{R}_g$.

(ii) *If $C \in \mathbf{C}'$, the ch. p. $\pi_c^*(\lambda)$ of $A_c|\mathscr{S}^*$ can be factored as*

$$\pi_c^*(\lambda) = \pi_g(\lambda)\pi_c^M(\lambda). \tag{4.11}$$

Here the roots of π_g are fixed for $C \in \mathbf{C}'$ and each root is a root of α_g; the roots of π_c^M can be assigned as any symmetric set of $d(\mathscr{R}^M)$ complex numbers by suitable choice of $C|\mathscr{R}^M$, $C \in \mathbf{C}'$.

(iii) *Write $\mathscr{S} = \mathscr{S}_1 + \cdots + \mathscr{S}_k$. The ch. p. $\pi_c(\lambda)$ of $A_c|\mathscr{S}$ can be factored as*

$$\pi_c(\lambda) = \pi_{1c}(\lambda) \cdots \pi_{kc}(\lambda)\pi_c^*(\lambda), \tag{4.12}$$

where

$$\begin{aligned} d_i \equiv \deg \pi_{ic} &= d((\mathscr{R}_i + \mathscr{R}_0)/\mathscr{R}_0), \qquad i \in J, \\ \deg \pi_c^* &= d(\mathscr{R}_0). \end{aligned} \tag{4.13}$$

The roots of π_{ic}, $i \in J$, can be assigned as any symmetric set of d_i complex numbers by suitable choice of $C \in \mathbf{C}'$, independent of $C|\mathscr{R}^M$.

Proof.

(i) By Theorem 2.1, EXT determines the $\mathscr{S}_i$ such that

$$\mathscr{S}^* = \mathscr{R}_0 = \bigcap_i \sum_{j \neq i} (\mathscr{R}_j \cap (\mathscr{R}^M \oplus \mathscr{R}_g)).$$

Since $\mathscr{R}^M \subset \mathscr{R}^* \subset \bigcap_i \mathscr{N}_i$, maximality of the $\mathscr{R}_j$ implies $\mathscr{R}_j \supset \mathscr{R}^M$, $j \in J$, so that $\mathscr{R}^M \subset \mathscr{S}^*$ and ((i), Theorem 4.1) follows.

(ii) If $C \in \mathbf{C}'$, $A_c\mathscr{S}^* \subset \mathscr{S}^* \subset \mathscr{V}^M \subset \mathscr{E}$, so that $A_c|\mathscr{S}^* = (A + BC)|\mathscr{S}^*$ and $C|\mathscr{S}^*$ has an extension $C_1 : \mathscr{E} \to \mathscr{U}$ such that $C_1 \in \mathbf{C}$ and $A_{c_1}|\mathscr{S}^* = A_c|\mathscr{S}^*$. By ((i), Theorem 4.1) and Lemma 4.1 (with $\mathscr{R} = \mathscr{R}^M$; $\mathscr{V} = \mathscr{V}^M$), the ch. p. of $A_{c_1}|\mathscr{S}^*$ factors as in (4.11), and the roots of $\pi_{c_1}^M$ are freely assignable by suitable choice of $C_1|\mathscr{R}^M$, $C_1 \in \mathbf{C}$, hence by suitable choice of $C|\mathscr{S}^*$, $C \in \mathbf{C}'$.

(iii) The expression (4.12) and assignability of the roots of π_{ic} follow by Theorem 1.2 applied to the $\mathscr{S}_i$; (4.13) follows by the fact (Theorem 2.1) that $(\mathscr{S}_i + \mathscr{S}^*) \cap \mathscr{E}' = 0$ and $\mathscr{S}^* = \mathscr{R}_0$, hence

$$(\mathscr{S}_i + \mathscr{S}^*)/\mathscr{S}^* \approx \frac{(\mathscr{S}_i + \mathscr{S}^* + \mathscr{E}')/\mathscr{E}'}{(\mathscr{S}^* + \mathscr{E}')/\mathscr{E}'} \approx (\mathscr{R}_i + \mathscr{R}_0)/\mathscr{R}_0.$$

Now suppose $\{\mathscr{E}', \mathscr{S}_i, i \in J\}$ is any solution of EDP, not necessarily determined by EXT. Then $\mathscr{S}^* \in \mathscr{J}'$ and $P\mathscr{S}^* \in \mathscr{J}$. By Lemma 4.3, the fixed eigenvalues of

$$A_c^* \equiv (A + (B + E')C)|\mathscr{S}^*, \qquad C \in \mathbf{C}'(\mathscr{S}^*),$$

coincide with the fixed eigenvalues of

$$(A + BC_0)|P\mathscr{S}^*, \qquad C_0 \in \mathbf{C}(P\mathscr{S}^*).$$

As shown in the proof of Theorem 1.1, $P\mathscr{S}_i \subset \mathscr{R}_i (\equiv \mathscr{R}_i^M)$, hence $P\mathscr{S}^* \subset \mathscr{R}^*$, and by maximality of $\mathscr{V}^M$, $P\mathscr{S}^* \subset \mathscr{V}^M$. Therefore, $C_0|P\mathscr{S}^*$ has an extension $C_0^M \in \mathbf{C}(\mathscr{V}^M)$. By Lemma 4.2,

$$\mathscr{V}^M = \mathscr{R}^M \oplus \mathscr{R}_g \oplus \mathscr{R}_b$$

with $\mathscr{R}^M \oplus \mathscr{R}_g$ given by (4.10). Since the fixed eigenvalues of A_c^* coincide with those of $(A + BC_0^M)|P\mathscr{S}^*$, it follows, if the fixed eigenvalues of A_c^* are all good, that

$$P\mathscr{S}^* \subset \mathscr{V} \equiv \mathscr{R}^M \oplus \mathscr{R}_g. \tag{4.14}$$

Since the extension constructed by EXT is minimal with respect to the properties (2.1) and (4.14), we have proved the following.

THEOREM 4.2. *The construction EXT yields a minimal solution of EDP, subject to* (2.1) *and the requirement that the fixed eigenvalues of*

$$(A + (B + E')C)|\mathscr{S}, \qquad C \in \mathbf{C}',$$

all be good.

Remark. Assuming as in [1] that $\{A|\mathscr{B}\} = \mathscr{E}$, we have that $\{A|\mathscr{B} \oplus \mathscr{E}'\} = \mathscr{E} \oplus \mathscr{E}'$. By the technique used in proving Theorem 1.2, it is straightforward to show that $(\mathscr{E} \oplus \mathscr{E}')/\mathscr{S}$ can be regarded as a c.s. (mod $\mathscr{S}$) for $(A, B + E')$, hence that the *only* fixed eigenvalues of A_c are those of A_c^*.

5. Example. Let $n = d(\mathscr{E}) = 5$ and let e_i, $1 \leqq i \leqq 5$, be the ith unit column vector, with 1 in the ith row and 0 elsewhere. Let

$$A = [e_4, e_1, e_3, e_3, e_4], \qquad B = [e_2, e_1 + e_5],$$

$H_1 = \text{row } e_1$, $H_2 = \text{row } e_5$. Writing $\{\cdot\}$ for the span of the vectors bracketed, we have

$$\mathscr{N}_1 = \{e_2, e_3, e_4, e_5\}, \qquad \mathscr{N}_2 = \{e_1, e_2, e_3, e_4\}.$$

It is easily checked that

$$\mathscr{R}_1^M = \mathscr{N}_2, \quad \mathscr{R}_2^M = \mathscr{N}_1, \quad \mathscr{B} \cap \mathscr{R}_1^M = \mathscr{B} \cap \mathscr{R}_2^M = \{e_2\}.$$

By Theorem 7.1 of [1], decoupling by state feedback is not possible. However, since (1.6) is satisfied, namely

$$\mathscr{R}_1^M + \mathscr{N}_1 = \mathscr{R}_2^M + \mathscr{N}_2 = \mathscr{E},$$

Theorem 1.1 asserts that decoupling is possible by use of dynamic compensation.

In this example, $d(\mathscr{B}) = 2 = k$, and according to § 3 any solution $\mathscr{S}_i$, $i = 1, 2$, of EDP must satisfy

$$P\mathscr{S}_i = \mathscr{R}_i^M, \qquad i = 1, 2. \tag{5.1}$$

By Theorem 3.1, a minimal extension has $d(\mathscr{E}') = n_M(\mathscr{V}^M)$ given by (2.5), where

$$\mathscr{V}^M = \max\,(\mathscr{I}, (\mathscr{R}^M)^*).$$

In this example,

$$(\mathscr{R}^M)^* = \mathscr{R}_1^M \cap \mathscr{R}_2^M = \{e_2, e_3, e_4\}$$

and one easily computes $\mathscr{V}^M = \{e_3, e_4\}$. By (2.4),

$$\mathscr{R}_0(\mathscr{V}^M) = \mathscr{R}_1^M \cap \mathscr{R}_2^M \cap \mathscr{V}^M = \mathscr{V}^M.$$

Then (2.5) gives $n_M(\mathscr{V}^M) = 1$, so that just one integrator is needed to achieve decoupling by dynamic compensation.

To determine the spectrum of $A + (B + E')C$ we follow the procedure EXT of § 4, and start by finding $\mathscr{R}^M = \max(\mathscr{C}, \mathscr{V}^M)$. Since $\mathscr{B} \cap \mathscr{V}^M = 0$, we have $\mathscr{R}^M = 0$. Since $A\mathscr{V}^M \subset \mathscr{V}^M$, we can take $A_c|\mathscr{V}^M = A|\mathscr{V}^M$; in our coordinate system

$$A|\mathscr{V}^M = \begin{bmatrix} A_{33} & A_{34} \\ A_{43} & A_{44} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix},$$

so that $\alpha(\lambda) = \lambda(\lambda - 1)$. If unstable eigenvalues are considered "bad," we have $\mathscr{R}_g = 0$ and $\mathscr{R}_b = \mathscr{V}^M$. Both the bad eigenvalues are fixed in the minimal extension determined first. To find the minimal extension subject to the constraint that all fixed eigenvalues be good, we set $\mathscr{V} = \mathscr{R}^M \oplus \mathscr{R}_g = 0$. By (2.4), $\mathscr{R}_0(\mathscr{V}) = 0$, and (2.5) gives

$$d(\mathscr{E}') = n_M(\mathscr{V}) = 3.$$

Exactly three compensating integrators are needed to achieve decoupling together with stability of the (extended) closed loop system matrix.

The reader may wish to investigate the possibilities with two compensating integrators.

6. Decoupling and open loop control. In previous sections and in [1], the apparently stringent restriction was imposed that feedback and dynamic compensation be linear. In particular, the definition of controllability subspace [1] was tied to a specific linear feedback structure. We now show that, with regard to decoupling, nothing is gained by considering more general types of control. To this end we show that *maximal* c.s. can be defined in an open loop sense without any assumptions on controller structure. Consider

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), && t \in T, \\ x(0) &= 0 \end{aligned} \tag{6.1}$$

on the time interval $T = [0, 1]$, and let $\mathscr{N} \subset \mathscr{E}$. Let $\mathbf{U}$ denote the class of m-vector-valued functions $u(\cdot)$, defined and continuous on T. Denote by $\phi: T \times \mathbf{U} \to \mathscr{E}$ the solution of (6.1), i.e.,

$$\phi(t, u) = \int_0^t e^{(t-s)A} Bu(s)\, ds, \qquad t \in T, \quad u \in \mathbf{U}.$$

THEOREM 6.1. *Let $\mathscr{X}$ be the set of states $x \in \mathscr{N}$ such that, for some $u \in \mathbf{U}$,*

$$\phi(t, u) \in \mathscr{N}, \quad t \in T; \qquad \phi(1, u) = x.$$

Then $\mathscr{X} = \mathscr{R}^M \equiv \max(\mathscr{C}, \mathscr{N})$.

Thus $\mathscr{R}^M$ is characterized as the largest set of states in $\mathscr{N}$ which can be reached from the zero state, by any control whatever, without leaving $\mathscr{N}$.

Proof. Let

$$\mathscr{R}^M = \{A + BC|\{BK\}\},$$

and write $\hat{A} \equiv A + BC, \hat{B} = BK$. We claim that

$$\mathscr{R}^M = \{R\}, \tag{6.2}$$

where

$$R = \int_0^1 e^{(1-t)\hat{A}}\hat{B}\hat{B}' e^{(1-t)\hat{A}'} dt$$

(here and below, maps are represented by matrices, and a prime denote transpose). In fact, $z \in \mathscr{N}(R)$ implies $z' e^{(1-t)\hat{A}}\hat{B} = 0$, $t \in T$, i.e., $z'\hat{A}^{j-1}\hat{B} = 0, j = 1, \cdots, n$, so $z \in (\mathscr{R}^M)^\perp$. Thus $\mathscr{N}(R) \subset (\mathscr{R}^M)^\perp$, so that $\mathscr{R}^M \subset \{R\}$, and the reverse inclusion is obvious.

To show $\mathscr{R}^M \subset \mathscr{X}$, let $x \in \mathscr{R}^M$, and note from (6.2) that $x = Rw$ for some $w \in \mathscr{E}$. Set

$$v(t) = \hat{B}' e^{(1-t)\hat{A}'}w, \qquad t \in T.$$

Then the equation

$$\dot{x}(t) = \hat{A}x(t) + \hat{B}v(t), \qquad t \in T,$$
$$x(0) = 0$$

implies $x(T) \subset \mathscr{R}^M \subset \mathscr{N}$ and $x(1) = x$, where $x(T) \equiv \{x(t): t \in T\}$. Put

$$u(t) = Kv(t) - Cx(t), \qquad t \in T.$$

Then $u \in \mathbf{U}$; $\phi(t, u) \in \mathscr{N}$, $t \in T$; $\phi(1, u) = x$; and so $x \in \mathscr{X}$.

To show $\mathscr{X} \subset \mathscr{R}^M$, let $\mathscr{V} = \max(\mathscr{I}, \mathscr{N})$. By [1, Theorem 3.1], $\mathscr{V} = \mathscr{V}^n$, where $\mathscr{V}^0 = \mathscr{N}$ and

$$\mathscr{V}^{\mu+1} = \mathscr{V}^\mu \cap A^{-1}(\mathscr{V}^\mu + \mathscr{B}), \qquad \mu = 0, 1, \cdots, \quad A^{-1}\mathscr{V} \equiv \{x: Ax \in \mathscr{V}\}.$$

If $x \in \mathscr{X}$, then for some $u \in \mathbf{U}$, (6.1) yields

$$x(T) \subset \mathscr{N}, \qquad x(1) = x.$$

Thus $x(T) \subset \mathscr{V}^0$. If $x(T) \subset \mathscr{V}^\mu$, then $\dot{x}(T) \subset \mathscr{V}^\mu$, so $Ax(T) = (\dot{x} - Bu)(T) \subset \mathscr{V}^\mu + \mathscr{B}$; hence

$$x(T) \subset \mathscr{V}^\mu \cap A^{-1}(\mathscr{V}^\mu + \mathscr{B}) = \mathscr{V}^{\mu+1},$$

and by induction $x(T) \subset \mathscr{V}$. Let $C \in \mathbf{C}(\mathscr{V})$. Then

$$\dot{x}(t) = (A + BC)x(t) + Bv(t), \qquad t \in T,$$

where $v(t) = u(t) - Cx(t)$. Thus

$$Bv(T) = (\dot{x} - (A + BC)x)(T) \subset \mathscr{V}$$

so that $\{Bv(t)\} \subset \mathscr{B} \cap \mathscr{V}$, $t \in T$. Hence, for $t \in T$,

$$x(t) = \int_0^t \exp[(t-s)(A+BC)]Bv(s)\,ds \in \{A+BC|\mathscr{B} \cap \mathscr{V}\} = \mathscr{R}^M.$$

We now pose an *open loop decoupling problem* (ODP) as follows. *Given* (6.1), *and* (1.2) *defined for* $t \in T$, *together with arbitrary vectors* $y_i \in \mathscr{H}_i$, $i \in J$, *find controls* $u_i \in \mathbf{U}$, $i \in J$, *such that*

$$H_i\phi(1, u_i) = y_i, \qquad i \in J, \tag{6.3}$$

$$H_j\phi(T, u_i) = 0, \qquad i, j \in J;\quad j \neq i. \tag{6.4}$$

Under these conditions each u_i affects only the output $y_i(\cdot)$, and $y_i(1) = y_i$.

THEOREM 6.2. *Write* $\mathscr{N}_i \equiv \mathscr{N}(H_i)$, $i \in J$. *ODP is solvable for arbitrary* $y_i \in \mathscr{H}_i$, $i \in J$, *if and only if*

$$\mathscr{R}_i^M + \mathscr{N}_i = \mathscr{E}, \qquad i \in J, \tag{6.5}$$

where

$$\mathscr{R}_i^M = \max\left(\mathscr{C}, \bigcap_{j \neq i} \mathscr{N}_j\right), \qquad i \in J. \tag{6.6}$$

Proof. If (6.5) is true, then $H_i\mathscr{R}_i^M = \mathscr{H}_i$, and there is $x_i \in \mathscr{R}_i^M$ with $H_i x_i = y_i$. By Theorem 6.1, there is $u_i \in \mathbf{U}$ such that

$$\phi(1, u_i) = x_i, \qquad \phi(T, u_i) \subset \bigcap_{j \neq i} \mathscr{N}_j,$$

i.e.,

$$H_i\phi(1, u_i) = y_i, \qquad H_j\phi(T, u_i) = 0, \qquad j \neq i.$$

Conversely, if (6.5) fails, then for some $i \in J$ there is $y \in \mathscr{H}_i$ such that $y \notin H_i\mathscr{R}_i^M$. Therefore any control $u \in \mathbf{U}$, such that $H_i\phi(1, u) = y$, has the property $\phi(1, u) \notin \mathscr{R}_i^M$. By Theorem 6.1,

$$\phi(t, u) \notin \bigcap_{j \neq i} \mathscr{N}_j$$

for some $t \in T$; i.e., for this t, $H_j\phi(t, u) \neq 0$ for some $j \in J$, $j \neq i$, and (6.4) fails.

Comparing Theorem 6.2 with Theorem 1.1 we have the following.

COROLLARY. *ODP is solvable if and only if EDP is solvable, namely, if and only if* (6.5) *is true.*

In the definition of ODP the choice $\mathbf{U}$ for the class of admissible controls, and the choice in (6.3) of common endpoint $t = 1$, are obviously not crucial. In fact, we have shown implicitly that a wide class of dynamic decoupling problems is equivalent to the EDP of § 1.

Concluding remark. Taken with its predecessor [1], the present article provides effective machinery for the formulation and solution of the decoupling problem. The results prescribe the synthesis of dynamic compensation by which decoupling can be realized, and clarify the conditions under which such compensation exists. Nevertheless, further aspects of the problem remain for

investigation. These include computer implementation, sensitivity analysis, and perhaps most important, a deeper account of algebraic structure.

Appendix. We collect here some auxiliary results; verifications, when straightforward, are omitted.

(i) Let $\mathscr{V}_i$, $i \in J$, be arbitrary subspaces. Let

$$\mathscr{V}_i^* \equiv \sum_{j \neq i} \mathscr{V}_j, \qquad \mathscr{V}^* \equiv \bigcap_i \mathscr{V}_i^*.$$

Then

$$\text{(A.1)} \qquad \mathscr{V}^* = \sum_i \mathscr{V}_i \cap \mathscr{V}^* = \sum_i \mathscr{V}_i \cap \mathscr{V}_i^* = \sum_{i \neq j} \mathscr{V}_i \cap \mathscr{V}_i^*, \qquad j \in J.$$

(ii) If $\mathscr{U}_i \equiv \mathscr{V}_i + \mathscr{V}^*$, $i \in J$, then $\mathscr{U}^* = \mathscr{V}^*$.

(iii) If $\mathscr{X} \equiv \bigcap_i \sum_{j \neq i} \mathscr{V}_j \cap \mathscr{Y}$ for some $\mathscr{Y}$, then

$$\text{(A.2)} \qquad \mathscr{X} = \bigcap_i \sum_{j \neq i} \mathscr{V}_j \cap \mathscr{X}.$$

(iv) By definition, the $\mathscr{V}_i$, $i \in J$, are mutually *independent* if and only if $\mathscr{V}^* = 0$, i.e., $\mathscr{V}_i \cap \mathscr{V}_i^* = 0$, $i \in J$. More generally:

PROPOSITION A.1. *$\mathscr{V}^*$ is the smallest subspace $\mathscr{V}_0$ such that the factor spaces $(\mathscr{V}_i + \mathscr{V}_0)/\mathscr{V}_0$ are independent subspaces of $\mathscr{E}/\mathscr{V}_0$.*

Proof. Independence of the factor spaces is equivalent to

$$\text{(A.3)} \qquad \mathscr{V}_0 = \sum_i [(\mathscr{V}_i + \mathscr{V}_0) \cap (\mathscr{V}_i^* + \mathscr{V}_0)].$$

From (A.3), $\mathscr{V}^* = \lim \mathscr{V}^\mu$, $\mu = 0, 1, \cdots$, where

$$\text{(A.4)} \qquad \mathscr{V}^0 = 0, \qquad \mathscr{V}^{\mu+1} = \sum_i [(\mathscr{V}_i + \mathscr{V}^\mu) \cap (\mathscr{V}_i^* + \mathscr{V}^\mu)].$$

By (ii), $\mathscr{V}^*$ satisfies (A.3), and (A.4) implies that any solution $\mathscr{V}_0$ of (A.3) contains $\mathscr{V}^*$.

(v) By Proposition A.1,

$$\begin{aligned} \sum_i d(\mathscr{V}_i/(\mathscr{V}_i \cap \mathscr{V}^*)) &= \sum_i d((\mathscr{V}_i + \mathscr{V}^*)/\mathscr{V}^*), \\ &= d(\sum_i (\mathscr{V}_i + \mathscr{V}^*)/\mathscr{V}^*), \\ &= d((\sum_i \mathscr{V}_i)/\mathscr{V}^*) \end{aligned}$$

so that

$$\text{(A.5)} \qquad \Delta[\mathscr{V}_i, J] \equiv \sum_i d(\mathscr{V}_i) - d(\sum_i \mathscr{V}_i) = \sum_i d(\mathscr{V}_i \cap \mathscr{V}^*) - d(\mathscr{V}^*).$$

(vi) If $\mathscr{U}$, $\mathscr{V}$, $\mathscr{W}$ are arbitrary spaces,

$$\text{(A.6)} \qquad \frac{(\mathscr{U} + \mathscr{W}) \cap (\mathscr{V} + \mathscr{W})}{\mathscr{U} \cap \mathscr{V} + \mathscr{W}} \approx \frac{(\mathscr{U} + \mathscr{V}) \cap \mathscr{W}}{\mathscr{U} \cap \mathscr{W} + \mathscr{V} \cap \mathscr{W}}.$$

PROPOSITION A.2. *Let* $\mathscr{S}_i, i \in J, \mathscr{V}, \mathscr{E}'$ *be such that* $\mathscr{V} \cap \mathscr{E}' = 0, \mathscr{S}^* \subset \mathscr{V} \oplus \mathscr{E}'$. *Define* $\mathscr{S} \equiv \sum_i \mathscr{S}_i$ *and*

$$\mathscr{S}_0 \equiv \bigcap_i \sum_{j \neq i} (\mathscr{S}_j + \mathscr{E}') \cap (\mathscr{V} \oplus \mathscr{E}').$$

Then

$$d(\mathscr{E}') = \delta_1 + \delta_2 + \rho, \tag{A.7}$$

where

$$\delta_1 \equiv \Delta[(\mathscr{S}_i + \mathscr{S}_0 + \mathscr{E}')/(\mathscr{S}_0 + \mathscr{E}'), J], \tag{A.8}$$

$$\delta_2 \equiv \Delta[((\mathscr{S}_i + \mathscr{E}') \cap (\mathscr{S}_0 + \mathscr{E}') + \mathscr{S}^*)/(\mathscr{S}^* + \mathscr{E}'), J], \tag{A.9}$$

$$\begin{aligned} \rho \equiv & \sum_i d[((\mathscr{S}_i + \mathscr{S}^*) \cap \mathscr{E}')/(\mathscr{S}_i \cap \mathscr{E}' + \mathscr{S}^* \cap \mathscr{E}')] \\ & + \sum_i d[(\mathscr{S}_i \cap \mathscr{E}')/(\mathscr{S}_i \cap \mathscr{S}^* \cap \mathscr{E}')] \\ & + d(\mathscr{S}^* \cap \mathscr{E}') + d(\mathscr{E}'/(\mathscr{S} \cap \mathscr{E}')). \end{aligned} \tag{A.10}$$

Proof. The proof is a direct computation, starting from the easy identity

$$d(\mathscr{E}') = d(\mathscr{S}) - d((\mathscr{S} + \mathscr{E}')/\mathscr{E}') + d(\mathscr{E}'/(\mathscr{S} \cap \mathscr{E}'))$$

and using (ii) and (A.1)–(A.6); from (A.2) note especially

$$\mathscr{S}_0 = \bigcap_i \sum_{j \neq i} [(\mathscr{S}_j + \mathscr{E}') \cap (\mathscr{S}_0 + \mathscr{E}')]. \tag{A.11}$$

PROPOSITION A.3. *If in* (A.7), $d(\mathscr{E}') = \delta_1$, *then* $\mathscr{S}^* + \mathscr{E}' = \mathscr{S}_0, (\mathscr{S}_i + \mathscr{S}^*) \cap \mathscr{E}' = 0, i \in J$, *and* $\mathscr{E}' \subset \mathscr{S}$.

Proof. $\rho = 0$ implies $\mathscr{S} \cap \mathscr{E}' = \mathscr{E}'$, i.e., $\mathscr{E}' \subset \mathscr{S}$; also $\mathscr{S}^* \cap \mathscr{E}' = 0$, hence $\mathscr{S}_i \cap \mathscr{E}' = 0, i \in J$; so that, from the first summation in (A.10), $(\mathscr{S}_i + \mathscr{S}^*) \cap \mathscr{E}' = 0, i \in J$. Also, $\delta_2 = 0$ implies that the bracketed factor spaces in (A.9) are independent; by Proposition A.1 and (A.11),

$$\mathscr{S}^* + \mathscr{E}' \supset \bigcap_i \sum_{j \neq i} ((\mathscr{S}_j + \mathscr{E}') \cap (\mathscr{S}_0 + \mathscr{E}')) = \mathscr{S}_0. \tag{A.12}$$

By (A.2) and the definitions of $\mathscr{S}^*, \mathscr{S}_0$,

$$\mathscr{S}_0 \supset \bigcap_i \sum_{j \neq i} \mathscr{S}_j \cap \mathscr{S}^* = \mathscr{S}^*,$$

hence the reverse inclusion holds in (A.12), so $\mathscr{S}^* + \mathscr{E}' = \mathscr{S}_0$.

PROPOSITION A.4. *For arbitrary* $\mathscr{U}, \mathscr{V}, \mathscr{W}$, *if*

$$\mathscr{U} \cap (\mathscr{V} + \mathscr{W}) = \mathscr{U} \cap \mathscr{V} + \mathscr{U} \cap \mathscr{W},$$

then

$$\mathscr{V} \cap (\mathscr{U} + \mathscr{W}) = \mathscr{U} \cap \mathscr{V} + \mathscr{V} \cap \mathscr{W}.$$

PROPOSITION A.5. *For arbitrary* $\mathscr{U}, \mathscr{V}$ *and a map* T,

$$T(\mathscr{U} \cap \mathscr{V}) \subset (T\mathscr{U}) \cap (T\mathscr{V})$$

with equality if and only if

$$(\mathscr{U} + \mathscr{V}) \cap \mathscr{N}(T) = \mathscr{U} \cap \mathscr{N}(T) + \mathscr{V} \cap \mathscr{N}(T).$$

REFERENCES

[1] W. M. WONHAM AND A. S. MORSE, *Decoupling and pole assignment in linear multivariable systems: A geometric approach*, this Journal, 8 (1970), pp. 1–18.

[2] E. G. GILBERT, *The decoupling of multivariate systems by state feedback*, this Journal, 7 (1969), pp. 50–63.

SIAM J. CONTROL
Vol. 8, No. 3, August 1970

OBSERVABILITY OF NONLINEAR SYSTEMS*

YE. YA. ROĬTENBERG†

Consider a system described by the differential equation

$$\dot{x}_j = \sum_{k=1}^{n} A_{jk}(t)x_k + \varphi_j(x_1, \cdots, x_n) + q_j(t), \qquad j = 1, \cdots, n, \tag{1}$$

where $x_1, \cdots, x_n$ are the phase coordinates of the system, and $A_{jk}(t)$, $j, k = 1, \cdots, n$, are variable coefficients. The $\varphi_j(x_1, \cdots, x_n)$, $j = 1, \cdots, n$, are nonlinear functions, Lipschitz continuous in all arguments in a certain closed region. The $q_j(t)$ are external forces.

Let us assume that the phase coordinates $x_1, \cdots, x_n$ are inaccessible to direct observation; only the linear combination

$$y = \sum_{k=1}^{n} c_k x_k \tag{2}$$

is accessible. The initial condition of the system (1) is assumed unknown. From observations of $y(t)$ over some finite time interval we wish to find the initial condition of the system; that is, we wish to find the values $x_1(0), \cdots, x_n(0)$, or, alternatively, the state of the system at some subsequent time $t = t^*$, that is, the values $x_1(t^*), \cdots, x_n(t^*)$. This is the observability problem [1] for system (1).

Introducing

$$A = \left\| \begin{matrix} A_{11}(t) \cdots A_{1n}(t) \\ \cdot \quad \cdot \quad \cdot \quad \cdot \\ A_{n1}(t) \cdots A_{nn}(t) \end{matrix} \right\|, \quad x = \left\| \begin{matrix} x_1 \\ \cdot \\ \cdot \\ x_n \end{matrix} \right\|, \quad \varphi(x) = \left\| \begin{matrix} \varphi_1(x_1, \cdots, x_n) \\ \cdot \\ \cdot \\ \varphi_n(x_1, \cdots, x_n) \end{matrix} \right\|,$$

$$q(t) = \left\| \begin{matrix} q_1(t) \\ \cdot \\ \cdot \\ q_n(t) \end{matrix} \right\|, \quad c = \|c_1, \cdots, c_n\|, \tag{3}$$

the scalar system (1) can be replaced by the matrix equation

$$\dot{x} = A(t)x + \varphi(x) + q(t). \tag{4}$$

Expression (2) becomes

$$y = cx.$$

Together with the system (1), we shall consider an auxiliary controlled system

$$\dot{\zeta}_j = \sum_{k=1}^{n} A_{jk}(t)\zeta_k + \varphi_j(\zeta_1, \cdots, \zeta_n) + q_j(t) + u_j, \qquad j = 1, \cdots, n. \tag{5}$$

* Originally published in *Vestnik Moskovskogo Universiteta, Matematika, Mekhanika*, 1969, no. 2, pp. 22–29. Submitted October 11, 1968. This translation into English has been prepared by R. N. and N. B. McDonough.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† Department of Differential Equations, Moscow University, Moscow, U.S.S.R.

Here $\zeta_1, \cdots, \zeta_n$ are phase coordinates and $u_j, j = 1, \cdots, n$ are applied controls. We will also introduce the linear combination of phase coordinates of (5) analogous to (2):

$$\eta = \sum_{k=1}^{n} c_k \zeta_k. \tag{6}$$

Introducing vectors

$$\zeta = \begin{Vmatrix} \zeta_1 \\ \cdot \\ \cdot \\ \cdot \\ \zeta_n \end{Vmatrix}, \qquad u = \begin{Vmatrix} u_1 \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{Vmatrix}, \tag{7}$$

we can replace the scalar system (5) by the matrix equation

$$\dot{\zeta} = A(t)\zeta + \varphi(\zeta) + q(t) + u. \tag{8}$$

Using (3) and (7), (6) becomes

$$\eta = c\zeta.$$

The initial state $\zeta(0)$ of (5) will be arbitrary. The controls $u_j = u_j(t)$, $j = 1, \cdots, n$, are to be chosen such that for a certain time $t = t^*$ we have

$$\rho(t) = \left\{ \sum_{j=1}^{n} [\zeta_j(t) - x_j(t)]^2 \right\}^{1/2} < \mu^* \quad \text{for } t \geqq t^*, \tag{9}$$

where μ^* is an arbitrarily small assigned number. All phase coordinates $\zeta_1, \cdots, \zeta_n$ of the system (5) are accessible to observation. Thus if we succeed in finding controls $u_j(t), j = 1, \cdots, n$, which assure condition (9), the observability problem for system (1) will have been solved with the assigned precision.

Let us now turn to a solution of this problem. From (8) and (4) it follows that the vector

$$z = \zeta - x \tag{10}$$

satisfies the differential equation

$$\dot{z} = A(t)z + \varphi(\zeta) - \varphi(x) + u. \tag{11}$$

The scalar system

$$\dot{z}_j = \sum_{k=1}^{n} A_{jk}(t)z_k + \varphi_j(\zeta_1, \cdots, \zeta_n) - \varphi_j(x_1, \cdots, x_n) + u_j, \qquad j = 1, \cdots, n, \tag{12}$$

corresponds to the matrix equation (11). From (10),

$$\zeta_k = x_k + z_k, \qquad k = 1, \cdots, n.$$

On expanding $\varphi(\zeta_1, \cdots, \zeta_n)$ into a Taylor series about $x_1, \cdots, x_n$, the system (12) can be written

$$\dot{z}_j = \sum_{k=1}^{n} A_{jk}(t)z_k + \sum_{k=1}^{n} f_{jk}(x_1(t), \cdots, x_n(t))z_k + R_j(t, z_1, \cdots, z_n) + u_j, \tag{13}$$

$$j = 1, \cdots, n,$$

where

$$f_{jk}(x_1(t), \cdots, x_n(t)) = \left[\frac{\partial \varphi_j(\zeta_1, \cdots, \zeta_n)}{\partial \zeta_k}\right]_{\zeta_1 = x_1(t), \cdots, \zeta_n = x_n(t)}. \tag{14}$$

In equations (13), $R_j(t, z_1, \cdots, z_n)$ denote terms of second and higher orders in $z_1, \cdots, z_n$. Corresponding to the original system (12), we have the first approximation

$$\dot{z}_j = \sum_{k=1}^{n} A_{jk}(t) z_k + \sum_{k=1}^{n} f_{jk}(x_1(t), \cdots, x_n(t)) z_k + u_j, \qquad j = 1, \cdots, n. \tag{15}$$

Controls u_j will be assumed to be of the form

$$u_j = b_j v, \qquad j = 1, \cdots, n, \tag{16}$$

where v is a scalar time function, accessible to observation,

$$v = \eta - y, \tag{17}$$

and $b_j, j = 1, \cdots, n$, are certain constants to be found. From (6), (2) and (10),

$$v = \sum_{k=1}^{n} c_k z_k,$$

and expressions (16) become

$$u_j = b_j \sum_{k=1}^{n} c_k z_k. \tag{18}$$

Equations (15) take the form

$$\dot{z}_j = \sum_{k=1}^{n} A_{jk}(t) z_k + \sum_{k=1}^{n} f_{jk}(x_1(t), \cdots, x_n(t)) z_k + b_j \sum_{k=1}^{n} c_k z_k, \qquad j = 1, \cdots, n. \tag{19}$$

Here $x_1(t), \cdots, x_n(t)$ are certain functions determined by the differential equations (1). Since the initial state $x(0)$ of the system (1) is unknown, the functions $x_k(t), k = 1, \cdots, n$, are unknown. Nevertheless, having assigned a region of possible initial states of system (1), we can find the region of possible values of the functions $x_k(t), k = 1, \cdots, n$. Then we can find upper and lower bounds of the functions

$$\mu_{jk}(t) = f_{jk}(x_1(t), \cdots, x_n(t)), \qquad j, k = 1, \cdots, n, \tag{20}$$

on the set of possible motions of system (1):

$$\alpha_{jk} \leqq \mu_{jk}(t) \leqq \beta_{jk}. \tag{21}$$

With (20), the differential equations (19) become

$$\dot{z}_j = \sum_{k=1}^{n} A_{jk}(t) z_k + \sum_{k=1}^{n} \mu_{jk}(t) z_k + b_j \sum_{k=1}^{n} c_k z_k, \qquad j = 1, \cdots, n. \tag{22}$$

We have here a system of linear differential equations with variable coefficients $\mu_{jk}(t), j, k = 1, \cdots, n$, the lower and upper bounds of which are known. The functions $\mu_{jk}(t)$ themselves are unknown.

Let us now consider how to assure a sufficiently rapid decrease of the difference between the solutions of systems (5) and (1). If we introduce

$$\mu(t) = \begin{Vmatrix} \mu_{11}(t) \cdots \mu_{1n}(t) \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \mu_{n1}(t) \cdots \mu_{nn}(t) \end{Vmatrix}, \qquad b = \begin{Vmatrix} b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{Vmatrix}, \tag{23}$$

the scalar system (22) can be replaced by the matrix equation

$$\dot{z} = [A(t) + \mu(t) + bc]z. \tag{24}$$

Let us now introduce new variables $\gamma_j, j = 1, \cdots, n$, related to z_j by

$$z_j = e^{-\sigma t}\gamma_j, \qquad \sigma > 0. \tag{25}$$

On introducing

$$\gamma = \begin{Vmatrix} \gamma_1 \\ \cdot \\ \cdot \\ \cdot \\ \gamma_n \end{Vmatrix},$$

equation (25) becomes

$$z = e^{-\sigma t}\gamma. \tag{26}$$

Since from (26),

$$\dot{z} = -\sigma e^{-\sigma t}\gamma + e^{-\sigma t}\dot{\gamma},$$

equation (24) can be written

$$\dot{\gamma} = [A(t) + \sigma E + \mu(t) + bc]\gamma, \tag{27}$$

where E is the identity matrix. We have the following theorem.

THEOREM 1. *For the observability of system* (1) *it is sufficient to select the vector* b, *whose elements are the coefficients of the controls* (18), *such that it is possible to construct for the system* (27) *a Lyapunov system* V, *that is, a sign definite function whose derivative* dV/dt, *constructed using* (27), *is a sign definite function of sign opposite to that of* V, *for any point of the region of the* n^2*-dimensional space* $(\mu_{11}, \cdots, \mu_{1n}, \mu_{21}, \cdots, \mu_{2n}, \cdots, \mu_{n1}, \cdots, \mu_{nn})$ *determined by the inequalities* (21).

Proof. According to (25), (18) and (20) the nonlinear system (13) transforms into

$$\dot{\gamma}_j = \sum_{k=1}^{n} [A_{jk}(t) + \sigma E_{jk} + \mu_{jk}(t) + b_j c_k]\gamma_k + Q_j(t, \gamma_1, \gamma_2, \cdots, \gamma_n), \tag{28}$$

$$j = 1, \cdots, n,$$

where $Q_j(t, \gamma_1, \cdots, \gamma_n)$ are terms of second and higher orders in $\gamma_1, \cdots, \gamma_n$. According to a theorem of Malkin [2, p. 375], if over the region

$$t > 0, \qquad |\gamma_j| \leqq H, \tag{29}$$

the $Q_j(t, \gamma_1, \cdots, \gamma_n)$ satisfy

$$|Q_j(t, \gamma_1, \cdots, \gamma_n)| < a[|\gamma_1| + \cdots + |\gamma_n|], \tag{30}$$

then, under the hypothesis of the theorem, the solution $\gamma_j = 0, j = 1, \cdots, n$, of (28) will be asymptotically stable if the constant a is sufficiently small. Then from Lyapunov's theorem on asymptotic stability, we can conclude that, for any $\varepsilon > 0$, there exists a number $T(\varepsilon)$ such that

$$[\gamma_1^2(t) + \cdots + \gamma_n^2(t)]^{1/2} < \varepsilon \quad \text{for } t \geqq T(\varepsilon). \tag{31}$$

Since from (25),

$$z_j = e^{-\sigma t}\gamma_j, \qquad \sigma > 0, \quad j = 1, \cdots, n,$$

from (31) it follows that, for a suitable choice of σ, we can assure fulfillment of condition (9):

$$\rho(t) = [z_1^2(t) + \cdots + z_n^2(t)]^{1/2} < \mu^* \quad \text{for } t \geqq t^*$$

for given μ^* and t^*.

The observability conditions for system (1), with the above estimates for the functions $Q_j(t, \gamma_1, \cdots, \gamma_n)$ in (28), are given in the following theorem.

THEOREM 2. *If the nonlinear functions* $Q_j(t, \gamma_1, \cdots, \gamma_n), j = 1, \cdots, n$, *in* (28) *satisfy*

$$|Q_j(t, \gamma_1, \gamma_2, \cdots, \gamma_n)| < \frac{(1 - q)v_3(\gamma_1^2 + \cdots + \gamma_n^2)^{1/2}}{nv_4}, \qquad 0 < q < 1, \tag{32}$$

then for the observability of system (1) *it suffices to select a vector* b, *whose elements are the coefficients of the controls* (18), *such that, for the linear system* (27), *it is possible to construct a Lyapunov function* W *satisfying*

$$\begin{gathered} v_1(\gamma_1^2 + \cdots + \gamma_n^2) \leqq W \leqq v_2(\gamma_1^2 + \cdots + \gamma_n^2), \\ \frac{dW}{dt} \leqq -qv_3(\gamma_1^2 + \cdots + \gamma_n^2), \\ \left|\frac{\partial W}{\partial \gamma_j}\right| \leqq v_4(\gamma_1^2 + \cdots + \gamma_n^2)^{1/2} \end{gathered} \tag{33}$$

($v_1, \cdots, v_4$ *are positive constants, and* $0 < q < 1$) *at any point of the region of the* n^2*-dimensional space* $(\mu_{11}, \cdots, \mu_{1n}, \mu_{21}, \cdots, \mu_{2n}, \cdots, \mu_{n1}, \cdots, \mu_{nn})$ *determined by the inequalities* (21).

Proof. According to a theorem of N. N. Krasovskii [3, p. 102], it follows from conditions (32) and (33) that the solution of the system (28), for any initial condition $\gamma_j(t_0)$, t_0 satisfies

$$[\gamma_1^2(t) + \cdots + \gamma_n^2(t)]^{1/2} \leqq B[\gamma_1^2(t_0) + \cdots + \gamma_n^2(t_0)]^{1/2}e^{-\alpha(t-t_0)} \quad \text{for } t \geqq t_0, \tag{34}$$

where

$$B = \frac{v_2}{v_1}, \qquad \alpha = \frac{qv_3}{v_2}.$$

Since from (25),

$$z_j = e^{-\sigma t}\gamma_j, \qquad \sigma > 0, \quad j = 1, \cdots, n,$$

by taking $t_0 = 0$ in (34) we find that for $t > 0$,

$$\rho(t) = [z_1^2(t) + \cdots + z_n^2(t)]^{1/2}$$

will decrease at least as fast as $e^{-\sigma t}$. With an appropriate selection of σ this assures $\rho(t) < \mu^*$ for $t \geqq t^*$ for the assigned values of μ^* and t^*.

As an example, let us examine the system

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -x_1 - \varepsilon x_2 - \sin x_1. \end{aligned} \tag{35}$$

Let us assume that the phase coordinate

$$y = x_1 \tag{36}$$

is available. In this case, matrices (3) are

$$A = \begin{Vmatrix} 0 & 1 \\ -1 & -\varepsilon \end{Vmatrix}, \quad x = \begin{Vmatrix} x_1 \\ x_2 \end{Vmatrix}, \quad \varphi(x) = \begin{Vmatrix} \varphi_1(x_1, x_2) \\ \varphi_2(x_1, x_2) \end{Vmatrix} = \begin{Vmatrix} 0 \\ -\sin x_1 \end{Vmatrix},$$

$$c = \| 1 \quad 0 \|.$$

The auxiliary system (5) here has the form

$$\begin{aligned} \dot{\zeta}_1 &= \zeta_2 + u_1, \\ \dot{\zeta}_2 &= -\zeta_1 - \varepsilon\zeta_2 - \sin \zeta_1 + u_2. \end{aligned} \tag{37}$$

The linear combination of phase coordinates of system (37) analogous to (36) is

$$\eta = \zeta_1.$$

The vector

$$z = \zeta - x$$

has the form

$$z = \begin{Vmatrix} z_1 \\ z_2 \end{Vmatrix} = \begin{Vmatrix} \zeta_1 - x_1 \\ \zeta_2 - x_2 \end{Vmatrix}. \tag{38}$$

According to (17) and (23),

$$v = \zeta_1 - x_1 = z_1, \qquad b = \begin{Vmatrix} b_1 \\ b_2 \end{Vmatrix}.$$

From (16), the controls u_j, $j = 1, 2$, will be

$$\begin{aligned} u_1 &= b_1 z_1, \\ u_2 &= b_2 z_1. \end{aligned} \tag{39}$$

According to (14) and (20), the functions $\mu_{jk}(t)$, $j, k = 1, 2$, here are

$$\mu_{11}(t) \equiv 0, \qquad \mu_{12}(t) \equiv 0, \qquad \mu_{22}(t) \equiv 0,$$

$$\mu_{21}(t) = \left[\frac{\partial \varphi_2(\zeta_1, \zeta_2)}{\partial \zeta_1}\right]_{\zeta_1 = x_1(t), \zeta_2 = x_2(t)} = -\cos x_1(t),$$

and thus estimates (21) will be

$$-1 \leqq \mu_{21}(t) \leqq 1. \tag{40}$$

Equations (19) become

$$\begin{aligned} \dot{z}_1 &= z_2 + b_1 z_1, \\ \dot{z}_2 &= -z_1 - \varepsilon z_2 + \mu_{21}(t) z_1 + b_2 z_1. \end{aligned} \tag{41}$$

Let us now change to the new variables γ_j, $j = 1, 2$, related to z_j by (25).

According to (41), equations (27) become

$$\begin{aligned} \dot{\gamma}_1 &= (\sigma + b_1)\gamma_1 + \gamma_2, \\ \dot{\gamma}_2 &= [-1 + \mu_{21}(t) + b_2]\gamma_1 + (-\varepsilon + \sigma)\gamma_2. \end{aligned} \tag{42}$$

As the Lyapunov function for the system (42) we shall use

$$V = -\tfrac{1}{2}(\gamma_1^2 + \gamma_2^2).$$

Its derivative

$$\frac{dV}{dt} = -\gamma_1 \dot{\gamma}_1 - \gamma_2 \dot{\gamma}_2$$

becomes

$$\frac{dV}{dt} = a_{11}\gamma_1^2 + 2a_{12}\gamma_1\gamma_2 + a_{22}\gamma_2^2, \tag{43}$$

after substitution of $\dot{\gamma}_1$ and $\dot{\gamma}_2$ from (42), and where

$$a_{11} = -(b_1 + \sigma), \quad a_{12} = -\tfrac{1}{2}[b_2 + \mu_{21}(t)], \quad a_{22} = \varepsilon - \sigma. \tag{44}$$

The discriminant of the quadratic form (43) is

$$D = \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix}. \tag{45}$$

That the trivial solution of the system (42) be asymptotically stable requires that, for any time t, the principal minors of the discriminant (45) be greater than some arbitrarily small positive number l:

$$a_{11} > l, \quad a_{12}a_{22} - a_{12}^2 > l. \tag{46}$$

Using (44), conditions (46) become

$$-(b_1 + \sigma) > l, \quad -(b_1 + \sigma)(\varepsilon - \sigma) - \tfrac{1}{4}[b_2 + \mu_{21}(t)]^2 > l. \tag{47}$$

Taking into account the estimate (40), b_1 and b_2 should be chosen so as to satisfy conditions (47).

As an example, let us consider the case $\varepsilon = 5$, $\sigma = 4$. The values $b_1 = -8$, $b_2 = 1.5$ satisfy conditions (47) for any function $\mu_{21}(t)$ with bounds (40). The difference between the phase trajectories of systems (35) and (37) (with controls (39) and initial conditions $x_1(0) = 0.5$, $x_2(0) = 0.3$, $\zeta_1(0) = -0.3$, $\zeta_2(0) = -0.1$), at $t^* = 1.5$ sec. is $z_1(t^*) = -5.19 \times 10^{-5}$, $z_2(t^*) = -1.33 \times 10^{-4}$; and at $t^{**} = 3$ sec. the difference is $z_1(t^{**}) = -2 \times 10^{-8}$, $z_2(t^{**}) = -6 \times 10^{-8}$.

REFERENCES

[1] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), no. 1, pp. 102–119.

[2] I. G. MALKIN, *The Theory of the Stability of Motion*, Nauka, Moscow, 1966.

[3] N. N. KRASOVSKII, *Certain Problems of the Theory of the Stability of Motion*, Fizmatgiz, Moscow, 1959.

SIAM J. CONTROL
Vol. 8, No. 3, August 1970

FILTERING FOR LINEAR DISTRIBUTED PARAMETER SYSTEMS*

HAROLD J. KUSHNER†

1. Introduction. In this paper, we develop a filtering theory for linear parabolic systems which are driven by white noise, and where white noise corrupted observations are taken.

Consider the purely formal equation

$$\dot{U}(x,t) = \mathscr{L}U(x,t) + \sigma(x,t)\xi_t, \tag{1}$$

where $\mathscr{L}$ is a parabolic operator and ξ_t is Gaussian "white noise." An interpretation of (1), which is an extension of the Ito interpretation of stochastic ordinary differential equations, is stated in Lemma 3.

Lemma 3 requires some results concerning criteria guaranteeing smoothness of random surfaces and these, based on [1], [2], are reviewed in § 2. The first filtering problem treated (§ 3) deals with a first boundary problem for (1) with internal observations of the form (9), or, formally, the data, for $s \leqq t$,

$$\text{observation} = \dot{y}_s = \int H(x,s)U(x,s)\,dx + \psi_s$$

is available at time t, where ψ_s is "white Gaussian noise."

In order to give physical meaning to (1), Lemma 3 (based on § 2, [1], [2]) gives conditions under which continuous "paths" $U(x,t)$ exist, which also have sufficiently many spacial derivatives for $\mathscr{L}U(x,t)$ to be continuous. If the surface $U(x,t)$ is smooth, it would be nice to know that the conditional mean $M(x,t)$ is also. Lemma 4 shows that there is a version of $M(x,t)$ which is continuous and has as many spacial derivatives as does $U(x,t)$ and also that the conditional covariance is smooth. Theorem 1 shows that $M(x,t)$ satisfies (15), an equation of the form (1), where ξ_t is replaced by a term due to the observation, exactly as is the ordinary differential equation case. Also, the covariance satisfies (16), the relevant "Ricatti" equation.

Section 4 treats a second boundary value problem, where noisy observations are taken on the boundary, and a boundary disturbance drives the system.

2. Smoothness results on random surfaces. Let z_t be a normalized Wiener process, D a bounded open domain in E^n with closure $\bar{D}$ and a continuous and piecewise uniformly differentiable boundary and write $\bar{R} = \bar{D} \times [0,T]$. Let $D_t = \partial/\partial t$, $D_i = \partial/\partial x_i$, $D_i^l = \partial^l/\partial x_j^l$. Let $f(x,t)$ be a stochastic process on $\bar{D} \times [0,T] = \bar{R}$. The parenthesis in $(D_i f(x,t))$ denotes the "mean square" derivative of $f(x,t)$

* Received by the editors February 24, 1969, and in revised form October 2, 1969.

† Division of Applied Mathematics, Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912. This research was supported in part by the National Aeronautics and Space Administration under Grant NGL-40-002-015, and in part by the Air Force Office of Scientific Research under Grant AFOSR 693–67.

with respect to x_i, if it exists. Define the norm

$$\|g(x)\|_{W_{l,p}(\bar{D})} = \sum_{k=0}^{l} \sum_{l_1+\cdots+l_n=k} \|D_1^{l_1} \cdots D_n^{l_n} g(x)\|_{L_p(\bar{D})}, \tag{2}$$

where $\psi \in L_p(\bar{D})$ means that $\int_D |\psi(x)|^p \, dx \equiv \|\psi\|^p_{L_p(\bar{D})} < \infty$. References [1] and [2], from which Lemmas 1 and 2 are taken, give conditions on the expectations of integrals of powers of the mean square derivatives, which guarantee that $f(x, t)$ has a (with probability 1) continuous version on $\bar{R}$, and perhaps several continuous derivatives with respect to components of x. The proof of Lemma 1 is contained in [2].

LEMMA 1. *Let the boundary ∂D of D have the property that any line intersects it only finitely often. Let the functions*

$$\begin{aligned} &\alpha(x, t, s), \quad \{D_i\alpha(x, t, s)\}, \quad \{D_iD_j\alpha(x, t, s)\}, \\ &\{D_iD_jD_k\alpha(x, t, s)\}, \qquad \{D_iD_jD_kD_l\alpha(x, t, s)\} \end{aligned} \tag{3}$$

be defined on $\bar{D} \times [0, T] \times [0, T] = \bar{R} \times [0, T]$, continuous in (x, t) for each s, and bounded (in absolute value) by a square integrable function of s. Let f be any function in the set (3), *and let $z(t)$ be a Wiener process. Then $\int_0^T f^2(x, t, s)\, ds \leqq M < \infty$ for some real number M, and $\int_0^t f(x, t, s)\, dz_s$ can be defined to be a separable and measurable process with parameter (x, t). There is a null set N and a separable and measurable version of $\int_0^t \alpha(x, t, s)\, dz_s = \psi(x, t)$ which, for $\omega \notin N$, is continuous in (x, t) and has three continuous (in (x, t)) derivatives with respect to the components of x. These derivatives are equal to continuous (for $\omega \notin N$), separable and measurable versions of $\int_0^t D_i\alpha(x, t, s)\, dz_s$, $\int_0^t D_iD_j\alpha(x, t, s)\, dz_s$, $\int_0^t D_iD_jD_k\alpha(x, t, s)\, dz_s$, respectively.*

Let in addition, for some real numbers $K < \infty$, $\beta > 0$,

$$\begin{aligned} &E\left\{\int_0^{t+\Delta} f(x, t+\Delta, s)\, dz_s - \int_0^t f(x, t, s)\, dz_s\right\}^2 \\ &\quad = \int_0^t [f(x, t+\Delta, s) - f(x, t, s)]^2\, ds + \int_t^{t+\Delta} f^2(x, t+\Delta, s)\, ds \leqq K\Delta^\beta, \end{aligned} \tag{4}$$

where f is any member of (3). *Let g be any member of the first three sets of* (3). *Then the continuous version (for $\omega \notin N$) of $\int_0^t g(x, t, s)\, dz_s = \phi(x, t)$ is Hölder continuous on $\bar{R}$, i.e., there is some $K(\omega) < \infty$ w.p.1 and a real $\gamma > 0$ so that*

$$|\phi(x+\delta, t+\Delta) - \phi(x, t)| \leqq K(\omega)[|\Delta|^\gamma + |\delta|^\gamma],$$

where $|\cdot|$ refers to the Euclidean norm.

LEMMA 2. *Let $f(x,t)$ be a process on $\bar{R}$ which is continuous in probability together with its mean square derivatives up to order l on $\bar{R}$. Let $pl > n$, $p > 1$, and suppose that*[1] *for $0 \leqq s \leqq t \leqq T$,*

$$E\|f(\cdot,t) - f(\cdot,s)\|^q_{W_{l,p}(\bar{D})} \leqq K|t-s|^{1+\alpha} \tag{5}$$

for some real $K < \infty$ and $1 \leqq q < \infty$ and $\alpha > 0$. Then there is a w.p.1 continuous version of $f(\cdot,\cdot)$ on $\bar{R} \times [0,T]$, and the version is Hölder continuous in t, uniformly in x, w.p.1.

If $0 < m < l - n/p$, then the mean square derivatives of order $\leqq m$ have continuous versions on $\bar{R}$ w.p.1, and $f(x,t)$ has w.p.1 a continuous version whose first m x-derivatives coincide with the mean square square derivatives.

For proof, see Theorem 4 in [1].

3. Filtering for a stochastic first boundary value problem.

3.1. System model. The first system with which we will deal has the representation[2]

$$dU(x,t) = \left[\mathscr{L}U(x,t) + \int k(y,x,t)U(y,t)\,dy\right]dt + \sigma(x,t)\,dz, \tag{6}$$

where

$$\mathscr{L} = \sum a_{ij}(x,t)D_iD_j + \sum b_i(x,t)D_i \tag{7}$$

and (A1)–(A7) hold.

(A1) ∂D (the boundary of D) has a local representation with Hölder continuous 4th derivatives.

(A2) The coefficients of $\mathscr{L}$ and their first two derivatives are Hölder continuous in $\bar{R}$.

(A3) $\sum a_{ij}\xi_i\xi_j \geqq K\sum \xi_i^2$ for some real $\infty > K > 0$.

(A4) σ and its first four x-derivatives are Hölder continuous on $\bar{R}$.

(A5) σ and $\mathscr{L}\sigma$ go to zero as $x \to \partial D$.

(A6) $k(y,x,t)$ is bounded, measurable and Hölder continuous in x,t, uniformly in y, and $k(y,x,t) \to 0$ as $x \to \partial D$.

(A7) $U(x,0)$ is Gaussian for each x, has a bounded variance, has Hölder continuous second derivatives, and $U(x,0)$ and $\mathscr{L}U(x,0) \to 0$ as $x \to \partial D$. $U(x,0)$ is independent of z_t and of w_t (to be introduced below).

In [2], Lemmas 1 and 2 are applied to (6) to give it a precise definition which is summarized in the following lemma.

---

[1] Recall that (5) is equivalent to

$$E\left[\int_{\bar{D}} |(D_1^{l_1}\cdots D_n^{l_n}\{f(x,t) - f(x,s)\})|^p\,dx\right]^{q/p} \leqq K|t-s|^{1+\alpha}$$

for all $l_1 + \cdots + l_n \leqq k \leqq l$, for $0 \leqq s \leqq t \leqq T$.

[2] For notational simplicity, we let the "driving term" be $\sigma(x,t)\,dz$. It could be $\sum \sigma_i(x,t)\,dz_i$, where the z_i are independent. See [2, Lemma 2.2].

LEMMA 3[3] (see [2, Lemma 3.2] for proof). *Assume* (A1)–(A7). *Then there is a random function* $U(x, t)$ *on* $(0, T] \times \bar{D}$ *so that a version (for* $\omega \notin N$, *a null set) of the uniformly (in* $(0, T] \times D$*) mean square continuous functions*

$$U(x,t), (D_iU(x,t)), \cdots, (D_iD_jD_kU(x,t)) \tag{8}$$

are continuous on $(0, T] \times \bar{D}$ *w.p.*1; *these versions of the mean square derivatives are true derivatives.* $U(x,t)$ *and* $\mathscr{L}U(x,t) \to 0$ *as* $x \to \partial D$ *(for* $\omega \notin N$*),* $U(x,t) \to U(x,0)$ *(for* $\omega \notin N$, *and uniformly in* x*) as* $t \to 0$. *The first three sets of* (8) *are Hölder continuous in* t *for* $\omega \notin N$. $U(x,t)$ *is a Markov process (with values in a state space of functions with Hölder continuous second derivatives). For any finite set of arguments* $\{x_\alpha, t_\alpha, \alpha = 1, \cdots, s\}$, *the random variables in* (8) $\{U(x_\alpha, t_\alpha), \cdots, (D_iD_jD_kU(x_\alpha, t_\alpha)), \alpha = 1, \cdots, s\}$ *are jointly Gaussian. The random variables in* (8) *have uniformly (over* (x,t)*) bounded variances. The variances of* $U(x,t)$ *and of* $\mathscr{L}U(x,t)$ *tend to zero as* $x \to \partial D$. $U(x,t)$ *is nonanticipative with respect to the* z_t *process, and the Itô differential of* $U(x,t)$ *satisfies* (6). $U(x,t)$ *satisfies the condition* (5) *of Lemma* 2 *for* $m = 3$, $l = 4$, *and all large* p, *and some finite* q *and* $\alpha > 0$. $(D_iD_jD_kD_lU(x,t))$ *is also uniformly mean square continuous in* $(0, T) \times R$.

3.2. The filtering problem. Let $\tilde{w}_t$ be a normalized Wiener process independent of the z_t process, and suppose that:

(A8) $H(x,t)$ is a vector-valued function which is defined and continuous on $\bar{R}$.

(A9) $B(t)$ is continuous on $[0, T]$ and $B(t)B'(t) = \Sigma_t$ is strictly positive definite on $[0, T]$.

Define $w_s = \int_0^s B(\tau)\, d\tilde{w}_\tau = \int_0^s \Sigma_\tau^{1/2}\, d\tilde{w}_\tau$. Suppose that the data

$$y(s) = \int_0^s \left[\int H(x,\tau)U(x,\tau)\, dx \right] d\tau + \int_0^s B(\tau)\, d\tilde{w}_\tau \equiv \int_0^s h_\tau\, d\tau + w_s, \qquad s \leqq t, \tag{9}$$

is available at time t. All introduced σ-algebras are assumed to be complete with respect to whatever measures are imposed on them; let[4] $\mathscr{F}_t$ be the minimal σ-algebra determined by $y(s)$, $s \leqq t$. Let μ_1 be the measure determined by the processes $U(x,s)$, $s \leqq t$, and $dy(s) = h_s\, ds + dw_s$, $s \leqq t$, and μ_0 the measure determined by the processes $U(x,t)$ and $dy(s) = dw_s$, $s \leqq t$. Let E_i^t denote the expectation with respect to μ_i, and conditioned on $\mathscr{F}_t$.

[3] The smoothness in (A1), (A2), (A4) gives a $U(x,t)$ with continuous third x-derivatives, hence Hölder continuous second derivatives. In the control problem in [2], we wanted $U(x,t)$ to have Hölder continuous second derivatives. If only continuous second derivatives are required, then the differentiability requirements in (A1), (A2), (A4) can be reduced by one.

[4] To be more precise, let Ω be a function space with generic element $\omega = (\omega', \omega'')$, where ω' is a member of the space of bounded functions on $\bar{R}$, and ω'' is a member of the space of bounded functions on $[0, T]$, with values in the Euclidean m-dimensional space E^m, where m is the dimension of w_t and y_t. The terminology is used later. See Part 1 of the proof of Theorem 1.

Define

$$M(x,t) = E_1^t U(x,t),$$

$$P(x,y,t) = E_1^t(U(x,t) - M(x,t))(U(y,t) - M(y,t))$$
$$= E_1(U(x,t) - M(x,t))(U(y,t) - M(y,t)).$$

The equality of the two expressions for $P(x, y, t)$ follows from the Gaussianess of $(U(x,t), U(y,t))$.

Let $\mathscr{L}_x P(x,y,t)$ denote $\mathscr{L}$ operating on $P(x,y,t)$ as a function of x. Lemma 4 proves that there is a version of the estimate $M(x,t)$ which is, w.p.1, as smooth as the signal $U(x,t)$.

LEMMA 4. *Assume* (A1)–(A9). *Then (excluding a null set independent of* (x,t)*) there are on* $(0,T] \times \bar{D}$ *continuous versions of the first four sets of the continuous in quadratic mean functions*

$$M(x,t), (D_iM(x,t)), (D_iD_jM(x,t)), (D_iD_jD_kM(x,t))(D_iD_jD_kD_lM(x,t)); \tag{10}$$

also $M(x,t) \to M(x,0)$ *as* $t \to 0$ *(and also in quadratic mean) and the first three sets of mean square derivatives are true derivatives and* $E_1^t \mathscr{L} U(x,t) = \mathscr{L} M(x,t)$; *also* $M(x,t)$ *and* $\mathscr{L}M(x,t) \to 0$ *as* $x \to \partial D$, *and the first three sets of* (9) *have Hölder continuous versions.* $P(x,y,t)$ *has continuous third derivatives in the components of* x *and* y *on* $(0,T] \times D$, *and* $P(x,y,t)$ *and* $\mathscr{L}_x P(x,y,t)$ *and* $\mathscr{L}_y P(x,y,t) \to 0$ *as* $x \to \partial D$ *or* $y \to \partial D$. $P(x,y,t) \to P(x,y,0)$ *as* $t \to 0$.

Proof. $M(x,t)$ and the $(D_iM(x,t)), \cdots, (D_iD_jD_kD_lM(x,t))$ exist and are mean square continuous in x, uniformly in (x,t) in $(0,T] \times D$. Also $E_1^t(D_iU(x,t)) = (D_iM(x,t))$ w.p.1 (as well as for the next three derivatives) for each x, t in $(0,T] \times D$. These assertions easily follow from estimates of the following type: let e_i be the ith coordinate direction in E^n, where n is the dimension of x. Then

$$E_1\left|\frac{M(x + e_i\Delta, t) - M(x,t)}{\Delta} - E_1^t(D_iU(x,t))\right|^2$$

$$= E_1\left|E_1^t\left\{\frac{U(x + e_i\Delta, t) - U(x,t)}{\Delta} - (D_iU(x,t))\right\}\right|^2 \tag{11}$$

$$\leqq E_1\left|\frac{U(x + e_i\Delta, t) - U(x,t)}{\Delta} - (D_iU(x,t))\right|^2 \to 0$$

as $\Delta \to 0$, uniformly for x, t in $D \times (0,T]$.

Furthermore, $M(x,t)$ also satisfies the estimates (12) and (13):

$$E_1|M(x,t)|^u = E_1|E_1^t U(x,t)|^u \leqq E_1|U(x,t)|^u,$$
$$E_1|(D_iD_jD_kD_lM(x,t))|^u \leqq E_1|(D_iD_jD_kD_lU(x,t))|^u \tag{12}$$

and

$$E_1|M(x,t) - M(x,s)|^u = E_1|E_1^t U(x,t) - E_1^s U(x,s)|^u$$
$$\leqq KE_1|E_1(U(x,t) - U(x,s))|^u + KE_1|E_1^t U(x,t) - E_1^s U(x,t)|^u \tag{13}$$
$$\leqq KE_1|U(x,t) - U(x,s)|^u + \varepsilon_u(x,t,s),$$

where

$$\varepsilon_u(x,t,s) = KE_1|E_1^t U(x,t) - E_1^s U(x,t)|^u.$$

In deriving the last inequality $|E_1^t f|^u \leqq E_1^t |f|^u$ was used.

We will next show that $\varepsilon_2(x, t, s) \leqq K_1|t - s|$. From this, and the Gaussian property, it follows that $\varepsilon_{2m}(x, t, s) \leqq K_m|t - s|^m$ for some real sequence K_m. Define $\Delta = (t - s)/n$, $N_n = [n/(t - s) - 1]$ (nearest integer),

$$Y_{ni} = \frac{1}{\Delta}\int_{i\Delta+s}^{(i+1)\Delta+s} h_s\, ds + W_{ni} = H_{ni} + W_{ni},$$

$$W_{ni} = \frac{1}{\Delta}\int_{i\Delta+s}^{(i+1)\Delta+s} .$$

Define $E_1^n U(x, t) = E_1[U(x, t)|y_\tau, \tau \leqq s; Y_{n0}, \cdots, Y_{nN_n}]$. Then $E_1^n U(x, t) \to E_1^t U(x, t)$ in probability. Since, for some subsequence,

$$(E_1^n U(x, t) - E_1^s U(x, t))^2 \to (E_1^t U(x, t) - E_1^s U(x, t))^2 \quad \text{w.p.1},$$

Fatous' lemma gives

$$E_1|E_1^t U(x, t) - E_1^s U(x, t)|^2 \leqq \liminf_n E_1|E_1^n U(x, t) - E_1^s U(x, t)|^2. \tag{14}$$

We only need show that the expectation on the right side of (14) satisfies the asserted inequality for $\varepsilon_2(x, t, s)$. Next, we may write $E_1^n U(x, t)$ in the form

$$E_1^n U(x, t) = E_1^s U(x, t) + \sum_{\alpha=0}^{N_n} Q_{n\alpha}(Y_{n\alpha} - E_1^s Y_{n\alpha})$$

for some sequence $\{Q_{n\alpha}\}$. The variance of the sum is

$$e_n \equiv \Sigma_{uy}(n)\Sigma_{yy}^{-1}(n)\Sigma_{uy}'(n),$$

where

$$Y_n' = [(Y_{n0} - E_1^s Y_{n0}), \cdots, (Y_{nN_n} - E_1^s Y_{nN_n})],$$

$$\Sigma_{uy}(n) = E_i U(x, t) Y_n',$$

$$\Sigma_{yy}(n) = E_1 Y_n Y_n'.$$

Since the $\{W_{n\alpha}\}$ are mutually independent in α and have a covariance bounded below (in sense of positive definite matrices) by cI/Δ, for some $c > 0$, and the terms of Y_n have uniformly bounded (in all indices n, α) covariances, $\Sigma_{yy}^{-1}(n) \geqq c_1 I\Delta + O(\Delta^2)$ where $O(\Delta^2)$ is a matrix with entries of the order of Δ^2, uniformly in all indices. Thus

$$\begin{aligned} e_n &\leqq \text{const.} \cdot \Delta \cdot \text{number of diagonal entries in } \Sigma_{yy}^{-1} \\ &\quad + \text{const.} \cdot \Delta^2 \cdot \text{total number of entries in } \Sigma_{yy}^{-1} \\ &\leqq \text{const.}\left[\Delta \cdot \frac{(t - s)}{\Delta 2} + \Delta^2 \cdot \frac{(t - s)^2}{\Delta 2}\right] \end{aligned}$$

which yields the desired bound, since the constant can be made independent of n.

Estimates (11), (12), (13) imply that the $M(x,t), \cdots, (D_iD_jD_kD_lM(x,t))$ are uniformly mean square continuous in $(0,T] \times D$.

The statements concerning the continuity of $M(x,t)$ and its derivatives then follow from Lemmas 2 and 3, since by the estimates (12), (13) (and obvious similar estimates for $(D_iM(x,t)), \cdots, (D_iD_jD_kD_lM(x,t))$, if $U(x,t)$ satisfies Lemma 2 for $m = 3$, so does $M(x,t)$. Since (13) is valid for $s = 0$, $M(x,t) \to M(x,0)$.

$M(x,t)$ and $\mathscr{L}M(x,t) \to 0$ as $x \to \partial D$ since both $U(x,t)$ and $\mathscr{L}U(x,t)$ and their variances $\to 0$ as $x \to \partial D$.

The asserted smoothness of $P(x,y,t)$ and its boundary properties follows from the continuity in quadratic mean of the elements of (10), $\mathscr{L}M(x,t)$, $\mathscr{L}U(x,t)$ and (10′):

$$U(x,t), \cdots, (D_iD_jD_kD_lU(x,t)) \tag{10′}$$

(see, for example, Loeve [6, § 34.2] for the type of calculations which are required).

THEOREM 1. *Assume* (A1)–(A9). *Then there is a version of* $M(x,t)$ *which has the Itô differential w.p.*1:

$$\begin{aligned} dM(x,t) = &\left[\mathscr{L}M(x,t) + \int k(\xi,x,t)M(\xi,t)\,d\xi\right]dt \\ &+ \left[dy - \int H(\xi,t)M(\xi,t)\,d\xi\right]' \Sigma_t^{-1}\left[\int H(\xi,t)P(\xi,x,t)\,d\xi\right] \end{aligned} \tag{15}$$

*and, for this version, w.p.*1, $M(x,t)$ *and* $\mathscr{L}M(x,t) \to 0$ *as* $x \to \partial D$. *Furthermore,* $P(x,y,t)$ *satisfies*

$$\begin{aligned} P_t(x,y,t) = &[\mathscr{L}_x + \mathscr{L}_y]P(x,y,t) \\ &+ \int k(y,\xi,t)P(x,\xi,t)\,d\xi + \int k(\xi,x,t)P(\xi,y,t)\,d\xi \\ &+ \sigma(x,t)\sigma(y,t) - \left[\int H(\xi,t)P(x,\xi,t)\,d\xi\right]' \\ &\cdot \Sigma_t^{-1}\left[\int H(\xi,t)P(\xi,y,t)\,d\xi\right]. \end{aligned} \tag{16}$$

$P(x,y,t)$, $\mathscr{L}_xP(x,y,t)$ *and* $\mathscr{L}_yP(x,y,t) \to 0$ *as* $x \to \partial D$.

Proof. For the sake of keeping a framework which will allow a generalization (not proved here) to nonlinear systems, we take a slightly more general approach than necessary. The nonlinear problem for ordinary stochastic Ito equations was treated in [3]; however, here we follow a slightly different approach, due to Zakai [4], which gives the result under weaker conditions than those required in [3].

Part 1. μ_0 and μ_1 are absolutely continuous with respect to one another and $d\mu_1/d\mu_0 = \exp R_t$, where

$$R_t = -\frac{1}{2}\int_0^t h_s'\Sigma_s^{-1}h_s\,ds + \int_0^t h_s'\Sigma_s^{-1}\,dy_s.$$

This statement can be verified in the following way. Define v_0, v_1, v_u as the measures determined by $dy_s = h_s\,ds + dw_s$, $s \leqq t$, $dy_s = dw_s$, $s \leqq t$, and $U(x, s)$, $x \in \bar{D}$, $s \leqq t$, respectively. Let $\mathscr{B}_n$ be the σ-algebra determined by $U(x, s)$, $(x, s) \in \bar{D} \times [0, T]$. $U(x, s)$ is continuous in quadratic mean and separable [2]; hence $\mathscr{B}_u$ is countably generated. Suppose that h_s, $s \leqq t$, is a known function. Then, for any Borel set A of suitable dimension,

$$P_1\{(y_{t_1}, \cdots, y_{t_n}) \in A | h_0 \text{ deterministic}\} = \int_{A'} \left(\frac{dv_1}{dv_0}\right)_h dv_0,$$

where A' is the inverse image of A and

$$\left(\frac{dv_1}{dv_0}\right)_h \equiv \exp R_t.$$

It then follows that

$$P_1\{(y_{t_1}, \cdots, y_{t_n}) \in A | \mathscr{B}_u\} = \int_{A'} \left(\frac{dv_1}{dv_0}\right)_h dv_0.$$

This implies that $(dv_1/dv_0)_h = d\mu_1/d\mu_0$.

Next, following Zakai [4], note that if $E_1|f(\omega, t)| < \infty$, then (see Loeve [5, § 24.4])

$$E_1^t f(\omega, t) = \frac{E_0^t f(\omega, t)(d\mu_1/d\mu_0)}{E_0^t(d\mu_1/d\mu_0)} = \frac{E_0^t f(\omega, t) \exp R_t}{E_0^t \exp R_t}. \tag{17}$$

Part 2. Write

$$F_t = \left[\mathscr{L}U(x, t) + \int k(y, x, t)U(y, t)\,dy\right].$$

Then, w.p.1, by virtue of Lemma 4,

$$E_1^t F_t = \mathscr{L}M(x, t) + \int k(y, x, t)M(y, t)\,dy.$$

In (17), let $f(\omega, t) = U(x, t)$. Both $U(x, t)$ and $\exp R_t$ are stochastic integrals and

$$d[\exp R_s] = (\exp R_s)h_s'\Sigma_s^{-1}\,dy_s.$$

Then Itô's lemma, applied to (17), yields

$$E_1^t U(x, t) = M(x, t) = \frac{E_0^t\left[U(x, 0) + \int_0^t d[U(x, s) \exp R_s]\right]}{E_0^t\left[1 + \int_0^t (\exp R_s)h_s'\Sigma_s^{-1}\,dy_s\right]} \equiv \frac{A_t}{B_t},$$

where we use

$$d[U(x,t)\exp R_t] = E_0^t \int_0^t \Big[U(x,s)(\exp R_s)h_s'\Sigma_s^{-1}\,dy_s + (\exp R_s)(F_s\,ds + \sigma(x,s)\,dz_s)\Big] + E_0^t U(x,0)$$

and where $dy\,dz = 0$ is used to eliminate the $(dU(x,s))(d\exp R_s)$ term from A_t.

As in Kushner [3] or Zakai [4], it can be shown below that,[5] w.p.1,

$$E_0^t \int_0^t (\exp R_s)[F_s\,ds + \sigma(x,s)\,dz_s] = \int_0^t [E_0^s(\exp R_s)F_s]\,ds,$$

$$E_0^t \int_0^t U(x,s)[(\exp R_s)h_s'\Sigma_s^{-1}]\,dy_s = \int_0^t [E_0^s U(x,s)(\exp R_s)h_s'\Sigma_s^{-1}]\,dy_s, \tag{18}$$

$$E_0^t \int_0^t (\exp R_s)h_s'\Sigma_s^{-1}\,dy_s = \int_0^t [E_0^s(\exp R_s)h_s'\Sigma_s^{-1}]\,dy_s,$$

where the second integrals are well-defined w.p.1. Assuming (18) now, we proceed exactly as in [3] and get

$$dM(x,t) = \frac{dA_t}{B_t} - \frac{A_t\,dB_t}{B_t^2} + \frac{A_t(dB_t)^2}{B_t^3} - \frac{(dA_t)(dB_t)}{B_t^2}, \tag{19}$$

where

$$\begin{aligned}
(dB_t)^2 &= E_0^t[(\exp R_t)h_t']\Sigma_t^{-1}[E_0^t(\exp R_t)h_t]\,dt,\\
(dA_t)(dB_t) &= [E_0^t U(x,t)(\exp R_t)h_t']\Sigma_t^{-1}[E_0^t(\exp R_t)h_t]\,dt,\\
dA_t &= (E_0^t U(x,t)\cdot \exp R_t\cdot h_t'\Sigma_t^{-1})\,dy_t + (E_0^t F_t\cdot\exp R_t)\,dt,\\
dB_t &= (E_0^t \exp R_t\cdot h_t'\Sigma_t^{-1})\,dy_t.
\end{aligned}$$

Equation (15) is obtained by substituting in (19) and using the fact (see (17)) that $E_1^t f = [E_0^t f\exp R_t]/E_0^t \exp R_t$.

Part 3. Similarly, dP is calculated from the expression

$$dP(x,y,t) = dE_1^t U(x,t)U(y,t) - dM(x,t)M(y,t).$$

To get $dE_1^t U(x,t)U(y,t)$, repeat the procedure starting with (17), where we now let $f(\omega,t) = U(x,t)U(y,t)$, and use the w.p.1 equalities

$$\begin{aligned}
E_1^t U(x,t)\mathscr{L}_y U(y,t) &= M(x,t)\mathscr{L}_y M(y,t)\\
&\quad + E_1^t(U(x,t) - M(x,t))\mathscr{L}_y(U(y,t) - M(y,t))\\
&= M(x,t)\mathscr{L}_y M(y,t) + \mathscr{L}_y P(x,y,t).
\end{aligned}$$

The details are straightforward and are omitted.

[5] The demonstration of (14) by the method of [3] requires more stringent conditions on R_s and U, than by the method of [4]. The method of [4] is applicable under the conditions of the hypothesis of Theorem 1. The method of [3] may also be applied, by applying it to a suitable sequence of bounded $F_s^\varepsilon, h_s^\varepsilon$ which converges to F_s and h_s in probability.

4. The second boundary value problem. Now, we consider the equations

$$dU(x,t) = [\mathscr{L}U(x,t) - f(x,t)]\,dt - \sigma(x,t)\,dz, \tag{20a}$$

$$U_V(x,t) + \beta(x,t)U(x,t) = g(x,t) + v(x,t)r(t), \tag{20b}$$

$$\mathscr{L} = \sum a_{ij}(x,t)D_iD_j + \sum b_i(x,t)D_i,$$

where $U_V(x,t)$ is the conormal derivative[6] $\partial U/\partial V = \lim_{y\to x, y\in D} \partial U(y,t)/\partial V(x)$ at x on ∂D, and (B1)–(B8) are assumed.

(B1) $\sum a_{ij}(x,t)\xi_i\xi_j \geqq K\sum \xi_i^2$ for some real $K > 0$.

(B2) $a_{ij}(x,t)$ and $b_i(x,t)$ are Hölder continuous in R.

(B3) $f(x,t)$ is continuous, and Hölder continuous in x, uniformly in t.

(B4) ∂D has a local representation with Hölder continuous derivatives.

(B5) Real-valued $g(x,t)$ and row-vector-valued $v(x,t)$ are continuous on $\bar{R}$ and r is the Gaussian random process satisfying $dr = A(t)r\,dt + G(t)\,d\tilde{z}$, where $\tilde{z}_t$ is independent of the z_t and w_t processes introduced earlier, and of $U(x,0)$. $A(t)$ and $G(t)$ are bounded continuous functions.

(B6) The observations $dy = [\int_{\partial D} H(\xi,t)U(\xi,t)\,dS_\xi]\,dt + dw$ are taken, where $H(\xi,t)$ is continuous on $\partial D \times [0,T]$, and w_t is independent of $U(x,0)$, and dS_ξ is the differential surface measure on ∂D. Also Σ_t satisfies (A9), where $dw = \Sigma_t^{1/2}\,d\tilde{w}$, and $\tilde{w}_t$ is a normalized Wiener process.

(B7) Denote $\alpha(x,t,s) = \int_D \Gamma(x,\xi;t,s)\sigma(\xi,s)\,d\xi$, where Γ is the fundamental solution of $D_tU = \mathscr{L}U$. Let $\sigma(\xi,s)$ be uniformly continuous. Let $\gamma(x,t,s)$ represent either $\alpha(x,t,s)$, $D_i\alpha(x,t,s)$ or $D_iD_j\alpha(x,t,s)$. Let, uniformly in $\bar{R}$,

$$\int_t^{t'} \gamma^2(x,t',\tau)\,d\tau + \int_0^t [\gamma(x,t',\tau) - \gamma(x,t,\tau)]^2\,d\tau \leqq K|t'-t|^\beta \tag{21}$$

for some real K and $\beta > 0$. Let $D_iD_jD_k\gamma(x,t,s)$ satisfy (21) uniformly for x,t,t' in any compact subset of $D \times [0,T]^2$.

(B8) Let $U(x,0)$ be differentiable w.p.1, and let $a_{ij}(x,0)$ be continuously differentiable in some neighborhood of ∂D.

LEMMA 5. *Assume* (B1)–(B8). *Then there is a random function* $U(x,t)$ *which has a version with the following properties w.p.*1 (*where the null set does not depend on* x,t):

(a) $U(x,t)$ *is continuous*[7] *on* $\bar{R}$ (*also in quadratic mean*); $(D_iU(x,t))$ *is continuous on compact subsets of* $\bar{D} \times (0,T]$ (*also in quadratic mean*).

(b) *The* $(D_iD_jU(x,t))$ *are continuous on compact subsets of* $D \times (0,T]$.

(c) $U(x,t)$ *has an Itô differential which satisfies* (20a), *for* $t > 0$.

(d) $U(x,t)$ *satisfies the boundary condition* (20b), *and* $U(x,t) \to U(x,0)$ *as* $t \to 0$.

[6] $V(x)$ is the conormal direction at the point x on ∂D.

[7] $D_iU(x,t)$ on ∂D is defined as $\lim_{y\to x, y\in D} D_iU(y,t)$.

(e) *The variances of* $U(x,t)$, $(D_iU(x,t))$ *(in compact subsets of* $\bar{D} \times (0,T]$*) and* $(D_iD_jU(x,t))$ *(in compact subsets of* $D \times (0,T]$*) are uniformly bounded.*

(f) $U(x,t)$ *is nonanticipative with respect to the* z_t *and* $\tilde{z}_t$ *processes.*

Proof. The treatment in Friedman [6, Theorem 2, p. 144 and Corollary 2, p. 147] will be followed, with the few modifications required by the stochastic nature of the problem taken into account. Define

$$\gamma(x,t) = \int_0^t dz_s \alpha(x,t,s), \qquad \gamma_i(x,t) = \int_0^t dz_s D_i\alpha(x,t,s),$$

$$\gamma_{ij}(x,t) = \int_0^t dz_s D_iD_j\alpha(x,t,s), \qquad \gamma_{ijk}(x,t) = \int_0^t dz_s D_iD_jD_k\alpha(x,t,s).$$

Let $k(x,t) = \int_0^t dz_s \rho(x,t,s)$. Then,

$$Ek^2(x,t) = \int_0^t dt \rho^2(x,t,s),$$

(22) $Ek^{2n}(x,t) = K_n[Ek^2(x,t)]^n$ for some real K_n,

$$E[k(x,t') - k(x,t)]^2 = \int_t^{t'} ds \rho^2(x,t',s) + \int_0^t ds[\rho(x,t',s) - \rho(x,t,s)]^2.$$

Note also that $\gamma_i(x,t)$ is the mean square derivative of $\gamma_0(x,t)$ with respect to the ith coordinate of x in $D \times (0,T]$, and $\gamma_{ijk}(x,t)$ is the mean square derivative of $\gamma_{jk}(x,t)$ with respect to the ith coordinate of x in $D \times (0,T]$.

Then, by the estimates (22), (B7) and Lemma 2, there is a version of $\gamma_0(x,t)$ which (w.p.1) is continuous on $\bar{R}$; it has continuous derivatives $D_i\gamma_0(x,t) = (D_i\gamma_0(x,t)) = \gamma_i(x,t)$ on $\bar{R}$ and continuous second derivatives $D_iD_j\gamma_0(x,t) = \gamma_{ij}(x,t) = (D_iD_j\gamma_0(x,t))$ in compact subsets of $D \times [0,T]$. Furthermore, for $(x,t) \in \partial D \times (0,T]$, $\partial/\partial V(x) = \sum \varphi_i(x) D_i$, where the φ_i are Hölder continuous. Hence, the function

$$\frac{\partial}{\partial V(x)} \int_0^t dz_s \alpha(x,t,s) \equiv \gamma_V(x,t)$$

also has a continuous w.p.1 version on $\partial D \times (0,T]$ and, in fact, can be identified with $\int_0^t dz_s[\partial\alpha(x,t,s)/\partial V(x)]$. Next $D_t\alpha(x,t,s) = \mathscr{L}\alpha(x,t,s)$, $s < t$, $x \notin \partial D$, and, by (B7), $\int_0^t (D_t\alpha(x,t,s))^2 \, ds \leqq K < \infty$ on $\bar{R}$. Also $\alpha(x,t,s)$ is continuous on $\bar{R}$ and tends to $\sigma(x,t)$ as $s \uparrow t$. Hence

$$d\int_0^t \alpha(x,t,s)\, dz_s = \sigma(x,t)\, dz_t + \left[\int_0^t \alpha_t(x,t,s)\, dz_s\right] dt$$

$$= \sigma(x,t)\, dz_t + \mathscr{L}\int_0^t \alpha(x,t,s)\, dz_s\, dt.$$

From what has been said, the function $F(x,t)$ defined by

$$F(x,t) = \int_D \frac{\partial \Gamma(x,\xi;t,0)}{\partial V(x)} U(\xi,0)\,d\xi - \int_0^t d\tau \int \frac{\partial \Gamma(x,\xi;t,\tau)}{\partial V(x)} f(\xi,\tau)\,d\xi$$
$$+ \beta(x,t)\int_D \Gamma(x,\xi;t,0)U(\xi;0)\,d\xi$$
$$- \beta(x,t)\int_0^t d\tau \int_D \Gamma(x,\xi;t,\tau)f(\xi,\tau)\,d\tau - \beta(x,t)\gamma_0(x,t)$$
$$- g(x,t) - v(x,t)r(t)$$

is continuous and uniformly bounded w.p.1 on $\partial D \times (0,T]$ (see Friedman [6, p. 145], where continuity is shown for a similar deterministic problem). Then, there is a continuous (and uniformly bounded w.p.1) solution on $\partial D \times [0,T]$ to the equation (see Friedman [6, (3.6), p. 145])

$$\varphi(x,t) = 2\int_0^t d\tau \int_{\partial D \times [0,T]} \left[\frac{\partial \Gamma(x,\xi;t,\tau)}{\partial V(x)} + \beta(x,t)\Gamma(x,\xi;t,\tau)\right]$$
$$\cdot \varphi(\xi,\tau)\,dS_\xi + 2F(x,t),$$

where dS_ξ is the differential surface measure on ∂D. Finally (see [6, Theorem 2, p. 144 and Corollary 2, p. 147]), it is evident that the function

$$U(x,t) = \int_0^t d\tau \int_{\partial D} \Gamma(x,\xi;t,\tau)\varphi(\xi,\tau)\,dS_\xi + \int_D \Gamma(x,\xi;t,0)U(\xi,0)\,d\xi$$
$$- \gamma_0(x,t) - \int_0^t d\tau \int \Gamma(x,\xi;t,\tau)f(\xi,\tau)\,d\xi$$

has the properties required. In particular, $F(x,t)$ is a nonanticipative functional of the z_t and $\tilde{z}_t$ processes, which implies that $\varphi(x,t)$ and, in turn, $U(x,t)$, are also nonanticipative. This completes the proof.

Now, redefine μ_1 to be the measure determined by $U(x,s)$, $s \leqq t$, and dy_s given by (B6) for $s \leqq t$, and $dr(s)$, $s \leqq t$, given by (B5). Let μ_0 be the measure determined by $U(x,s)$, $r(s)$, $s \leqq t$, and $w(s)$, $s \leqq t$.

Let $R(t)$ denote the vector $E_1^t r(t)$, $P_R(t)$ denote the covariance matrix $E_1^t(r(t) - E_1^t r(t))(r(t) - E_1^t r(t))'$ and $P_{MR}(x,t)$ denote the covariance $E_1^t(U(x,t) - E_1^t U(x,t))(r(t) - E_1^t r(t))$.

THEOREM 2. *Assume* (B1)–(B8). *Then there is a version of* $M(x,t)$ *such that w.p.*1 : $M(x,t)$ *and its first mean square (or true) derivative are continuous w.p.*1 *on* $\bar{R}$ *and* $\bar{D} \times (0,T]$, *respectively. The second mean square (or true) derivatives of* $M(x,t)$ *are continuous in* $D \times (0,T]$ *and* $M(x,t)$ *has an Itô differential which satisfies*

$$dM(x,t) = [\mathscr{L}M(x,t) - f(x,t)]dt$$
$$+ \left[dy - \int_{\partial D} H(\xi,t)M(\xi,t)\,dS_\xi\right]' \Sigma_t^{-1} \left[\int_{\partial D} H(\xi,t)P(\xi,x,t)\,dS_\xi\right]. \tag{23a}$$

Also

$$\frac{\partial M(x,t)}{\partial V(x)} + \beta(x,t)M(x,t) = g(x,t) + R(x,t), \tag{23b}$$

$$\begin{aligned} dR(t) &= AR(t)\,dt \\ &\quad + \left[dy - \int_{\partial D} H(\xi,t)M(\xi,t)\,dS_\xi \right]' \Sigma_t^{-1} \left[\int_{\partial D} H(\xi,t)P_{MR}(\xi,t)\,dS_\xi \right], \end{aligned} \tag{24}$$

$$\begin{aligned} \dot{P}(x,y,t) &= (\mathscr{L}_x + \mathscr{L}_y)P(x,y,t) + \sigma(x,t)\sigma(y,t) \\ &\quad - \left[\int_{\partial D} H(\xi,t)P(x,\xi,t)\,dS_\xi \right]' \Sigma_t^{-1} \left[\int_{\partial D} H(\xi,t)P(\xi,y,t)\,dS_\xi \right], \end{aligned} \tag{25}$$

$$\begin{aligned} \dot{P}_R(t) &= A'P_R(t) + P_R(t)A + G(t)G'(t) \\ &\quad - \left[\int_{\partial D} H(\xi,t)P_{MR}(\xi,t)\,dS_\xi \right]' \Sigma_t^{-1} \left[\int_{\partial D} H(\xi,t)P_{MR}(\xi,t)\,dS_\xi \right], \end{aligned}$$

$$\begin{aligned} \dot{P}_{MR}(x,t) &= \mathscr{L}P_{MR}(x,t) + AP_{MR}(x,t) \\ &\quad - \left[\int_{\partial D} H(\xi,t)P(x,\xi,t)\,dS_\xi \right]' \Sigma_t^{-1} \left[\int_{\partial D} H(\xi,t)P_{MR}(\xi,t)\,dS_\xi \right], \end{aligned}$$

$P(x, y, t)$ satisfies the boundary conditions for (x, t) on $\partial D \times (0, T]$,

$$\frac{\partial P_V(x,y,t)}{\partial V(x)} + \beta(x,t)P(x,y,t) = v(x,t)P_{MR}(x,t),$$

and $P_{MR}(x, t)$ satisfies the (vector) boundary conditions

$$\frac{\partial P_{MR}(x,t)}{\partial V(x)} + \beta(x,t)P_{MR}(x,t) = v(x,t)P_R(t).$$

Proof. The details are very similar to those of Theorem 1 and Lemma 4 and are omitted. Only the boundary conditions will be discussed. By Lemma 5, and a result similar to that of Lemma 4, it is easy to show that there is a version of $M(x, t)$ so that (w.p.1) $M(x, t)$ is continuous on $\bar{R}$ (and in quadratic mean) $(D_i M(x, t)) = D_i M(x, t)$ is continuous on $\bar{D} \times (0, T]$ (and in quadratic mean). Similarly, for $x \in \partial D$, it can be shown that $\partial M(y, t)/\partial V(x)$ and $\partial U(y, t)/\partial V(x)$ are continuous in quadratic mean on $\bar{D} \times \partial D \times (0, T]$ (as functions of (x, y, t)). Then $E_1^t(\partial U(y, t)/\partial V(x)) = \partial M(y, t)/\partial V(x)$, where the last term is defined on ∂D by $\lim_{y \to x, y \in D} \partial M(y, t)/\partial V(x) \equiv \partial M(x, t)/\partial V(x)$. Also $\lim_{y \to x,\, y \in D} \partial U(y, t)/\partial V(x)$ satisfies (w.p.1)

$$\begin{aligned} E_1^t \left[\frac{\partial U(x,t)}{\partial V(x)} + \beta(x,t)U(x,t) - v(x,t)r(t) - g(x,t) \right] \\ = \frac{\partial M(x,t)}{\partial V(x)} + \beta(x,t)M(x,t) - v(x,t)R(t) - g(x,t). \end{aligned}$$

The equation

$$E_1^t[U(y,t) - M(y,t)]\left[\frac{\partial U(x,t)}{\partial V(x)} + \beta(x,t)U(x,t) - v(x,t)r(t) - g(x,t)\right] = 0$$

implies

$$\frac{\partial P(x,y,t)}{\partial V(x)} + \beta(x,t)P(x,y,t) - v(x,t)P_{RM}(y,t) = 0.$$

Also

$$E_1^t[r(t) - R(t)]\left[\frac{\partial U(x,t)}{\partial V(x)} + \beta(x,t)U(x,t) - v(x,t)r(t) - g(x,t)\right] = 0$$

implies

$$\frac{\partial P_{MR}(x,t)}{\partial V(x)} + \beta(x,t)P_{MR}(x,t) - v(x,t)P_R(t) = 0.$$

This completes the details of the proof.

REFERENCES

[1] H. J. KUSHNER, *An application of the Sobolev imbedding theorems to criteria for the continuity of processes with a vector parameter*, Ann. Math. Statist., 40 (1969), pp. 517–526.

[2] ———, *On the optimal control of a system governed by a linear parabolic equation with white noise inputs*, this Journal, 6 (1968), pp. 596–614.

[3] ———, *Dynamical equations for nonlinear filtering*, J. Differential Equations, 3 (1967), pp. 179–190.

[4] M. ZAKAI, *On the optimal filtering of diffusion processes*, Z. Wahrscheinlichkeitstheorie, 11 (1969), pp. 230–243.

[5] M. LOEVE, *Probability Theory*, 3rd ed., Van Nostrand, Princeton, 1963.

[6] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, N.J., 1964.

SIAM J. CONTROL
Vol. 8, No. 3, August 1970

CONTROL PROBLEMS WITH FUNCTIONAL RESTRICTIONS*

J. WARGA†

1. Introduction. We shall study a class of control problems in which restrictions are imposed on functions of controls with values in real topological vector spaces. This class includes, among others, unilateral [1], [2], [3], [4] and minimax [5] problems of the calculus of variations, the more general unilateral and minimax versions of control problems defined by integral [6] or functional-differential [7] equations, and certain pursuit and evasion games. We shall discuss some of these applications elsewhere.

Certain abstract models of control problems with restrictions in real topological vector spaces have been considered by L. W. Neustadt [8], [4] and R. V. Gamrelidze and G. L. Kharatishvili [9] who have derived necessary conditions for minimum. Neustadt has applied his results to several problems, mostly defined by ordinary differential equations [3], [4]. These authors' respective assumptions and necessary conditions (see, e.g., [4, Condition 3.1, p. 63 and Theorem 3.1, p. 64] and [9, §§ 4 and 5, p. 244]) are expressed in terms of certain mappings that must be appropriately defined for particular problems. Our approach is to study both existence and necessary conditions for our model in terms of certain continuity and differentiability properties of the functions of controls that define the problem. To do so, we imbed, as in [11], the set $\mathscr{R}$ of "measurable original controls" in a larger set $\mathscr{S}$ of "measurable relaxed controls." The reason for this is twofold: first, because, as it is well known (especially in the control theory of ordinary differential equations), the existence of an "original minimizing point" can be rarely guaranteed whereas "relaxed minimizing points" do exist in rather general situations and can be suitably "approximated" by measurable original controls [11]; secondly, because we shall derive necessary conditions for minimum that apply both to "original" and to "relaxed" minimizing points and that can be formulated in a natural manner in $\mathscr{S}$ but would be rather artificial in $\mathscr{R}$ alone.

Relaxed controls [11] represent an extension of the concept of generalized curves introduced in 1937 by L. C. Young [12] (and which established the basis for the existence theory in variational problems) and of the related concept of relaxed controls on an interval (Warga [13], McShane [14], Ghouila–Houri [15], Nishiura [16]). Lemma 3.1 below is an application of the basic separation theorem for convex sets which was used in a similar context by Neustadt [4], [8]. Lemma 3.2 supplies, for all problems defined by functions of controls with the required continuity and differentiability properties, a canonical mapping of the kind postulated by Neustadt in [4, Condition 3.1, p. 63]. The fixed-point argument of Lemma 3.3 is analogous to prior arguments of this kind (as by Pontryagin et al. in [10]), and the convex set $\{Dx(\bar{q}; q - \bar{q}) | q \in Q\}$ is patterned after McShane's [17] "convex set of variations" which has been a basic variational tool for thirty years.

* Received by the editors May 12, 1969, and in revised form October 21, 1969.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115. This research was supported by the National Aeronautics and Space Administration under Grant NGR 22-011-020.

2. Existence and necessary conditions. Before discussing our problem in complete generality, we shall briefly illustrate by an example how our model applies to a particular control problem. Let us consider the differential equation

$$(\dot{y}^1(t), \cdots, \dot{y}^n(t)) = \dot{y}(t) = g(t, y(t), \rho(t), p) \quad \text{a.e. in } T = [t_0, t_1],$$

where p belongs to some given metric set P with a regular probability measure μ defined on P, ρ can be chosen from the class $\mathscr{R}$ of measurable functions from T to some subset $\tilde{R}$ of the Euclidean k-space E_k, $g: T \times E_n \times E_k \times P \to E_n$, and the initial value $y(t_0)$ is restricted to some given set B. Assume that, for every $\rho \in \mathscr{R}$, $b \in B$ and $p \in P$, there exists a unique absolutely continuous solution $t \to y(\rho, b, p)(t)$ of the differential equation such that $y(\rho, b, p)(t_0) = b$ and the function $p \to y(\rho, b, p)(t_1): P \to E_n$ is continuous. Let A be a given subset of E_l, $h_0: E_n \to E_1$, $h_1: E_n \to E_m$, $h_2: E_n \to E_l$ and let, henceforth, the origin of any vector space that we consider be represented by 0 if the nature of the space is clear from the context. We wish to determine a control (function) $\bar{\rho} \in \mathscr{R}$ and a control (parameter) $\bar{b} \in B$ that minimize $\int_P h_0(y(\rho, b, p)(t_1))\mu(dp)$ on the set

$$\left\{(\rho, b) \in \mathscr{R} \times B \,\middle|\, \int_P h_1(y(\rho, b, p)(t_1))\mu(dp) = 0, h_2(y(\rho, b, P)(t_1)) \subset A\right\}.$$

(Such a problem would arise if we wished to minimize the expected value (with respect to μ) of a cost functional with restrictions placed on variances or higher moments and with the range of a controlled function restricted to A.)

We shall now reformulate this particular problem in more abstract terms. Let V denote the Banach space of continuous functions from P to E_l, with the usual norm, and let $C = \{v(\cdot) \in V | v(P) \subset A\}$. For $(\rho, b) \in \mathscr{R} \times B$, we set

$$x_0(\rho, b) = \int_P h_0(y(\rho, b, p)(t_1))\mu(dp) \in E_1,$$

$$x_1(\rho, b) = \int_P h_1(y(\rho, b, p)(t_1))\mu(dp) \in E_m$$

and

$$x_2(\rho, b)(\cdot) = h_2(y(\rho, b, \cdot)(t_1)) \in V.$$

We are thus given a function $x = (x_0, x_1, x_2): \mathscr{R} \times B \to E_1 \times E_m \times V$ and a subset C of V, and we wish to minimize x_0 on the set $\{(\rho, b) \in \mathscr{R} \times B | x_1(\rho, b) = 0, x_2(\rho, b) \in C\}$. (The functions x_0, x_1 and x_2 are the "functions of controls" referred to in § 1.)

In other control problems that we have in mind, the set T may be multi-dimensional, the function x may be determined by other functional relations (say by integral or partial differential equations), the set P may consist of "adverse" controls, or coincide with T, etc.

We now turn to the general problem. Let T and R be compact metric spaces, with a positive, finite, regular, complete, and nonatomic measure dt defined on T, $\tilde{R}$ a dense subset of R, m a nonnegative integer, $m' = \max(2, m + 1)$, V a real

topological vector space, C a subset of V, and B a subset of some vector space with a topology whose relativization to all m'- or lower-dimensional subsets is Euclidean. We denote by $\mathscr{R}$ the class of measurable functions ρ from T to $\tilde{R}$, and assume given a function $x = (x_0, x_1, x_2): \mathscr{R} \times B \to E_1 \times E_m \times V$. We wish to study certain properties of an "original minimizing point" $(\bar{\rho}, \bar{b}) \in \mathscr{R} \times B$ that minimizes x_0 on the set $\{(\rho, b) \in \mathscr{R} \times B | x_1(\rho, b) = 0, x_2(\rho, b) \in C\}$.

For reasons that were mentioned in the Introduction, we imbed $\mathscr{R}$ in a set $\mathscr{S}$ of "measurable relaxed controls." We denote by S the set of regular Borel probability measures on R, and let $\mathscr{S}$ be the class of functions $\sigma: T \to S$ with the property that $t \to \int_R c(r)\sigma(dr; t)$ is measurable for every continuous $c: R \to E_1$. We define $\mathscr{R}$ as a subset of $\mathscr{S}$ by identifying $\rho: T \to \tilde{R}$ with $\sigma_\rho: T \to S$ such that $\sigma_\rho(t)$ is a measure concentrated at $\rho(t)$ with probability 1 for almost all $t \in T$. (Observe that if ρ is measurable, then $\sigma_\rho \in \mathscr{S}$, and conversely.)

Next we define a topology in $\mathscr{S}$. To do so, we consider the Banach space $\mathscr{B}$ of functions $(t, r) \to \phi(t, r): T \times R \to E_1$, measurable in t for each r, continuous in r for each t, and with the norm $|\phi| = \int_T \sup_{r \in R} |\phi(t, r)|\, dt < \infty$. (It follows easily from known theorems that $\mathscr{B}$ is isomorphic to the Banach space $L^1(T, C(R))$ of L^1 functions from T to the space of continuous scalar functions on R [18, Theorem 11, p. 149 and Theorem 22, p. 117]). We identify every $\sigma \in \mathscr{S}$ with the element $l_\sigma \in \mathscr{B}^*$ (the topological dual of $\mathscr{B}$) defined by

$$l_\sigma(\phi) = \langle \sigma, \phi \rangle = \int_T dt \int_R \phi(t, r)\sigma(dr; t) \qquad (\phi \in \mathscr{B}).$$

We then choose for $\mathscr{B}^*$ the weak star topology in $\mathscr{B}^*$ (the $\mathscr{B}$ topology of $\mathscr{B}^*$) and for its subset $\mathscr{S}$, the corresponding relative topology. We have shown [11, Theorem 2.4, p. 631 and Theorem 2.5, p. 632] that $\mathscr{S}$ is closed and $\mathscr{R}$ is dense[1] in $\mathscr{S}$, and it follows from known theorems [18, Theorem 2, p. 424 and Theorem 1, p. 426] that $\mathscr{S}$ is metric and compact.

We shall assume henceforth that the definition of x has been extended from $\mathscr{R} \times B$ to $\mathscr{S} \times B$, and we choose the product topology for $\mathscr{S} \times B$. The extension of x to $\mathscr{S} \times B$ is fairly simply accomplished for control problems defined by functional equations; e.g., in the illustrative problem described at the beginning of this section, $y(\sigma, b, p)(\cdot)$ is the solution of the equation

$$y(t) = b + \int_{t_0}^{t} d\tau \int_R g(\tau, y(\tau), r, p)\sigma(dr; \tau) \qquad (t_0 \leqq t \leqq t_1).$$

We say that $(\bar{\sigma}, \bar{b}) \in \mathscr{S} \times B$ is a *relaxed minimizing point* if it minimizes x_0 on the set $\{(\sigma, b) \in \mathscr{S} \times B | x_1(\sigma, b) = 0, x_2(\sigma, b) \in C\}$. The following existence and approximation theorem is an immediate consequence of the fact that $\mathscr{S}$ is closed and sequentially compact and $\mathscr{R}$ is dense in $\mathscr{S}$.

THEOREM 2.1. *Let B be closed and sequentially compact and C closed, and assume that there exists a point $(\sigma', b') \in \mathscr{S} \times B$ such that $x_1(\sigma', b') = 0$ and*

[1] The proof, in [11], that $\mathscr{R}$ is dense in $\mathscr{S}$ presupposes that $\tilde{R} = R$ but it remains valid if $\tilde{R}$ is dense in R. Furthermore, the statement that $\mathscr{R}$ is dense in $\mathscr{S}$ also follows from Lemma 3.2 below.

$x_2(\sigma', b') \in C$ *and that* x *is continuous in some neighborhood of the set*

$$\{(\sigma, b) \in \mathscr{S} \times B \mid x_1(\sigma, b) = 0, x_2(\sigma, b) \in C, x_0(\sigma, b) \leqq x_0(\sigma', b')\}.$$

Then there exists a relaxed minimizing point $(\bar{\sigma}, \bar{b})$ *and a sequence* $\{\rho_i\}_{i=1}^{\infty}$ *in* $\mathscr{R}$ *such that* $\lim_{i\to\infty} \rho_i = \bar{\sigma}$ *and* $\lim_{i\to\infty} x(\rho_i, \bar{b}) = x(\bar{\sigma}, \bar{b})$.

Remark. We observe that a relaxed minimizing point, as an element of $\mathscr{S} \times B$, may, but need not, belong to $\mathscr{R} \times B$. If it does, then it is also an original minimizing point. On the other hand, examples can be given [13, p. 118] of problems that admit different original and relaxed minimizing points.

We shall now state certain necessary conditions for minimum that are satisfied by both original and relaxed minimizing points and that we shall prove in § 4. These conditions generalize the Weierstrass E-condition (the maximum principle) and the transversality conditions.

We set

$$\mathscr{T}_n = \left\{\theta = (\theta^1, \cdots, \theta^n) \mid \theta^j \geqq 0, \sum_{j=1}^{n} \theta^j \leqq 1\right\} \subset E_n \quad (n = 1, 2, \cdots),$$

$$Q = \mathscr{S} \times B, \qquad q = (\sigma, b)$$

and

$$Dx(\bar{q}; q - \bar{q}) = \lim_{\alpha \to +0} \frac{1}{\alpha}(x(\bar{q} + \alpha(q - \bar{q})) - x(\bar{q})).$$

We denote by $|\cdot|$ the norm in a normed linear space, and say that the function

$$\theta \to x\left(\bar{q} + \sum_{j=1}^{n} \theta^j(q_j - \bar{q})\right) : \mathscr{T}_n \to E_1 \times E_m \times V$$

(where $q_j \in Q$) has a derivative at 0 if

$$\frac{1}{|\theta|}\left(x\left(\bar{q} + \sum_{j=1}^{n} \theta^j(q_j - \bar{q})\right) - x(\bar{q}) - \sum_{j=1}^{n} \theta^j Dx(\bar{q}; q_j - \bar{q})\right)$$

exists and converges to 0 in $E_1 \times E_m \times V$ as $|\theta| \to 0$, $\theta \in \mathscr{T}_n$. We similarly define the existence of a derivative of x_1, x_2 or $x_{1,2} = (x_1, x_2)$.

THEOREM 2.2. *Let* B *be a convex set and* C *a convex body. Let, furthermore,* $\bar{q} = (\bar{\sigma}, \bar{b})$ *be either a relaxed or an original minimizing point and assume that for every subset* $\{b_1, \cdots, b_{m'}\}$ *of* B, *the function*

$$(\sigma, \theta) \to x\left(\sigma, \bar{b} + \sum_{j=1}^{m'} \theta^j(b_j - \bar{b})\right) : \mathscr{S} \times \mathscr{T}_{m'} \to E_1 \times E_m \times V$$

is continuous and, for every subset $\{q_1, \cdots, q_{m'}\}$ *of* Q, *the function*

$$\theta \to x\left(\bar{q} + \sum_{j=1}^{m'} \theta^j(q_j - \bar{q})\right) : \mathscr{T}_{m'} \to E_1 \times E_m \times V$$

has a derivative at 0. *Then there exists a nonvanishing continuous linear functional* l *on* $E_1 \times E_m \times V$ *such that*

$$l((v_0, v_1, v_2)) = \lambda_0 v_0 + \lambda_1 v_1 + l_2(v_2) \quad \text{for } v_0 \in E_1, v_1 \in E_m \quad \text{and} \quad v_2 \in V, \quad \lambda_0 \geqq 0,$$

$$l(Dx(\bar{q}; q - \bar{q})) \geqq 0 \quad \text{for all } q \in Q,$$

and

$$l_2(v_2) \leqq l_2(x_2(\bar{q})) \quad \textit{for all } v_2 \in C.$$

Remark. For purposes of exposition and to emphasize the main ideas, we have stated the above theorem, and its proof, in a somewhat simplified form. We can, however, strengthen this theorem in several ways, and we shall list a few of these. First of all, if $\mathscr{R}_s$ is the class of simple measurable functions from T to $\tilde{R}$ (having only a finite number of values) and the point $\bar{q} = (\bar{\sigma}, \bar{b})$, while not necessarily an original or relaxed minimizing point, minimizes x_0 on the set $\{(\rho, b) \in \mathscr{R}_s \times B | x_1(\rho, b) = 0, x_2(\rho, b) \in C\}$, then the conclusion of Theorem 2.2, and its proof, remain valid. Secondly, a modification of Lemma 3.2 along the lines of [11, Lemma 4.4, p. 634] permits one to prove a generalization of Theorem 2.2 in which $\mathscr{R}$ and $\mathscr{S}$ are replaced, respectively, by

$$\mathscr{R}^{\#} = \{\rho \in \mathscr{R} | \rho(t) \in R^{\#}(t) \quad (t \in T)\}$$

and

$$\mathscr{S}^{\#} = \{\sigma \in \mathscr{S} | \sigma(\bar{R}^{\#}(t); t) = 1 \quad (t \in T)\},$$

where $R^{\#}(\cdot)$ is a given mapping from T to the class of nonempty subsets of R that satisfies certain conditions (see [11, Assumption 2.3, p. 631]) and $\bar{R}^{\#}(t)$ is the closure of $R^{\#}(t)$. Finally, a simple reinterpretation of Theorem 2.2 permits one to conclude that this theorem remains valid in the case where $\bar{q}$ is a relaxed minimizing point if the assumption that x is continuous in some neighborhood of $\bar{q}$ in Q is replaced by the weaker assumption that, for every subset $\{q_1, \cdots, q_m\}$ of Q, the function $\theta \to x(\bar{q} + \sum_{j=1}^{m'} \theta^j(q_j - \bar{q})) : \mathscr{T}_{m'} \to E_1 \times E_m \times V$ is continuous in some neighborhood of 0 in $\mathscr{T}_{m'}$. These, and similar, remarks are also applicable to Theorem 2.3 below.

The following theorem is essentially a refinement of Theorem 2.2. It applies to certain "unorthodox" problems such as the one considered by Neustadt [4, pp. 87–91] who has shown, in [8], [3] and [4], that in certain classes of problems linear approximations can be usefully replaced by convex approximations.

For $(v_0, v_1, v_2) \in E_1 \times E_m \times V$, we write $v_{0,1} = (v_0, v_1)$ and $v_{1,2} = (v_1, v_2)$.

THEOREM 2.3. *Let the assumptions be the same as in Theorem* 2.2 *except that the condition*

$$\theta \to x\left(\bar{q} + \sum_{j=1}^{m'} \theta^j(q_j - \bar{q})\right) : \mathscr{T}_{m'} \to E_1 \times E_m \times V \tag{2.3.1}$$

has a derivative at 0 *for every subset* $\{q_1, \cdots, q_{m'}\}$ *of* Q

is replaced by

$$\theta \to x_{1,2}\left(\bar{q} + \sum_{j=1}^{m'} \theta^j(q_j - \bar{q})\right) : \mathscr{T}_{m'} \to E_m \times V \tag{2.3.1'}$$

has a derivative at 0 *for every subset* $\{q_1, \cdots, q_{m'}\}$ *of* Q *and* $x_0(q) = g(x_{1,2}(q))(q \in Q)$, *where* g *is a continuous covex function from* $E_m \times V$ *to* E_1.

Let $\hat{g}(v_{1,2}) = g(x_{1,2}(\bar{q}) + v_{1,2}) - g(x_{1,2}(\bar{q}))$ *for* $v_{1,2} \in E_m \times V$. *Then there exists a nonvanishing continuous linear functional* l *on* $E_1 \times E_m \times V$ *such that*

$$l((v_0, v_{1,2})) = \lambda_0 v_0 + \lambda_1 v_1 + l_2(v_2) = \lambda_0 v_0 + l_{1,2}(v_{1,2})$$

for $v_0 \in E_1, v_1 \in E_m$ *and* $v_2 \in V$, $\lambda_0 \geqq 0$,

$$\lambda_0 \hat{g}(Dx_{1,2}(\bar{q}; q - \bar{q})) + l_{1,2}(Dx_{1,2}(\bar{q}: q - \bar{q})) \geqq 0 \quad \textit{for all } q \in Q$$

and

$$l_2(v_2) \leqq l_2(x_2(\bar{q})) \quad \textit{for all } v_2 \in C.$$

3. Auxiliary lemmas. We shall denote by $|A|$ the measure of $A \subset T$ and by H°, ∂H and $\bar{H}$ the interior, the boundary, and the closure, respectively, of a set H. If A is a subset of some E_l, with 0 as a limit point, $\mathcal{Y}$ a topological vector space, and $h: A \to \mathcal{Y}$, then we say that $h(a) = o(|a|)$ if $\lim_{|a| \to 0} h(a)/|a| = 0$.

We observe that $\mathcal{S}$ is a convex subset of $\mathcal{B}^*$ and that, for $\mathcal{T}_n$ defined as in Theorems 2.1 and 2.2 and $\sigma_j \in \mathcal{S}$ $(j = 0, 1, \cdots, n)$, the function $\theta \to \sigma_0 + \sum_{j=1}^{n} \theta^j \cdot (\sigma_j - \sigma_0): \mathcal{T}_n \to \mathcal{S}$ is continuous. Furthermore, if $\sigma_i \in \mathcal{S}$ $(i = 0, 1, 2, \cdots)$ and $\lim_{i \to \infty} |\{t \in T | \sigma_i(t) \neq \sigma_0(t)\}| = 0$, then $\lim_{i \to \infty} \sigma_i = \sigma_0$ (but not conversely).

Finally, we note that if $\theta \to x(\bar{q} + \sum_{j=1}^{m'} \theta^j (q_j - \bar{q})): \mathcal{T}_{m'} \to E_1 \times E_m \times V$ has a derivative at 0 for all $\{q_1, \cdots, q_{m'}\} \subset Q$, then $Dx(\bar{q}; q - \bar{q})$ exists for all $q \in Q$ and $Dx(\bar{q}; \alpha_1 q_1 + \alpha_2 q_2 - \bar{q}) = \alpha_1 Dx(\bar{q}; q_1 - \bar{q}) + \alpha_2 Dx(\bar{q}; q_2 - \bar{q})$ for all $q_1, q_2 \in Q$ provided $\alpha_1 \geqq 0, \alpha_2 \geqq 0, \alpha_1 + \alpha_2 = 1$. Then the set $\{Dx(\bar{q}; q - \bar{q}) | q \in Q\}$ is convex whenever B is convex. Similar remarks apply when x is replaced by $x_{1,2}$.

LEMMA 3.1. *Let W be a convex subset of $E_1 \times E_m \times V$ containing $(0, 0, 0)$ and $(1, 0, 0)$, C' a nonempty, open and convex subset of V, and $0 \in \bar{C}'$. Then either there exist $\lambda_0 \geqq 0$, $\lambda_1 \in E_m$ and $l_2 \in V^*$, not all 0, and such that*

$$\lambda_0 \cdot w_0 + \lambda_1 \cdot w_1 + l_2(w_2) \geqq 0 \qquad \textit{for all } (w_0, w_1, w_2) \in W$$

and

$$l_2(c) \leqq 0 \quad \textit{for all } c \in C',$$

or there exist points $(\xi_0^i, \xi_1^i, \xi_2^i) = \xi^i \in W$ and numbers $\beta^i > 0$ $(i = 0, \cdots, m)$, such that $\sum_{i=0}^{m} \beta^i = 1$, $\xi_2^i \in C'$, the set $\{\xi_{0,1}^0, \cdots, \xi_{0,1}^m\}$, where $\xi_{0,1}^i = (\xi_0^i, \xi_1^i)$, is linearly independent, $\xi_0^i < 0$ and $\sum_{i=0}^{m} \beta^i \xi_1^i = 0$.

Proof. Let $W_{0,1} = \{\xi_{0,1} \in E_1 \times E_m | \xi = (\xi_0, \xi_1, \xi_2) \in W\}$. Then either

(i) $0 \in \partial W_{0,1}$, or

(ii) $0 \in W_{0,1}^\circ$ and there exists some $\bar{\xi} = (\bar{\xi}_0, \bar{\xi}_1, \bar{\xi}_2) \in W$ such that $\bar{\xi}_0 < 0$, $\bar{\xi}_1 = 0$ and $\bar{\xi}_2 \in C'$, or

(iii) $0 \in W_{0,1}^\circ$ and every $\xi = (\xi_0, \xi_1, \xi_2)$ with $\xi_0 < 0$, $\xi_1 = 0$ and $\xi_2 \in C'$ is outside W.

If (i) holds, then the first alternative of the lemma is satisfied with $l_2 = 0$ and (λ_0, λ_1) an inward normal to $W_{0,1}$ at 0 ($\lambda_0 \geqq 0$ because $(1, 0) \in W_{0,1}$).

If (ii) holds (hence, $0 \in W_{0,1}^\circ$), then there exist $\eta^i \in W$ and $\beta^i > 0$ $(i = 0, \cdots, m)$, such that $\sum_{i=0}^{m} \beta^i = 1$, $\eta_0^i < 0$, $\{\eta_{0,1}^0, \cdots, \eta_{0,1}^m\}$ is linearly independent and $\sum_{i=0}^{m} \beta^i \eta_1^i = 0$. Since $\bar{\xi} \in W$ and $\bar{\xi}_2 \in C'$, there exists some $\theta \in (0, 1]$ such that $\xi^i = \theta \eta^i + (1 - \theta)\bar{\xi} \in W (i = 0, \cdots, m), \xi_0^i < 0$, $\sum_{i=0}^{m} \beta^i \xi_1^i = 0$, $\{\xi_{0,1}^0, \cdots, \xi_{0,1}^m\}$ is independent, and $\xi_2^i \in C'$.

Finally, we consider the case (iii). Let $W_2' = \{w_2 | w = (w_0, w_1, w_2) \in W, w_0 < 0, w_1 = 0\}$. Then W_2' is nonempty and convex in V and $W_2' \cap C'$ is empty.

Since C' is nonempty, open and convex, there exists a nonzero continuous linear functional l'_2 and a real α such that

$$l'_2(w_2) \leqq \alpha \leqq l'_2(c) \quad \text{for all } w_2 \in W'_2 \quad \text{and} \quad c \in C'. \tag{3.1.1}$$

Since $0 \in \bar{C}' \cap \overline{W}'_2$, we have $\alpha = 0$.

Now let $W^{\#} = \{(l'_2(w_2), w_0, w_1) | w = (w_0, w_1, w_2) \in W\}$, let 0_m be the origin in E_m, and let $H = \{(\xi'_0, \xi_0, 0_m) | \xi'_0 > 0, \xi_0 < 0\}$. Then, by (3.1.1), $H \cap W^{\#}$ is empty and, since both H and $W^{\#}$ are nonempty convex sets in $E_1 \times E_1 \times E_m$, there exist real λ'_0 and λ_0 and $\lambda_1 \in E_m$ such that $|\lambda'_0| + |\lambda_0| + |\lambda_1| \neq 0$ and

$$\begin{aligned} &\lambda'_0 l'_2(w_2) + \lambda_0 w_0 + \lambda_1 \cdot w_1 \geqq 0 \geqq \lambda'_0 \xi'_0 + \lambda_0 \xi_0 \\ &\qquad \text{for all } w = (w_0, w_1, w_2) \in W, \quad \xi'_0 > 0 \quad \text{and} \quad \xi_0 < 0. \end{aligned} \tag{3.1.2}$$

It follows that $\lambda'_0 \leqq 0$ and $\lambda_0 \geqq 0$. If we set $\lambda'_0 l'_2 = l_2$, then (3.1.1) and (3.1.2) imply the first alternative of the lemma.

LEMMA 3.2. *For every choice of n and $\sigma_j \in \mathscr{S}$ $(j = 0, 1, \cdots, n)$, and for all $\theta \in \mathscr{T}_n$, we can construct a sequence $\{\rho_i(\theta)\}_{i=1}^{\infty}$ in $\mathscr{R}$ that converges to $\tilde{\sigma}(\theta) = \sigma_0 + \sum_{j=1}^{n} \theta^j(\sigma_j - \sigma_0)$ in $\mathscr{S}$ uniformly in $\theta \in \mathscr{T}_n$ and such that $\theta \to \rho_i(\theta) : \mathscr{T}_n \to \mathscr{R}$ is continuous for every i.*

Proof. For every fixed positive integer i we can partition the compact and metric set R into disjoint Borel subsets R^i_k $(k = 1, \cdots, k_i)$ of diameters at most $1/i$. We can similarly partition T (leaving out, if necessary, a subset T_0 of measure 0) into Borel subsets T^i_l $(l = 1, \cdots, l_i)$ of diameters at most $1/i$ and each of positive measure. For each k, we choose a point $r^i_k \in R^i_k \cap \tilde{R}$.

Since the measure of T is nonatomic, for each l we can define subsets $T^i_l(\alpha)$ $(0 \leqq \alpha \leqq 1)$ of T^i_l such that $T^i_l(\alpha) \subset T^i_l(\beta)$ for $\alpha \leqq \beta$, $|T^i_l(\alpha)| = \alpha |T^i_l|$, $T^i_l(0)$ is empty and $T^i_l(1) = T^i_l$. Now let $\alpha^i_{l,j,k} = (1/|T^i_l|) \int_{T^i_l} \sigma_j(R^i_k; t)\, dt$ $(l = 1, \cdots, l_i, k = 1, \cdots, k_i,$ $j = 0, \cdots, n)$, and, for each $\theta \in \mathscr{T}_n$, let $\theta^0 = 1 - \sum_{j=1}^{n} \theta^j$. For each $\theta \in \mathscr{T}_n$ and each l, we can partition the interval $[0, 1]$ into disjoint intervals
$n, k = 1, \cdots, k_1)$ of lengths $\theta^j \alpha^i_{l,j,k}$, arranged in the lexicographic order of (j, k). Let $a^i_{l,j,k}(\theta)$, $b^i_{l,j,k}(\theta)$ be the endpoints of $I^i_{l,j,k}(\theta)$. We now set, for all $\theta \in \mathscr{T}_n$,

$$\rho_i(\theta)(t) = \begin{cases} r^i_k & \text{for } t \in T^i_{l,j,k}(\theta) \\ \qquad = T^i_l(b^i_{l,j,k}(\theta)) - T^i_l(a^i_{l,j,k}(\theta)) & \text{for all } l, j \text{ and } k, \\ r^1_1 & \text{for } t \in T_0. \end{cases}$$

These relations define $\rho_i(\theta)$ on T for all $\theta \in \mathscr{T}_n$.

We shall show that (i) the sequence $\{\rho_i(\theta)\}_{i=1}^{\infty}$ in $\mathscr{R}$ converges to $\tilde{\sigma}(\theta) = \sigma_0 + \sum_{j=1}^{n} \theta^j(\sigma_j - \sigma_0)$ in $\mathscr{S}$, uniformly in $\theta \in \mathscr{T}_n$, and (ii) for each i, the function $\theta \to \rho_i(\theta) : \mathscr{T}_n \to \mathscr{R}$ is continuous.

Statement (ii) follows from the observation that $|\{t \in T | \rho_i(\theta') \neq \rho_i(\theta'')\}|$ converges to 0 with $|\theta' - \theta''|$ because the $a^i_{l,j,k}$ and $b^i_{l,j,k}$ are continuous (in fact, linear) in θ.

To prove the assertion (i), we observe that the statement

$$\text{“}\lim_{i \to \infty} \sigma_i = \sigma\text{”}$$

is equivalent to the statement

$$\text{“}\lim_{i\to\infty}\int_T dt\int_R \phi(t,r)\sigma_i(dr;t)=\int_T dt\int_R \phi(t,r)\sigma(dr;t)\quad\text{for each } \phi\in\mathscr{B};\text{”}$$

furthermore, since finite sums of functions of the form $(t,r)\to f(t)c(t)$, where f and c are continuous, are dense in $\mathscr{B}$, it suffices to consider $\phi(t,r)=f(t)c(r)$. Thus, for any continuous $f:T\to E_1$ and $c:R\to E_1$, we must prove that

$$\lim_{i\to\infty}\int_T f(t)c(\rho_i(\theta)(t))\,dt=\int_T f(t)\,dt\int_R c(r)\tilde{\sigma}(\theta)(dr;t)$$

uniformly in $\theta\in\mathscr{T}_n$.

For each T_l^i, we choose a point $t_l^i\in T_l^i$. The symbol $O(\varepsilon)$ will represent a quantity whose absolute value does not exceed ε, $|f|_1=\int_T|f(t)|\,dt$ and $|c|_\infty=\sup_{r\in R}|c(r)|$. Let $\varepsilon>0$, and let $i_0=i_0(\varepsilon)$ be sufficiently large so that $|c(r)-c(r')|\leqq\varepsilon/(3|f|_1)$ and $|f(t)-f(t')|\leqq\varepsilon/(3|c|_\infty|T|)$ if distance $(r,r')\leqq 1/i_0$ and distance $(t,t')\leqq 1/i_0$. Then, for every $\theta\in\mathscr{T}_n$ and $i\geqq i_0$, and with summations taken for $j=0,\cdots,n$, $l=1,\cdots,l_i$ and $k=1,\cdots,k_i$, we have

$$\begin{aligned}
\int_T f(t)\,dt\int_R c(r)\tilde{\sigma}(\theta)(dr;t)&=\sum_j\theta^j\int_T f(t)\,dt\int_R c(r)\sigma_j(dr;t)\\
&=\sum_{j,k}\theta^j\int_T f(t)c(r_k^i)\sigma_j(R_k^i;t)\,dt+O(\varepsilon/3)\\
&=\sum_{l,j,k}f(t_l^i)c(r_k^i)\theta^j\int_{T_l^i}\sigma_j(R_k^i;t)\,dt+O(2\varepsilon/3)\\
&=\sum_{l,j,k}f(t_l^i)c(r_k^i)|T_{j,k,l}^i(\theta)|+O(2\varepsilon/3)\\
&=\sum_{l,j,k}f(t_l^i)\int_{T_{j,k,l}^i(\theta)}c(\rho_i(\theta)(t))\,dt+O(2\varepsilon/3)\\
&=\int_T f(t)c(\rho_i(\theta)(t))\,dt+O(\varepsilon).
\end{aligned}$$

Since i_0 was chosen independently of θ, statement (i) is thus verified.

LEMMA 3.3. *Let $C',\beta^i>0$ and $\xi^i\in E_1\times E_m\times V(i=0,\cdots,m)$ be as described in Lemma 3.1, let U be a given neighborhood of 0 in V, and let $\beta=(\beta^0,\cdots,\beta^m)$. For $h_0>0$, $\mathscr{T}'=\{\theta=(\theta^0,\cdots,\theta^m)|0\leqq\theta^i\leqq h_0\,(i=0,\cdots,m)\}$ and $k=1,2,\cdots$, let $\bar{y}=(\bar{y}_0,\bar{y}_1,\bar{y}_2)$, $y^k=(y_0^k,y_1^k,y_2^k)$, $\psi=(\psi_0,\psi_1,\psi_2)$ and $e^k=(e_0^k,e_1^k,e_2^k):\mathscr{T}'\to E_1\times E_m\times V$ be continuous and such that $\lim_{k\to\infty}y^k(\theta)=\bar{y}(\theta)$ and $\lim_{k\to\infty}e^k(\theta)=0$ uniformly in $\theta\in\mathscr{T}'$, $\psi(\theta)=o(|\theta|)$, and*

$$y^k(\theta)-\bar{y}(0)-\sum_{i=0}^m\theta^i\xi^i=\psi(\theta)+e^k(\theta)\qquad(k=1,2,\cdots,\quad\theta\in\mathscr{T}').\tag{3.3.1}$$

Then there exists $\gamma' > 0$ such that, for every $\gamma \in (0, \gamma']$, we can determine $\theta(\gamma) \in \mathscr{T}'$ and an integer $k(\gamma) > 0$ satisfying the following relations:

$$\psi(\theta(\gamma)) = o(\gamma), \quad \theta(\gamma) - \gamma\beta = o(\gamma), \quad \frac{1}{\gamma} e_2^{k(\gamma)}(\theta(\gamma)) \in U,$$

$$y_0^{k(\gamma)}(\theta(\gamma)) < \bar{y}_0(0), \quad y_1^{k(\gamma)}(\theta(\gamma)) = \bar{y}_1(0), \quad y_2^{k(\gamma)}(\theta(\gamma)) - \bar{y}_2(0) \in C'.$$

Proof. Since $\xi_2^i \in C'$ $(i = 0, \cdots, m)$, it follows that

$$\sum_{i=0}^{m} \beta^i \xi_2^i \in C'.$$

Thus there exists a neighborhood U_1 of 0 in V such that $\sum_{i=0}^{m} \beta^i \xi_2^i + U_1 \subset C'$. As in any topological vector space, we can determine a symmetric neighborhood U_2 of 0 in V such that $U_2 + U_2 \subset U_1$. We now set $U' = U \cap U_2$.

Let M be the matrix with columns $\xi_{0,1}^i$ $(i = 0, \cdots, m)$ which is clearly non-singular. Relation (3.3.1) implies, projecting both sides on $E_1 \times E_m$, that

$$y_{0,1}^k(\theta) - \bar{y}_{0,1}(0) = M\theta + \psi_{0,1}(\theta) + e_{0,1}^k(\theta) \qquad (k = 1, 2, \cdots, \quad \theta \in \mathscr{T}'). \tag{3.3.2}$$

We shall show that there exists $\bar{\gamma} > 0$ such that, for every $\gamma \in (0, \bar{\gamma}]$, we can choose an integer $k(\gamma) > 0$ and a point $\theta(\gamma) \in \mathscr{T}'$ satisfying the relations

$$y_{0,1}^{k(\gamma)}(\theta(\gamma)) - \bar{y}_{0,1}(0) = \gamma M\beta, \tag{3.3.3}$$

$$\frac{1}{\gamma} e_2^{k(\gamma)}(\theta(\gamma)) \in U' \tag{3.3.4}$$

and

$$\lim_{\gamma \to +0} \frac{1}{\gamma}(\theta(\gamma) - \gamma\beta) = 0. \tag{3.3.5}$$

Indeed, since $\psi_{0,1} = o(|\theta|)$ we can determine $h', 0 < h' \leqq h_0$, such that $|M^{-1}\psi_{0,1}(\theta)| \leqq \frac{1}{6}\beta_{\min}\theta_{\max}/\beta_{\max}$ if $0 \leqq \theta^i \leqq h'$ $(i = 0, \cdots, m$, where $\beta_{\min} = \min \beta^i$, $\beta_{\max} = \max \beta^i$ and $\theta_{\max} = \max \theta^i$ $(i = 0, \cdots, m)$. We set $\bar{\gamma} = 2h'/(3\beta_{\max})$ and choose, for every $\gamma \in (0, \bar{\gamma}]$, an integer $k(\gamma)$ sufficiently large so that $\gamma^{-1} e_2^{k(\gamma)}(\theta) \in U'$ for all $\theta \in \mathscr{T}'$ and

$$|M^{-1} e_{0,1}^{k(\gamma)}(\theta)| \leqq \min(\tfrac{1}{4}\gamma\beta_{\min}, \gamma^2). \tag{3.3.6}$$

Let $\mathscr{T}_\gamma'' = \{\theta \in E_{m+1} \mid |\theta^i - \gamma\beta^i| \leqq \frac{1}{2}\gamma\beta_{\min} (i = 0, 1, \cdots, m)\}$. Then $\mathscr{T}_\gamma'' \subset \mathscr{T}'$ and $\mathscr{T}_\gamma''$ is homeomorphic to a closed ball in E_{m+1}. The function

$$\theta \to \gamma\beta - M^{-1}(\psi_{0,1}(\theta) + e_{0,1}^{k(\gamma)}(\theta))$$

is continuous and maps $\mathscr{T}_\gamma''$ into itself; it admits, therefore, a fixed point $\theta(\gamma)$ which, in view of (3.3.2), satisfies relation (3.3.3). Relation (3.3.4) is satisfied because of the choice of $k(\gamma)$ and (3.3.5) is implied by $\psi(\theta) = o(|\theta|)$ and (3.3.6).

It follows that $\psi(\theta(\gamma)) = o(\gamma)$ and $\theta(\gamma) - \gamma\beta = o(\gamma)$. Furthermore, relations (3.3.1) and (3.3.4) imply that, for all $\gamma \in (0, \bar{\gamma}]$,

$$\frac{1}{\gamma}\left(y_2^{k(\gamma)}(\theta(\gamma)) - \bar{y}_2(0) - \sum_{i=0}^{m} \theta^i(\gamma)\xi_2^i\right) \in \frac{1}{\gamma}\psi_2(\theta(\gamma)) + U';$$

hence

$$y_2^{k(\gamma)}(\theta(\gamma)) - \bar{y}_2(0) - \gamma \sum_{i=0}^{m} \beta^i \xi_2^i + a(\gamma) \in \gamma U', \tag{3.3.7}$$

where $a(\gamma) = o(\gamma)$. Now let $\gamma' > 0$ be small enough so that $\gamma' \leqq \min(\bar{\gamma}, 1)$ and $a(\gamma)/\gamma \in U'$ for all $\gamma \leqq \gamma'$. Then relation (3.3.7) implies that

$$\frac{1}{\gamma}(y_2^{k(\gamma)}(\theta(\gamma)) - \bar{y}_2(0)) \in \sum_{i=0}^{m} \beta^i \xi_2^i + U_1 \subset C'.$$

Since $0 \in \bar{C}'$, it follows that

$$y_2^{k(\gamma)}(\theta(\gamma)) - \bar{y}_2(0) \in C'. \tag{3.3.8}$$

We recall that $M\beta = \sum_{i=0}^{m} \beta^i \xi_{0,1}^i$, $\sum_{i=0}^{m} \beta^i \xi_1^i = 0$ and $\xi_0^i < 0$ $(i = 0, \cdots, m)$. Thus relations (3.3.3) and (3.3.8) yield the remaining conclusions of the lemma.

4. Proofs of Theorems 2.2 and 2.3.

4.1. Proof of Theorem 2.2. Let W be the convex hull of $\{Dx(\bar{q}; q - \bar{q}) | q \in Q\} \cup \{(1, 0, 0)\} \subset E_1 \times E_m \times V$ and $C' = C^0 - x_2(\bar{q})$. Since $Dx(\bar{q}; q - \bar{q}) = 0$ for $q = \bar{q}$, the assumptions of Lemma 3.1 are satisfied. The first alternative of Lemma 3.1 yields directly the conclusion of Theorem 2.2. We shall assume, therefore, that the second alternative is valid, and we shall show that it leads to a contradiction.

Since $\xi^i \in W$ $(i = 0, \cdots, m)$, there exist points $q_i = (\sigma_i, b_i) \in Q$ $(i = 0, \cdots, m)$ such that ξ^i is a convex combination of $Dx(\bar{q}; q_i - \bar{q})$ and $(1, 0, 0)$. We can verify that the q_i can be chosen so that the set $\{Dx(\bar{q}; q_i - \bar{q}) | i = 0, \cdots, m\}$ has all the properties listed for $\{\xi^i | i = 0, \cdots, m\}$, and we may assume, therefore, that

$$\xi^i = Dx(\bar{q}; q_i - \bar{q}) \qquad (i = 0, \cdots, m). \tag{4.1.1}$$

For each $\theta \in \mathscr{T}_{m+1}$, we set $\tilde{\sigma}(\theta) = \bar{\sigma} + \sum_{i=0}^{m} \theta^i(\sigma_i - \bar{\sigma})$ and $\tilde{b}(\theta) = \bar{b} + \sum_{i=0}^{m} \theta^i \cdot (b_i - \bar{b})$.

By Lemma 3.2, we can construct a sequence $\{\rho_k(\theta)\}_{k=1}^{\infty}$ in $\mathscr{R}$ that converges to $\tilde{\sigma}(\theta)$ uniformly in $\theta \in \mathscr{T}_{m+1}$ and such that, for each k, $\theta \to \rho_k(\theta) : \mathscr{T}_{m+1} \to \mathscr{R}$ is continuous. By assumption, x is continuous when restricted to some neighborhood of $\bar{q} = (\bar{\sigma}, \bar{h})$ in $\mathscr{S} \times \tilde{b}(\mathscr{T}_{m+1})$ and, as previously mentioned, $\mathscr{S}$ is metric and compact; there exist, therefore, $h_0 > 0$ and $k_0 > 0$ such that, for $\mathscr{T}' = \{\theta \in E_{m+1} | 0 \leqq \theta^i \leqq h_0 \ (i = 0, \cdots, m)\}$, $\bar{y}(\theta) = x(\tilde{\sigma}(\theta), \tilde{b}(\theta))$, $y^k(\theta) = x(\rho_k(\theta), \tilde{b}(\theta))$ and $e^k(\theta) = y^k(\theta) - x(\tilde{\sigma}(\theta), \tilde{b}(\theta))$ $(k \geqq k_0)$, we have $\lim_{k \to \infty} e^k(\theta) = 0$ uniformly in $\theta \in \mathscr{T}'$ and the functions $\theta \to e^k(\theta)$ and $\theta \to y^k(\theta)$: $\mathscr{T}' \to E_1 \times E_m \times V$ are continuous for each k. We may assume that h_0 was chosen sufficiently small so that the function

$$\theta \to \psi(\theta) = x(\tilde{\sigma}(\theta), \tilde{b}(\theta)) - x(\bar{\sigma}, \bar{b}) - \sum_{i=0}^{m} \theta^i Dx((\bar{\sigma}, \bar{b}); (\sigma_i, b_i) - (\bar{\sigma}, \bar{b}))$$

is continuous in $\mathscr{T}'$, and, by assumption, $\psi(\theta) = o(|\theta|)$.

We set $U = V$. Then, in view of (4.1.1) and the definition of $y^k, e^k, \bar{y}$ and ψ, the assumptions of Lemma 3.3 are satisfied. The conclusion of Lemma 3.3 implies

that there exist $\theta' = \theta(\gamma')$, $\rho' = \rho_{k(\gamma')}(\theta') \in \mathscr{R}$ and $b' = \tilde{b}(\theta') \in B$ such that

$$x_0(\rho', b') < x_0(\bar{\sigma}, \bar{b}), \quad x_1(\rho', b') = 0, \quad x_2(\rho', b') \in C.$$

This shows that $\bar{q} = (\bar{\sigma}, \bar{b})$ can be neither a relaxed nor an original minimizing point, contrary to assumption.

4.2. Proof of Theorem 2.3. For $\xi_{1,2} = (\xi_1, \xi_2) \in E_m \times V$, let $\hat{g}(\xi_{1,2}) = g(x_{1,2}(\bar{q}) + \xi_{1,2}) - g(x_{1,2}(\bar{q}))$, $W = \{(\hat{g}(Dx_{1,2}(\bar{q}; q - \bar{q})) + a, \ Dx_{1,2}(\bar{q}; q - \bar{q})) \mid q \in Q, \ a \geqq 0\}$ and $C' = C^0 - x_2(\bar{q})$. Then the assumptions of Lemma 3.1 are satisfied and its first alternative yields the conclusion of Theorem 2.3. We shall assume, therefore, that the second alternative holds. Then there exist $q_i \in Q$ $(i = 0, \cdots, m)$ such that

$$\xi^i_{1,2} = Dx_{1,2}(\bar{q}; q_i - \bar{q}).$$

Furthermore, since $\xi^i \in W$, we have

$$\hat{g}(\xi^i_{1,2}) \leqq \xi^i_0 < 0 \quad (i = 0, \cdots, m) \quad \text{and} \quad \hat{g}\left(\sum_{i=0}^{m} \beta^i \xi^i_{1,2}\right) < 0. \tag{4.2.1}$$

We define $\mathscr{T}'$, $\tilde{\sigma}(\theta)$, $\tilde{b}(\theta)$ and $\rho_k(\theta)$ as in § 4.1. We then set $y^k_{1,2}(\theta) = x_{1,2}(\rho_k(\theta), \tilde{b}(\theta))$ and similarly define $\bar{y}_{1,2}(\theta)$, $e^k_{1,2}(\theta)$ and $\psi_{1,2}(\theta)$. Next we set

$$\bar{y}_0(\theta) = y^k_0(\theta) = \sum_{i=0}^{m} \theta^i \xi^i_0, \qquad e^k_0(\theta) = 0$$

and

$$\psi_0(\theta) = 0 \qquad (\theta \in \mathscr{T}', \quad k = 1, 2, \cdots).$$

Finally, in view of relation (4.2.1), we can determine $\varepsilon > 0$ and neighborhoods U and $\tilde{U}$ of 0 in V such that

$$\hat{g}\left(\sum_{i=0}^{m} \beta^i \xi^i_{1,2} + (v_1, v_2)\right) < 0 \quad \text{for } |v_1| \leqq \varepsilon, \ v_2 \in \tilde{U}, \tag{4.2.2}$$

and

$$U + U \subset \tilde{U}.$$

Then the assumptions of Lemma 3.3 are satisfied, and it follows that there exist $\gamma' > 0$, $\theta(\gamma) \in \mathscr{T}'$, $\rho(\gamma) = \rho_{k(\gamma)}(\theta(\gamma)) \in \mathscr{R}$ and $b(\gamma) = \tilde{b}(\theta(\gamma)) \in B$ $(0 < \gamma \leqq \gamma')$ such that $\psi(\theta(\gamma)) = o(\gamma)$, $\gamma^{-1} e^{k(\gamma)}_2(\theta(\gamma)) \in U$, $x_1(\rho(\gamma), b(\gamma)) = 0$ and $x_2(\rho(\gamma), b(\gamma)) \in C$. Furthermore, we have

$$y^{k(\gamma)}_1(\theta(\gamma)) - \bar{y}_1(0) = x_1(\rho(\gamma), b(\gamma)) - x_1(\bar{\sigma}, \bar{b}) = \gamma \sum_{i=0}^{m} \beta^i \xi^i_1 = 0$$

and

$$y^{k(\gamma)}_2(\theta(\gamma)) - \bar{y}_2(0) = x_2(\rho(\gamma), b(\gamma)) - x_2(\bar{\sigma}, \bar{b}) = \gamma \sum_{i=0}^{m} \beta^i \xi^i_2 + o(\gamma) + e^{k(\gamma)}_2(\theta(\gamma)).$$

Since $\hat{g}$ is continuous and convex and $\hat{g}(0) = 0$, these relations and (4.2.2) imply that for sufficiently small γ,

$$\frac{1}{\gamma}\hat{g}(x_{1,2}(\rho(\gamma), b(\gamma)) - x_{1,2}(\bar{\sigma}, \bar{b})) \leqq \hat{g}\left(\frac{1}{\gamma}(x_{1,2}(\rho(\gamma), b(\gamma)) - x_{1,2}(\bar{\sigma}, \bar{b}))\right) < 0.$$

Thus, for sufficiently small γ, $x_0(\rho(\gamma), b(\gamma)) = g(x_{1,2}(\rho(\gamma), b(\gamma))) < g(x_{1,2}(\bar{\sigma}, \bar{b}))$, $x_1(\rho(\gamma), b(\gamma)) = 0$ and $x_2(\rho(\gamma), b(\gamma)) \in C$, contradicting the assumption that $(\bar{\sigma}, \bar{b})$ is either a relaxed or an original minimizing point. Thus the second alternative of Lemma 3.1 is inadmissible and Theorem 2.3 is valid.

Acknowledgment. I wish to acknowledge with thanks several stimulating conversations with L. W. Neustadt.

REFERENCES

[1] J. WARGA, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112 (1964), pp. 432–455.

[2] ———, *Unilateral variational problems with several inequalities*, Michigan Math. J., 12 (1965), pp. 449–480.

[3] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems. II: Applications*, this Journal, 5 (1967), pp. 90–137.

[4] ———, *A general theory of extremals*, J. Comp. System Sci., 3 (1969), pp. 57–92.

[5] J. WARGA, *On a class of minimax problems in the calculus of variations*, Michigan Math. J., 12 (1965), pp. 289–311.

[6] ———, *Relaxed controls for functional equations*, J. Functional Anal., 5 (1970), pp. 71–93.

[7] H. T. BANKS, *Variational problems involving functional differential equations*, this Journal, 7 (1969), pp. 1–17.

[8] W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems. I: General theory*, this Journal, 4 (1966), pp. 505–527.

[9] R. V. GAMKRELIDZE AND G. L. KHARATISHVILI, *Extremal problems in linear topological spaces. I*, Math. Systems Theory, 3 (1967), pp. 229–256.

[10] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[11] J. WARGA, *Functions of relaxed controls*, this Journal, 5 (1967), pp. 628–641.

[12] L. C. YOUNG, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Sci. Lettres Varsovie, CL III, 30 (1937), pp. 212–234.

[13] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl. 4 (1962), pp. 111–128.

[14] E. C. MCSHANE, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.

[15] A. GHOUILA-HOURI, *Sur la generalisation de la notion de commande d'un systeme guidable*, Rev. Francaise Informat. Recherche Operationnelle, 1 (1967), no. 4, pp. 7–32.

[16] T. NISHIURA, *On an existence theorem for optimal control*, this Journal, 5 (1967), pp. 532–544.

[17] E. J. MCSHANE, *Necessary conditions in generalized-curve problems of the calculus of variations*, Duke Math. J., 7 (1940), pp. 1–27.

[18] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part I*, Interscience, New York, 1967.

SIAM J. CONTROL
Vol. 8, No. 3, August 1970

UNILATERAL AND MINIMAX CONTROL PROBLEMS DEFINED BY INTEGRAL EQUATIONS*

J. WARGA†

1. Introduction. We consider control problems defined by the Uryson-type integral equation

$$y(t) = \int_T g(t, \tau, y(\tau), \rho(\tau), b)\, d\tau \qquad (t \in T), \tag{1.1}$$

where T is a compact metric space with an appropriate measure dt, $y \in C(T, E_n)$ (the Banach space of continuous functions from T to the Euclidean n-space E_n), ρ is a control function and b a control parameter. For given sets $A_1 \subset E_{m_1}$ and $A \subset E_{m_2}$, functions $h^0 : E_n \to E_1$, $h^1 : E_n \to E_{m_1}$ and $h^2 : T \times E_n \to E_{m_2}$ and a point $t_1 \in T$, we require that the admissible solutions (y, ρ, b) of (1.1) satisfy the restrictions

$$h^1(y(t_1)) \in A_1 \tag{1.2}$$

and

$$h^2(t, y(t)) \in A \qquad (t \in T). \tag{1.3}$$

Our purpose is to study points (y, ρ, b) that minimize $h^0(y(t_1))$ in the class of admissible solutions as well as to investigate the corresponding relaxed problem.

We have studied a similar problem in [1] without the unilateral restriction (1.3) and with ρ replaced by a relaxed control (but with y in either $C(T, E_n)$ or $L^p(T, E_n)$). The unilateral problem (with y subject to restriction (1.3)) has been studied in the special case where T is an interval of the real axis and (1.1) is equivalent to an ordinary differential equation; we have obtained existence theorems and necessary conditions for a relaxed minimum in [2] and [3] and L. W. Neustadt has derived necessary conditions for an ordinary minimum in [4] and [5] (first results in this area having been obtained by R. V. Gamkrelidze [6], [7]). Necessary conditions for the unilateral problem have also been investigated by V. R. Vinokurov [8] for the special case of Volterra-type integral equations but his arguments are heuristic and some of the results inaccurate.

In the present paper we restate (§ 2) with slight modifications certain results of [1] that establish the existence of a relaxed minimizing control and show that it can be approximated by original (ordinary) controls. We then apply (§§ 3 and 5) the results of [1] and [9] to our present problem and derive necessary conditions that are satisfied by both relaxed and original minimizing solutions. We also describe (§ 4) certain minimax problems and problems with other functional restrictions to which our methods are applicable and for which they yield analogous results.

* Received by the editors May 12, 1969, and in revised form October 21, 1969.

† Department of Mathematics, College of Engineering, Northeastern University, Boston, Massachusetts 02115. This research was supported by the National Aeronautics and Space Administration under Grant 22-011-020, Supplement 2.

2. Existence and approximation of relaxed minimizing solutions. Let T and R be compact metric spaces, with a positive, finite, regular, complete and non-atomic measure dt defined on T, $\mathscr{R}$ the class of measurable mappings from T to a dense subset $\tilde{R}$ of R, and B a convex subset of a Banach space. We assume that the function $(t, \tau, v, r, b) \to g(t, \tau, v, r, b): T \times T \times E_n \times R \times B \to E_n$ is measurable in (t, τ) (with respect to the product measure $dtd\tau$ on $T \times T$) for every (v, r, b) and continuous in (v, r, b) for every (t, τ). We also assume that h^0, h^1 and h^2 are continuous.

Let S be the class of regular Borel probability measures on R, and $\mathscr{S}$ the class of mappings $\sigma: T \to S$ (relaxed controls) such that the function $t \to \int_R c(r)\sigma(dr; t): T \to E_1$ is measurable for every continuous $c: R \to E_1$. We identify each "original control" $\rho \in \mathscr{R}$ with the element $\sigma_\rho \in \mathscr{S}$ such that the measure $\sigma_\rho(t)$ is concentrated at the point $\rho(t)$ with probability 1 for almost all $t \in T$. We also identify all mappings σ in $\mathscr{S}$ that coincide a.e. in T. We choose for $\mathscr{S}$ the smallest topology containing sets of the form

$$\left\{\sigma \in \mathscr{S} \,\middle|\, \left| \int_T dt \int_R \phi(t,r)(\sigma(dr; t) - \sigma_1(dr; t)) \right| < \varepsilon \right\},$$

where $\varepsilon > 0$, $\sigma_1 \in \mathscr{S}$ and $(t, r) \to \phi(t, r)$ is real-valued, measurable in t for every r and continuous in r for every t, with $\int_T \sup_{r \in R} |\phi(t, r)|\, dt < \infty$. (This choice of topology for $\mathscr{S}$ is equivalent to that given in [10, pp. 630–631], [1, p. 74] and [9, p. 372].)

We set

$$f(t, \tau, v, s, b) = \int_R g(t, \tau, v, r, b)s(dr)$$

for all Borel measures s on R and all (t, τ, v, b), and consider the relation

$$y(t) = \int_T f(t, \tau, y(\tau), \sigma(\tau), b)\, d\tau \qquad (t \in T) \tag{2.0.1}$$

for $(y, \sigma, b) \in C(T, E_n) \times \mathscr{S} \times B$. When $\sigma = \rho \in \mathscr{R}$ (in the sense just defined), (2.0.1) and (1.1) coincide.

We say that the point $(y, \sigma, b) \in C(T, E_n) \times \mathscr{S} \times B$ is a *relaxed admissible solution* if it satisfies (2.0.1), (1.2) and (1.3). If, furthermore, $\sigma \in \mathscr{R}$, then we say that (y, σ, b) is an *original admissible solution.* A *relaxed* (respectively *original*) *minimizing solution* is one that yields the minimum of $h^0(y(t_1))$ among all relaxed (respectively original) admissible solutions.

The following existence and approximation theorems follow essentially from [1, Theorem 3.1, Theorem 3.2, p. 76].

THEOREM 2.1. *Let A and A_1 be closed and B compact. Then there exists a*

relaxed minimizing solution $(\bar{y}, \bar{\sigma}, \bar{b})$ *if the following conditions are satisfied*:

(i) *There exists at least one relaxed admissible solution*;

(ii) *There exists a positive function* ψ_0, *integrable on* T, *and such that, for every* $(\sigma, b) \in \mathscr{S} \times B$ *and measurable* $y: T \to E_n$ *that satisfy* (2.0.1) *a.e. in* T, *we have*

$$|g(t, \tau, y(\tau), r, b')| \leqq \psi_0(\tau) \qquad \text{on } T \times T \times R \times B$$

and the function $t \to \int_T f(t, \tau, y(\tau), \sigma(\tau), b)\, d\tau$ *is continuous.*

THEOREM 2.2. *Let the assumptions of Theorem* 2.1 *be satisfied, and let* $(\bar{y}, \bar{\sigma}, \bar{b})$ *be a relaxed minimizing solution. Assume, furthermore, that* $\bar{y}$ *is the unique solution in* $C(T, E_n)$ *of* (2.0.1) *for* $\sigma = \bar{\sigma}$ *and* $b = \bar{b}$ *and that* (1.1) *admits at least one solution* y *in* $C(T, E_n)$ *for* $b = \bar{b}$ *and every* $\rho \in \mathscr{R}$ *in some neighborhood of* $\bar{\sigma}$ *in* $\mathscr{S}$. *Then there exists a sequence* $\{\rho_i\}_{i=1}^{\infty}$ *in* $\mathscr{R}$ *and a sequence* $\{y_i\}_{i=1}^{\infty}$ *in* $C(T, E_n)$ *such that the* $(y_i, \rho_i, \bar{b})$ *satisfy* (1.1) *and* $\lim_{i \to \infty} y_i(t) = \bar{y}(t)$ *for all* $t \in T$.

3. Necessary conditions. Let $(\bar{y}, \bar{\sigma}, \bar{b})$ be either a relaxed or an original minimizing solution. We shall show that $(\bar{y}, \bar{\sigma}, \bar{b})$ satisfies certain conditions that generalize the Weierstrass E-condition (Pontryagin's maximum principle) and the transversality conditions.

We denote by $|\cdot|$ the norm in a Banach space. If $\mathscr{X}$ and $\mathscr{Y}$ are Banach spaces, $\Omega \subset \mathscr{X}$ and $H: \Omega \to \mathscr{Y}$, we say that a linear operator $H_\omega(\bar{\omega}): \mathscr{X} \to \mathscr{Y}$ is a derivative, with respect to Ω, of H at $\bar{\omega} \in \Omega$ if $|H(\omega) - H(\bar{\omega}) - H_\omega(\bar{\omega})(\omega - \bar{\omega})| = o(|\omega - \bar{\omega}|)$ as $\omega \to \bar{\omega}$, $\omega \in \Omega$. If H is a function of several arguments ω_1, ω_2, etc., we denote partial derivatives by H_{ω_1}, H_{ω_2}, etc.

Assumption 3.1.

(i) There exist derivatives $g_v(t, \tau, v, r, b)$ and $g_b(t, \tau, v, r, b)$ for all $(t, \tau, v, r, b) \in T \times T \times E_n \times R \times B$, and they are measurable in (t, τ) (with respect to the product measure $dtd\tau$ on $T \times T$) for every (v, r, b) and continuous in (v, r, b) for every (t, τ); furthermore, for each $(\tau, r, b) \in T \times R \times B$, the functions $v \to g(t, \tau, v, r, b)$ $(t \in T)$ are equicontinuous on every compact subset of E_n;

(ii) There exist a compact set $D \subset E_n$ containing $\{\bar{y}(t) | t \in T\}$ in its interior, a measurable function ψ_0 on T, a modulus of continuity Φ, and a neighborhood N of $(\bar{\sigma}, \bar{b})$ in $\mathscr{S} \times B$ such that, for $\hat{g} = g$, g_v and g_b, we have $\sup_{t \in T} \int_T \psi_0(\tau)\, d\tau < \infty$,

$$|\hat{g}(t, \tau, v, r, b)| \leqq \psi_0(\tau) \qquad \text{on } T \times D \times R \times B, \tag{3.1.1}$$

$$\int_T \sup_{D \times R \times B} |\hat{g}(t_1, \tau, v, r, b) - \hat{g}(t_2, \tau, v, r, b)| d\tau \leqq \Phi(\text{distance } (t_1, t_2)) \tag{3.1.2}$$

for all $t_1, t_2 \in T$, and (2.0.1) has a unique continuous solution $\tilde{y}(\sigma, b)$ for all $(\sigma, b) \in N$, with $\tilde{y}(\sigma, b)(t) \in D$ $(t \in T)$;

(iii) For $k(t, \tau) = f_v(t, \tau, \bar{y}(\tau), \bar{\sigma}(\tau), \bar{b})$ $(t, \tau \in T)$, the integral equation

$$w(t) = \int_T k(t, \tau) w(\tau)\, d\tau \qquad (t \in T)$$

has only the trivial solution $w(\cdot) = 0$;

(iv) The functions $v \to h_v^0(v)$, $v \to h_v^1(v)$ and $(t, v) \to h_v^2(t, v)$ exist and are continuous for $(t, v) \in T \times E_n$.

We shall state necessary conditions for minimum for the case where the set A of the unilateral restriction (1.3) is a convex body and the set A_1 of (1.2) contains only the origin of E_{m_1}. The case where A and A_1 are diffeomorphic, respectively, to a convex body in E_{m_2} and a convex subset of some E_l can be easily transformed to the form just mentioned.

We observe that, as a consequence of Assumption 3.1 (iii), there exists a resolvent kernel k^* of k, that is, a measurable real matrix-valued function $k^* = (k^{*i}j)$ $(i, j = 1, \cdots, n)$ on $T \times T$ such that $\sup_{t \in T} \int_T |k^*(t, \tau)|\, d\tau < \infty$ and, for every $z \in C(T, E_n)$, the relations

$$w(t) = \int_T k(t, \tau) w(\tau)\, d\tau + z(t) \qquad (t \in T)$$

and

$$w(t) = \int_T k^*(t, \tau) z(\tau)\, d\tau + z(t) \qquad (t \in T)$$

are equivalent for $w \in C(T, E_n)$ (see [1, Lemma 7.2, p. 90]).

THEOREM 3.2. *Let $(\bar{y}, \bar{\sigma}, \bar{b})$ be either a relaxed or an original minimizing solution, let Assumption* 3.1 *be satisfied, and let k^* be a resolvent kernel of k. We assume, furthermore, that A is a convex body in E_{m_2} and A_1 contains only the origin of E_{m_1}. Then there exist $\lambda^0 \geqq 0$, $\lambda \in E_{m_1}$, a nonnegative finite regular measure ω on T and an ω-integrable function $\Phi : T \to E_{m_2}$ such that*

(i) $\lambda^0 + |\lambda| + \int_T |\phi(\tau)| \omega(d\tau) \neq 0$;

(ii) *(Weierstrass E-condition) for* $\chi(dt) = \phi(t) \cdot h_v^2(t, \bar{y}(t)) \omega(dt)$ $(t \neq t_1, t \in T)$,

$$\chi(dt_1) = \lambda^0 h_v^0(\bar{y}(t_1)) + \lambda \cdot h_v^1(\bar{y}(t_1)) + \phi(t_1) \cdot h_v^2(t_1, \bar{y}(t_1)) \omega(dt_1)$$

and

$$\alpha(t, \tau, s) = f(t, \tau, \bar{y}(\tau), s, \bar{b}) + \int_T k^*(t, \theta) f(\theta, \tau, \bar{y}(\tau), s, \bar{b})\, d\theta \qquad (s \in S),$$

we have

$$\int_T \alpha(t, \tau, s) \cdot \chi(dt) \geqq \int_T \alpha(t, \tau, \bar{\sigma}(\tau)) \cdot \chi(dt) \tag{3.2.1}$$

for all $s \in S$ and almost all $\tau \in T$;

(iii) *(Support (transversality) conditions)*

$$\int_T \left[\int_T f_b(t, \tau, \bar{y}(\tau), \bar{\sigma}(\tau), \bar{b})(b - \bar{b})\, d\tau + \int_{T \times T} k^*(t, \theta) f_b(\theta, \tau, \bar{y}(\tau), \bar{\sigma}(\tau), \bar{b}) \cdot (b - \bar{b})\, d\tau\, d\theta \right] \cdot \chi(dt) \geqq 0 \qquad \text{for all } b \in B;$$

and

(iv) $\phi(t) \cdot h^2(t, \bar{y}(t)) \geqq \phi(t) \cdot a$ *for all* $a \in A$ *and* ω*-almost all* $t \in T$.

Remark. In the special case where T is an interval $[t_0, t_1]$ of the real axis and (1.1) is equivalent to an ordinary differential equation, the above results slightly generalize those of [4], [5], [6] and [7]. Our results are also applicable to the problem considered in [8].

4. Minimax problems and problems with other functional restrictions. Let P be a compact metric space, and let a function $\tilde{g}$ be defined in the same manner as the function g but depend, in addition, on a parameter $p \in P$. We consider solutions (y, σ, b) of the equation

$$y(t, p) = \int_T d\tau \int_R \tilde{g}(t, \tau, y(\tau, p), r, b, p)\sigma(dr; \tau) \qquad (t \in T, \quad p \in P), \tag{4.0.1}$$

which is a generalization of (2.0.1). With suitable assumptions, analogous to those of §§ 2 and 3, for every such solution (y, σ, b) the function $p \to y(t_1, p)$, where $t_1 \in T$ is given, is continuous. We can now consider a variety of problems. In one such problem, we wish to minimize $\sup_{p' \in P} y^1(t_1, p')$ subject to the conditions

$$y(t_1, p_1) \in A_1 \quad \text{and} \quad y(t_1, P) \in A,$$

where $p_1 \in P$, $A_1 \subset E_n$ and $A \subset E_n$ are given. This problem was previously considered in [11] in the special case where (4.0.1) is equivalent to an ordinary differential equation and only relaxed minimizing solutions are considered.

Another problem can be defined by choosing continuous functions h^0, h^1 and h^2 from $E_n \times B \times P$ into, respectively, E_1, E_{m_1} and E_{m_2} and a Borel measure μ on P, and minimizing $\int_P h^0(y(t_1, p), b, p)\mu(dp)$ subject to the conditions

$$\int_P h^1(y(t_1, p), b, p)\mu(dp) \in A_1 \quad \text{and} \quad h^2(y(t_1, p), b, p) \in A \qquad (p \in P). \tag{4.0.2}$$

This is a fairly natural setting for control problems in which μ is a known probability measure on P, and it is desired to minimize an expected value with restrictions placed on variances or higher moments and with the range of the function $p \to y(t_1, p)$ confined to a "permissible" region A.

We shall only present results concerning the latter problem since the minimax problem can be easily transformed into a special case of the second. Theorems 4.1 and 4.2 below can be proved using arguments that slightly extend those in [1], [9] and § 5, and their proofs will therefore be omitted.

THEOREM 4.1. *Let* A *and* A_1 *be closed,* B *compact, and* $\mathscr{A}$ *be the set of all the points* $(y, \sigma, b) \in C(T \times P, E_n) \times \mathscr{S} \times B$ *that satisfy* (4.0.1) *and* (4.0.2). *Assume that*

(i) $\mathscr{A}$ *is nonempty*;

(ii) *There exists a positive integrable function* ψ_0 *on* T *such that*

$$|\tilde{g}(t, \tau, y(\tau, p), r, b', p)| \leqq \psi_0(\tau) \qquad \text{on } T \times T \times R \times B \times P$$

for every $(y, \sigma, b) \in \mathscr{A}$;

$$\text{(iii)} \quad \lim_{(t,p)\to(\bar{t},\bar{p})} \int \sup_{D\times R\times B} |\tilde{g}(\bar{t}, \tau, v, r, b, \bar{p}) - \tilde{g}(t, \tau, v, r, b, p)|\, d\tau = 0$$

for all $(\bar{t}, \bar{p}) \in T \times P$ *and every compact subset* D *of* E_n;

(iv) *For every* $(\sigma, b) \in \mathscr{S} \times B$, *equation* (4.0.1) *has a unique solution* y.

Then there exists a point $(\bar{y}, \bar{\sigma}, \bar{b})$ *that minimizes* $\int_P h^0(y(t_1, p), b, p)\mu(dp)$ *among all* $(y, \sigma, b) \in \mathscr{A}$ *and a sequence* $\{(\rho_i, y_i)\}$ *in* $\mathscr{R} \times C(T \times P, E_n)$ *such that the* $(y_i, \rho_i, \bar{b})$ *satisfy* (4.0.1) *and* $\lim_i y_i = \bar{y}$ *uniformly on* $T \times P$.

THEOREM 4.2. *Let* $\mathscr{A}$ *be as described in Theorem* 4.1 *and let* $(\bar{y}, \bar{\sigma}, \bar{b})$ *be either an original or a relaxed minimizing solution (that is,* $(\bar{y}, \bar{\sigma}, \bar{b})$ *minimizes* $\int_P h^0(y(t_1, p), b, p)\mu(dp)$ *among all* $(y, \sigma, b) \in \mathscr{A}$ *for which* σ *is either restricted to* $\mathscr{R}$ *or allowed to range over* $\mathscr{S}$*). Let* $\tilde{g}$ *satisfy Assumption* 3.1 *for each* $p \in P$, *with* r *replaced by* (r, p) *in* 3.1 (i), (3.1.1) *and* (3.1.2), D, ψ_0 *and* Φ *independent of* p, *and* h_v^i *continuous for* $i = 0, 1, 2$, *and let* (4.0.1) *have a unique continuous solution* $(t, p) \to y(t, p)$ *for each* $(\sigma, b) \in \mathscr{S} \times B$. *Then there exist* $\lambda_0 \geqq 0$, $\lambda_1 \in E_{m_1}$, *a nonnegative finite regular measure* ω *on* P *and an* ω*-integrable function* $\phi: P \to E_{m_2}$ *such that*

(i) $\lambda_0 + |\lambda_1| + \int_P |\phi(p)|\omega(dp) \neq 0$;

(ii) *For*

$$\chi(dp) = \sum_{i=0}^{1} \lambda_i h_v^i(\bar{y}(t_1, p), \bar{b}, p)\mu(dp) + h_v^2(\bar{y}(t_1, p), \bar{b}, p) \cdot \phi(p)\omega(dp),$$

$$\alpha(\theta, p, s) = f(t_1, \theta, \bar{y}(\theta, p), s, \bar{b}, p) + \int_T k^*(t_1, \tau, p) f(\tau, \theta, \bar{y}(\theta, p), s, \bar{b}, p) d\tau$$

$$(\theta \in T, \quad p \in P, \quad s \in S)$$

and

$$\beta(p) = \int_T f_b(t_1, \tau, \bar{y}(\tau, p), \bar{\sigma}(\tau), \bar{b}, p)\, d\tau + \int_{T\times T} k^*(t_1, \tau, p) f_b(\tau, \theta, \bar{y}(\theta, p), \bar{\sigma}(\theta), p)\, d\tau\, d\theta,$$

we have:
(*Weierstrass E-condition*)

$$\int_P \chi(dp) \cdot \alpha(\theta, p, s) \geqq \int_P \chi(dp) \cdot \alpha(\theta, p, \bar{\sigma}(\theta))$$

for all $s \in S$ *and almost all* $\theta \in T$;

(Support (transversality) conditions)

$$\int_P \chi(dp) \cdot \beta(p)b \geqq \int_P \chi(dp) \cdot \beta(p)\bar{b}$$

for all $b \in B$; *and*

$$\phi(p) \cdot h^2(t_1, \bar{y}(t_1, p)) \geqq \phi(p) \cdot a$$

for all $a \in A$ *and* ω*-almost all* $p \in P$.

5. Proofs of Theorems 2.1, 2.2 and 3.2. We denote by $L^1(T, E_n)$ the Banach space of integrable functions from T to E_n.

5.1. Proof of Theorem 2.1. Assumptions (i) and (ii) of Theorem 2.1 and the arguments of [1, 6.2, p. 85] show that any sequence $\{(y_j, \sigma_j, b_j)\}_{j=1}^{\infty}$ in $L^1(T, E_n) \times \mathscr{S} \times B$ of solutions a.e. of (2.0.1) admits a subsequence, defined by $j \in J$, converging in $L^1(T, E_n) \times \mathscr{S} \times B$ to some $(\bar{y}, \bar{\sigma}, \bar{b})$ that satisfies (2.0.1) a.e. in T and such that

$$\lim_{j \in J} \int_T f(t, \tau, y_j(\tau), \sigma_j(\tau), b_j)\, d\tau = \int_T f(t, \tau, \bar{y}(\tau), \bar{\sigma}(\tau), \bar{b})\, d\tau$$

for all $t \in T$.

It follows then from assumption (ii) of Theorem 2.1 that the y_j and $\tilde{y}$ can be assumed continuous, $(\bar{y}, \bar{\sigma}, \bar{b})$ satisfies (2.0.1) for all $t \in T$ and

$$\lim_{j \in J} y_j(t) = \bar{y}(t) \qquad \text{for all } t \in T. \tag{5.1.1}$$

If we choose the (y_j, σ_j, b_j) among admissible relaxed solutions so that $h^0(y_j(t_1))$ converges to $\inf h^0(y(t_1))$ (among admissible relaxed solutions), then (5.1.1) shows that $\bar{y}$ satisfies (1.2) and (1.3) (because the sets A_1 and A are closed and h^1 and h^2 are continuous), and is, therefore, a relaxed minimizing solution. This completes the proof.

5.2. Proof of Theorem 2.2. As in [1, 6.3, p. 86], we choose a sequence $\{\rho_j\}_{j=1}^{\infty}$ in $\mathscr{R}$ converging to $\bar{\sigma}$ and such that (1.1) has a solution $(y_j, \dot{\rho}_j, \bar{b}) \in C(T, E_n) \times \mathscr{R} \times B$ for $j = 1, 2, \cdots$. It follows then from the arguments of 5.1 and the uniqueness assumption that $\lim_{j \to \infty} y_j(t) = \bar{y}(t)$ for all $t \in T$. This completes the proof.

5.3. Proof of Theorem 3.2. Let N be the neighborhood of $(\bar{\sigma}, \bar{b})$ referred to in Assumption 3.1 (ii), and let a sequence $\{(\sigma_j, b_j)\}_{j=1}^{\infty}$ in N converge to some $(\sigma', b') \in N$. Then an argument similar to those in 5.1 and 5.2 shows that

$$\lim_{j \to \infty} \tilde{y}(\sigma_j, b_j)(t) = \tilde{y}(\sigma', b')(t) \qquad \text{for all } t \in T.$$

Since (3.1.1) and (3.1.2) imply that $y(\sigma', b')(\cdot)$ and the $\tilde{y}(\sigma_j, b_j)(\cdot)$ form an equicontinuous family, it follows that $\lim_{j \to \infty} \tilde{y}(\sigma_j, b_j) = \tilde{y}(\sigma', b')$ in $C(T, E_n)$. The topology that we chose for $\mathscr{S}$ can be easily seen to be metric (being a relativization of the weak star topology in the topological dual of $L^1(T, C(R, E_1))$ [12, Theorem 2, p. 424 and Theorem 1, p. 426]. We conclude, therefore, that $(\sigma, b) \to \tilde{y}(\sigma, b)$ is continuous in N.

Let $m' = \max(m_1 + 1, 2)$ and

$$\mathscr{T} = \{(\theta^1, \cdots, \theta^{m'}) \in E_{m'} | \theta^j \geqq 0, \sum_{j=1}^{m'} \theta^j \leqq 1\}.$$

The arguments of [1, Lemma 7.1, p. 87] show that, for every fixed choice of $\{\sigma_1, \cdots, \sigma_{m'}\} \subset \mathscr{S}$ and $\{b_1, \cdots, b_{m'}\} \subset B$, there exists a neighborhood Γ of $(\bar{y}, 0)$ in $C(T, E_n) \times \mathscr{T}$ such that the function $\tilde{F}: \Gamma \to C(T, E_n)$, defined by

$$\tilde{F}(y, \theta)(t) = \int_T f\left(t, \tau, y(\tau), \bar{\sigma}(\tau) + \sum_{j=1}^{m'} \theta^j(\sigma_j(\tau) - \bar{\sigma}(\tau)), \bar{b} + \sum_{j=1}^{m'} \theta^j(b_j - \bar{b})\right) d\tau \qquad (t \in T),$$

exists and is continuous, the partial derivative $\tilde{F}_y$ exists and is continuous on Γ, and the (total) derivative $\tilde{F}_{(y,\theta)}(\bar{y}, 0)$ exists. Furthermore,

$$\begin{aligned}(\tilde{F}_y(\bar{y}, 0)\Delta y)(t) &= \int_T f_v(t, \tau, \bar{y}(\tau), \bar{\sigma}(\tau), \bar{b})\Delta y(\tau)\, d\tau \\ &= \int_T k(t, \tau)\Delta y(\tau)\, d\tau \qquad (\Delta y \in C(T, E_n), t \in T)\end{aligned}$$

and

$$\begin{aligned}\tilde{F}_{\theta^j}(\bar{y}, 0)(t) &= \int_T f(t, \tau, \bar{y}(\tau), \sigma_j(\tau) - \bar{\sigma}(\tau), \bar{b})\, d\tau \\ &\quad + \int_T f_b(t, \tau, \bar{y}(\tau), \bar{\sigma}(\tau), \bar{b})(b_j - \bar{b})\, d\tau \qquad (j = 1, \cdots, m', t \in T).\end{aligned} \tag{5.3.1}$$

By 3.1 (iii) and [1, Lemma 7.2, p. 90], the mapping $I - \tilde{F}_y(\bar{y}, 0)$ is a linear homeomorphism of $C(T, E_n)$ onto itself. It follows, by a slight modification of the implicit function theorem[1] [13, p. 265], that the equation

$$y(t) = \tilde{F}(y, \theta)(t) \qquad (t \in T)$$

has a unique solution

$$\eta(\theta) = \tilde{y}\left(\bar{\sigma} + \sum_{j=1}^{m'} \theta^j(\sigma_j - \bar{\sigma}), \bar{b} + \sum_{j=1}^{m'} \theta^j(b_j - \bar{b})\right)$$

for all θ in some neighborhood $\tilde{\mathscr{T}}$ of 0 in $\mathscr{T}$, with values in some neighborhood $\tilde{Y}$ of $\bar{y}$ in $C(T, E_n)$. Furthermore, $\theta \to \eta(\theta)$ is continuous and has a derivative at 0 defined by

$$\eta_\theta(0) = (I - \tilde{F}_y(\bar{y}, 0))^{-1}\tilde{F}_\theta(\bar{y}, 0),$$

[1] The implicit function theorem is proven in [13, p. 265] for the equation $f(x, y) = 0$, where f is defined on an open set. In our case, $\tilde{F}$ is defined for all y but with θ restricted to $\mathscr{T}$ (which is not open). This is the main reason for our having introduced, at the beginning of § 3, derivatives relative to a set. With this definition of derivatives, the arguments in [13, p. 265] (and in the required lemmas and theorems) can be used, with only slight modifications, to yield our assertion.

yielding

$$\eta_{\theta^j}(0)(t) = \tilde{F}_{\theta^j}(\bar{y}, 0)(t) + \int_T k^*(t, \tau)\tilde{F}_{\theta^j}(\bar{y}, 0)(\tau)\, d\tau \qquad (t \in T, j = 1, \cdots, m'). \tag{5.3.2}$$

Now let functions $x_0 : \mathscr{S} \times B \to E_1$, $x_1 : \mathscr{S} \times B \to E_{m_1}$ and $x_2 : \mathscr{S} \times B \to C(T, E_{m_2})$ be defined as follows:

for $(\sigma, b) \in N$, we set

$$x_0(\sigma, b) = h^0(\tilde{y}(\sigma, b)(t_1)), \qquad x_1(\sigma, b) = h^1(\tilde{y}(\sigma, b)(t_1))$$

and

$$x_2(\sigma, b)(t) = h^2(t, \tilde{y}(\sigma, b)(t)) \qquad (t \in T);$$

for $(\sigma, b) \notin N$, we set $x_0(\sigma, b) = h^0(\tilde{y}(\bar{\sigma}, \bar{b})(t_1)) + 1$, $x_1(\sigma, b) = 0$, $x_2(\sigma, b)(t) = 0\ (t \in T)$.

We have shown that the function $(\sigma, b) \to \tilde{y}(\sigma, b) : N \to C(T, E_n)$ is continuous and the function

$$\theta \to \eta(\theta) = \tilde{y}\left(\bar{\sigma} + \sum_{j=1}^{m'} \theta^j(\sigma_j - \bar{\sigma}), \bar{b} + \sum_{j=1}^{m'} \theta^j(b_j - \bar{b})\right) : \mathscr{T} \to \tilde{Y}$$

has a derivative at 0 defined by (5.3.1) and (5.3.2). It follows now, by Assumption 3.1 (iv), that the function $x = (x_0, x_1, x_2)$ is continuous in N and the function

$$\theta \to x\left(\bar{\sigma} + \sum_{j=1}^{m'} \theta^j(\sigma_j - \bar{\sigma}), \bar{b} + \sum_{j=1}^{m'} \theta^j(b_j - \bar{b})\right)$$

has a derivative at 0. Let

$$Dx(\bar{\sigma}, \bar{b}; (\sigma, b) - (\bar{\sigma}, \bar{b})) = \lim_{\alpha \to +0} \frac{1}{\alpha}(x(\bar{\sigma} + \alpha(\sigma - \bar{\sigma}), \bar{b} + \alpha(b - \bar{b})) - x(\bar{\sigma}, \bar{b}))$$

for $(\sigma, b) \in \mathscr{S} \times B$

and

$$C = \{w(\cdot) \in C(T, E_{m_2}) | w(t) \in A \ \ (t \in T)\}.$$

Then the assumptions of [9, Theorem 2.2, p. 373] are verified for $V = C(T, E_{m_2})$ and $m = m_1$, and we conclude that there exists a nonvanishing continuous linear functional l on $E_1 \times E_{m_1} \times C(T, E_{m_2})$ such that

$$l((v_0, v_1, v_2)) = \lambda_0 v_0 + \lambda_1 \cdot v_1 + l_2(v_2)$$

for $v_0 \in E_1$, $v_1 \in E_{m_1}$ and $v_2 \in C(T, E_{m_2})$, $\lambda_0 \geqq 0$,

$$l(Dx(\bar{\sigma}, \bar{b}; (\sigma, b) - (\bar{\sigma}, \bar{b}))) \geqq 0 \qquad \text{for all } (\sigma, b) \in \mathscr{S} \times B, \tag{5.3.3}$$

and

$$l_2(v_2) \leqq l_2(h^2(\cdot, \bar{y}(\cdot))) \qquad \text{for all } v_2 \in C. \tag{5.3.4}$$

We observe that if we choose $(\sigma_1, b_1) = (\sigma, b)$ and $(\sigma_j, b_j) = (\bar{\sigma}, \bar{b})$ $(j = 2, \cdots, m')$, then

$$Dx_i(\bar{\sigma}, \bar{b}; (\sigma, b) - (\bar{\sigma}, \bar{b})) = h_v^i(\bar{y}(t_1))\eta_{\theta^1}(0) \qquad (i = 0, 1) \tag{5.3.5}$$

and

$$Dx_2(\bar{\sigma}, \bar{b}; (\sigma, b) - (\bar{\sigma}, \bar{b}))(t) = h_v^2(t, \bar{y}(t))\eta_{\theta^1}(0) \qquad (t \in T). \tag{5.3.6}$$

Furthermore, since l_2 is a continuous linear functional on $C(T, E_{m_2})$, there exists a finite regular Borel vector measure $\mu = (\mu^1, \cdots, \mu^{m_2})$ such that

$$l_2(w(\cdot)) = \int_T w(t) \cdot \mu(dt) \qquad \text{for all } w \in C(T, E_{m_2}).$$

If we set $\omega(\tilde{T}) = \sum_{j=1}^{m_2}$ (variation of μ^j on $\tilde{T}$) for all Borel sets $\tilde{T} \subset T$, then, by the Radon-Nikodym theorem, there exists an ω-integrable $\phi : T \to E_{m_2}$ such that

$$\mu(\tilde{T}) = \int_{\tilde{T}} \phi(t)\omega(dt) \qquad \text{for all Borel sets } \tilde{T} \subset T.$$

Relation (i) of Theorem 3.2 follows from the conclusion that l is nonvanishing. Relations (5.3.i) ($i = 1, 2, 3, 5, 6$) yield, for $j = 1$ and $b = \bar{b}$,

$$\int_T \chi(dt) \cdot \left[\int_T f(t, \tau, \bar{y}(\tau), \sigma(\tau) - \bar{\sigma}(\tau), \bar{b})\, d\tau \right.$$
$$\left. + \int_T k^*(t, \theta)\, d\theta \int_T f(\theta, \tau, \bar{y}(\tau), \sigma(\tau) - \bar{\sigma}(\tau), \bar{b})\, d\tau \right] \geqq 0 \qquad \text{for all } \sigma \in \mathcal{S}. \tag{5.3.7}$$

We choose a denumerable dense subset $R_\infty = \{r_1, r_2, \cdots\}$ of R and set, for $i = 1, 2, \cdots$ and an arbitrary measurable subset E of T, $\sigma(t) = \bar{\sigma}(t)$ for $t \in T - E$, $\sigma(t) = \sigma_{r_i}(t)$ (a measure concentrated at r_i) for $t \in E$. By (3.1.1) and the properties of k^*, the integrand in (5.3.7) is absolutely integrable. Thus, interchanging the order of integration, we obtain

$$\int_E d\tau \int_T \chi(dt) \cdot \alpha(t, \tau, \sigma_{r_i}(\tau) - \bar{\sigma}(\tau)) \geqq 0$$

for $i = 1, 2, \cdots$ and all measurable subsets E of T. We conclude that

$$\int_T \chi(dt) \cdot \alpha(t, \tau, \sigma_{r_i}(\tau)) \geqq \int_T \chi(dt) \cdot \alpha(t, \tau, \bar{\sigma}(\tau))$$

for almost all $\tau \in T$ and $i = 1, 2, \cdots$, and it is now easy to deduce relation (3.2.1) (the details of the derivation are exactly as in [1, 7.3, p. 91]).

We can similarly deduce relation (iii) of Theorem 3.2 from relations (5.3.i) ($i = 1, 2, 3, 5, 6$), setting $j = 1$ and $\sigma = \bar{\sigma}$. Finally, relation (5.3.4) implies that

$$\int_T \phi(t) \cdot h^2(t, \bar{y}(t))\omega(dt) \geqq \int_T \phi(t) \cdot w(t)\omega(dt)$$

for every $w(\cdot) \in C$.

Relation (iv) of Theorem 3.2 then follows in a straightforward manner. This completes the proof of the theorem.

REFERENCES

[1] J. WARGA, *Relaxed controls for financial equations*, J. Functional Anal., 5 (1970), pp. 71–93.
[2] ———, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112 (1954), pp. 432–455.
[3] ———, *Unilateral variational problems with several inequalities*, Michigan Math. J., 12 (1965), pp. 449–480.
[4] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems. II: Applications*, this Journal, 5 (1967), pp. 90–137.
[5] ———, *A general theory of extremals*, J. Comput. and System Sci., 3 (1969), pp. 57–92.
[6] R. V. GAMKRELIDZE, *Optimal control processes with restricted phase coordinates*, Izv. Akad. Nauk SSSR Ser. Mat., 24 (1960), pp. 315–356.
[7] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
[8] V. R. VINOKUROV, *Optimal control of processes describable by integral equations. I*, Izv. Vyssh. Uchebn. Zaved. Matematika, 62 (1967), pp. 21–33; English transl., this Journal, 7 (1969), pp. 326–336.
[9] J. WARGA, *Control problems with functional restrictions*, this Journal, 8 (1970), pp. 360–371.
[10] ———, *Functions of relaxed controls*, this Journal, 5 (1967), pp. 628–641.
[11] ———, *On a class of minimax problems in the calculus of variations*, Michigan Math. J., 12 (1965), pp. 289–311.
[12] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part I*, Interscience, New York, 1967.
[13] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

SIAM J. CONTROL
Vol. 8, No. 3, August 1970

ON FIXED TIME CONTROL PROBLEMS IN A BANACH SPACE*

GUNTER H. MEYER†

1. Introduction. It is the purpose of this paper to present a rigorous derivation of the dynamic programming equation of Bellman for fixed time (i.e., free endpoint) control problems in an arbitrary Banach space. It will be shown that this dynamic programming equation is a generalization of the invariant imbedding equation for two-point boundary value problems which, for a given control, converts the underlying boundary value problem into an initial value problem. Furthermore, the dynamic programming equations for controlled multipoint boundary value problems will be derived.

The development of our theory for vector-valued functions produces only technical but no conceptual difficulties as compared to the scalar case, and in order to illustrate our approach we shall consider the following scalar problem:

$$\begin{aligned} J'(t) &= F_1(J, y, u, t), \qquad J(T) = \alpha, \\ y'(t) &= F_2(J, y, u, t), \qquad y(0) = \beta, \end{aligned} \tag{1.1}$$

where F_1 and F_2 are continuously differentiable, and where u belongs to a given set K of piecewise differentiable functions on $[0, T]$. Our intention is to maximize $J(0)$ through the proper choice of u.

For fixed $u \in K$, problem (1.1) can be imbedded into the family of initial value problems:

$$\begin{aligned} J'(t) &= F_1(J, y, u, t), \qquad J(T) = \alpha, \\ y'(t) &= F_2(J, y, u, t), \qquad y(T) = s. \end{aligned} \tag{1.2}$$

If these equations are considered as characteristic equations, then it is well known (see, e.g., Courant and Hilbert [3, pp. 62–69]) that for varying s the solutions of (1.2) generate an integral surface $J(t, x)$ for the so-called invariant imbedding equation corresponding to (1.1),

$$J_t(t, y) + J_y(t, y)F_2(J, y, u, t) = F_1(J, y, u, t), \qquad J(T, y) = \alpha. \tag{1.3}$$

We remark that $J_t(t, y)$ is discontinuous at the singularities of $u(t)$ because the characteristics have corners at these points. It now follows that if the integral surface $J(t, y)$ exists for $t \in [0, T]$ and all y, then the characteristic through $(J(0, \beta), \beta)$ is a solution of (1.1). Moreover, if $u \in K$ satisfies

$$\begin{aligned} J_t(t, \beta, \hat{u}) &= \inf_{u \in K} [F_1(J, \beta, u, t) - J_y(t, \beta, u)F_2(J, \beta, u, t)], \\ J(T, \beta, u) &= \alpha, \end{aligned} \tag{1.4}$$

* Received by the editors June 18, 1968, and in revised form September 27, 1969.

† Mobil Research and Development Corporation, Field Research Laboratory, Dallas, Texas 75221.

then it is seen that for any other control $v \in K$,

$$0 \geqq \int_0^T [J_t(t, \beta, \hat{u}) - J_t(t, \beta, v)]\, dt = J(0, \beta, v) - J(0, \beta, \hat{u}),$$

and hence that $J(0, \beta, v) \leqq J(0, \beta, \hat{u})$. Consequently, for $\hat{u}$ to be optimal it is sufficient that it satisfy (1.4). Equation (1.4) is exactly Bellman's dynamic programming equation (see [2, p. 264]).

Our results will be presented in two sections. The derivation of the dynamic programming equation for controlled Banach-space-valued two- and multipoint boundary value problems is discussed in § 2. The next and main section deals with linear controlled equations. In this case the partial differential dynamic programming equation reduces to two ordinary differential equations with the effect that the choice of optimal control and the solution of the resulting boundary value problem can be separated. In order to illustrate our theory several maximum principles derived by Friedman [4] for parabolic evolution equations will be rederived and reinterpreted from our point of view. Finally, the equations for linear controlled multipoint boundary value problems are established.

2. The dynamic programming equation. Our development will be based on an extension of the abstract characteristic theory which was used in [8] for the conversion of multipoint boundary value problems into initial value problems. For ease of exposition we shall restrict ourselves to fixed time control problems which for a given control function reduce to two-point boundary value problems. The application of our theory to controlled N-point boundary value problems is straightforward and will be outlined at the end of this section.

Let us introduce our notation. X, Y, and Z will always denote real Banach spaces, D is an open convex subset of a product space to be defined from case to case, and I is an open interval on the real line R. We shall consider differential equations of the following general form:

$$x' = G(t, x, y), \qquad y' = H(t, x, y), \tag{2.1}$$

where G and H are (bounded or unbounded) functions on $I \times D \subset R \times X \times Y$ with values in X and Y, respectively. A solution of (2.1) subject to given initial conditions

$$x(t_0) = x_0, \qquad y(t_0) = y_0, \tag{2.2}$$

with $(x_0, y_0) \in D$ and $t_0 \in \bar{I}$, is defined to be a pair $\{x(t), y(t)\}$ of strongly continuous functions from $\bar{I}$ to D which satisfy (2.1) Lebesgue almost everywhere on I.

In analogy to the finite-dimensional setting we shall associate the following partial differential equation with (2.1):

$$w_t(t, y) + w_y(t, y)H(t, w, y) = G(t, w, y), \tag{2.3}$$

where $w_y(t, y)$ denotes the Fréchet derivative of w with respect to y at the point (t, y). A solution of (2.3) through a given point $(t_0, x_0, y_0) \in \bar{I} \times D$ is defined to be a strongly continuous function w on $\bar{I} \times D \cap Y$ with values in $D \cap X$ which

satisfies (2.3) for all $y \in D \cap Y$ and almost all $t \in I$, as well as

$$w(t_0, y_0) = x_0. \tag{2.4}$$

If a solution $\{x(t), y(t)\}$ of (2.1), (2.2) remains imbedded in the solution $w(t, y)$ of (2.3), (2.4), i.e., if $\{w(t, y(t)), y(t)\} \equiv \{x(t), y(t)\}$, then (2.1) and (2.3) are said to be equivalent; carrying the analogy further, we shall call (2.1) the characteristic equations of (2.3), while $\{x(t), y(t)\}$ will be a characteristic of the integral surface $w(t, y)$. Equivalence can be assured under the following condition.

THEOREM 2.1. *Let w be a solution of* (2.3), (2.4). *Suppose also that the equation $y' = H(t, w(t, y)y)$, $y(t_0) = y_0$ has a solution $y(t)$ over I. Then $x(t) \equiv w(t, y(t))$ and $y(t)$ are solutions of* (2.1), (2.2), *i.e., $\{x(t), y(t)\}$ is a characteristic of $w(t, y)$.*

For the proof we observe that the chain rule and (2.3) yield $x'(t) = w_t + w_y y'(t) = G(t, w, y)$ a.e. on I. Let us now turn to the two-point boundary value problem

$$\begin{aligned} x' &= G(t, x, y), \qquad g(x(t_0), y(t_0)) = 0, \\ y' &= H(t, x, y), \qquad h(x(t_1), y(t_1)) = 0, \end{aligned} \tag{2.5}$$

with $t_0, t_1 \in \bar{I}$. For definiteness we shall assume that $t_0 < t_1$. As described in [8] the existence of a solution for (2.5) may be considered from the following point of view.

THEOREM 2.2. *Assume that the Cauchy problem*

$$w_t(t, y) + w_y(t, y)H(t, w, y) = G(t, w, y), \qquad g(w(t_0, y), y) = 0 \tag{2.6}$$

has a solution over $[t_0, t_1] \times D \cap Y$. Suppose further that the set $S \subset D \cap Y$ of solutions of $h(w(t, y), y) = 0$ is not empty. If for some $\hat{y} \in S$ the problem $y' = H(t, w(t, y), y)$, $y(t_1) = \hat{y}$ has a solution over $[t_0, t_1]$, then $\{x(t) \equiv w(t, y(t)), y(t)\}$ is a solution of (2.5).

For the proof we observe that $\{x(t), y(t)\}$ is a characteristic of w. We remark that in connection with boundary value problems, the partial differential equation (2.6) is known as the invariant imbedding equation.

We are now ready to consider control problems. Subsequently, K will denote a given set of functions defined on I with values in Z. Our object is to find a function $\hat{u} \in K$ such that a functional $J(t_1, u)$ is minimized over K, i.e., $J(t_1, \hat{u}) = \inf_{u \in K} J(t_1, u)$, where $\{J(t, u), x(t, u)\}$ is a solution of the two-point boundary value problem

$$\begin{aligned} J' &= F(t, J, x, u), \qquad f(J(t_0), x(t_0)) = 0, \\ x' &= G(t, J, x, u), \qquad g(J(t_1), x(t_1)) = 0. \end{aligned} \tag{2.7}$$

Here F and G are suitably defined real- and Banach-space-valued functions on $I \times D \times Z$, while f and g implicitly represent the boundary conditions. (D is now taken to be a convex subset of $R \times X$.) Theorem 2.2 can be rephrased to obtain a characterization of the functional to be minimized.

THEOREM 2.3. *Suppose that for all $u \in K$:*

(i) *the Cauchy problem*

$$J_t(t, x, u) + J_x(t, x, u)G(t, J, x, u) = F(t, J, x, u), \qquad f(J(t_0, x, u), x) = 0$$

has a solution $J(t, x, u)$ over $[t_0, t_1] \times D \cap X$;

(ii) *the solution set* S *of* $g(J(t_1, x, u), x) = 0$ *is not empty*;
(iii) *the characteristic through* $(J(t, \hat{x}, u), \hat{x})$ *exists over* $[t_0, t_1]$ *for each* $\hat{x} \in S$.
Then a necessary condition for u *to be optimal is that*

$$J(t_1, u) = \inf_{\substack{u \in K \\ \hat{x} \in S}} J(t_1, \hat{x}, u). \tag{2.8}$$

The proof follows because the characteristic through $(J(t_1, \hat{s}, u), \hat{x})$ is a solution of (2.7) so that condition (2.8) is certainly necessary for optimality. We also remark that under appropriate smoothness conditions on F, G and the boundary vaiues, the integral surface $J(t, x, u)$ can be generated with its characteristic equations (for details we refer to [7]). In this case all solutions of (2.7) are necessarily imbedded in the integral surface $J(t, x, u)$ so that condition (2.8) is both necessary and sufficient for optimality.

For given $u \in K$, the function $J(t, x, u)$ defines a real-valued surface over the t, x-coordinate system. We now observe that if there exists a control $\hat{u} \in K$ such that $J_t(t, x, \hat{u}(t)) \leqq J_t(t, x, u(t))$ for all $u \in K$, then $\hat{u}$ is certainly optimal. In fact, since we are only interested in $J(t, \hat{x}, u)$ for some $\hat{x} \in S$ it suffices to require that $J(t, x, u)$ decrease maximally along the sections $x = \hat{x}$. Consequently, we can write the following theorem.

THEOREM 2.4. *Assume that Theorem* 2.3 *applies and that* S *contains only one element* $\hat{x}$. *Then* $\hat{u}$ *is optimal if it satisfies the dynamic programming equation*

$$J_t(t, x, \hat{u}) = \inf_{u \in K} [F(t, J, x, u) - J_x(t, x, u) G(t, J, x, u)]. \tag{2.9}$$

We remark that even when the solution $J(t, x, u)$ of the invariant imbedding equation can be generated over $[t_0, t_1]$ with the characteristics, the optimal control need not satisfy (2.9). Equation (2.9) is only meaningful if, given two controls $u_1(t)$, $u_2(t)$, we can synthesize a third control $u_3(t)$ such that

$$J_t(t, x, u_3(t)) = \min \{J_t(t, x, u_1(t)), J_t(t, x, u_2(t))\}.$$

The characterization of $J(t_1, x, u)$ remains valid if no such synthesis is possible.

Let us now briefly indicate how this development can be extended to controlled multipoint boundary value problems. For ease of exposition we shall restrict ourselves to the three-point problem:

$$\begin{aligned} J &= F(t, J, x, y, u), & f(J(t_0), x(t_0), y(t_0)) &= 0, \\ x &= G(t, J, x, y, u), & g(J(t_1), x(t_1), y(t_1)) &= 0, \\ y &= H(t, J, x, y, u), & h(J(t_2), x(t_2), y(t_2)) &= 0. \end{aligned} \tag{2.10}$$

We shall suppose that $t_0 \leqq t_1 \leqq t_2 \in \bar{I}$ and that F, G, and H are suitably defined functions on $I \times D \times Z$, where D is a convex subset of $R \times X \times Y$, while u again belongs to a set K of functions from I to Z. As described in detail in [8], the equations (2.10) can be associated with two Cauchy problems:

$$
\begin{aligned}
&J_t(t,x,y,u) + J_x(t,x,y,u)G(t,J,x,y,u) + J_y(t,x,y,u)H(t,J,x,y,u) \\
&\qquad = F(t,J,x,y,u), \\
&f(J(t_0,x,y,u),x,y) = 0, \qquad t \in [t_0,t_1],
\end{aligned} \tag{2.11}
$$

$$
\begin{aligned}
&J_t(t,y,u) + J_y(t,y,u)H(t,J,x,y,u) = F(t,J,x,y,u), \\
&x_t(t,y,u) + x_y(t,y,u)H(t,J,x,y,u) = G(t,J,x,y,u), \\
&J(t_1,y,u) = J(t_1,x(t_1,y,u),y,u), \quad g(J(t_1,y,u),x(t_1,y,u)) = 0, \\
&\qquad t \in [t_1,t_2].
\end{aligned} \tag{2.12}
$$

In fact, if $J(t,x,y,u)$ exists over $[t_0,t_1] \times D \cap (X \times Y)$, and if $J(t,y,u)$ and $x(t,y,u)$ exist over $[t_1,t_2] \times D \cap Y$, and if the solution set S of $h(J(t_2,y,u), x(t_2,y,u)) = 0$ is not empty, and furthermore, if the characteristic through $(J(t_2,\hat{y},u), x(t_2,\hat{y},u), \hat{y}) \in D$ for $\hat{y} \in S$ exists over $[t_0,t_2]$, then this characteristic is a solution of (2.10). Under these conditions (2.10) and (2.11), (2.12) are equivalent, and the optimum control $\hat{u}$ is characterized by

$$
J(t_2,\hat{u}) = \inf_{\substack{u \in K \\ \hat{y} \in S}} J(t_2,\hat{y},u),
$$

where $J(t,u)$ and $J(t,y,u)$ are the solutions of (2.10) and (2.12), respectively. Moreover, if S consists of one element $\hat{y}$, we can again consider the decrease of J along the section $y = \hat{y}$ and obtain a sufficient condition for optimality.

THEOREM 2.5. *Assume that for all $u \in K$, the solution of* (2.10) *is imbedded in the integral surfaces $J(t,x,y,u)$ of* (2.11) *and $\{J(t,y,u), x(t,y,u)\}$ of* (2.12). *Then $\hat{u}$ is optimal if it satisfies the dynamic programming equations*

$$
\begin{aligned}
J_t(t,x,\hat{y},u) = \inf_{u \in K} \{&F(t,J,x,\hat{y},u) - J_x(t,x,\hat{y},u)G(t,J,x,\hat{y},u) \\
&- J_y(t,x,\hat{y},u)H(t,J,x,\hat{y},u)\}
\end{aligned}
$$

for $(t,x) \in [t_0,t_1] \times D \cap X$, and

$$
J_t(t,\hat{y},u) = \inf_{u \in K} \{F(t,J,x,\hat{y},u) - J_y(t,\hat{y},u)H(t,J,x(t,\hat{y},u)\hat{y},u)\}, \qquad t \in [t_1,t_2].
$$

As an illustration consider the scalar control problem:

$$
\begin{aligned}
J' &= F(t,J,x,y,u), \qquad J(t_0) = \alpha, \\
x' &= G(t,J,x,y,u), \qquad x(t_1) = \beta, \\
y' &= H(t,J,x,y,u), \qquad y(t_2) = \gamma.
\end{aligned}
$$

We shall assume that F, G, and H are continuously differentiable in the real variables J, x, and y, and that for fixed (J,x,y), the functions F, G, and H and their derivatives with respect to J, x, and y are bounded and Lebesgue measurable in t on $[t_0,t_2]$ for all $u \in K$. Using the concept of a Carathéodory solution one can readily verify that the construction of the integral surface $J(t,x,y,u)$ of (2.11) through the initial manifold $J(t_0,x,y,u) = \alpha$ with a shooting method remains

valid (see also [7]). Hence J exists near the initial manifold; it is continuously differentiable with respect to x and y and absolutely continuous in t. Moreover, if the derivatives with respect to J, x, and y are uniformly bounded, then a quantitative estimate of the interval of existence for $J(t, x, y, u)$ can be given. Let us suppose that $J(t, x, y, u)$ exists for all x, y and $t \in [t_0, t_1]$. We then have to solve the invariant imbedding equations (2.12) subject to the initial values

$$J(t_1, y, u) = J(t_1, \beta, y, u), \qquad x(t_1, y, u) = \beta.$$

If we also suppose that $J(t, y, u)$ and $x(t, y, u)$ exist over $[t_1, t_2]$ for all y and all $u \in K$, then the optimal control u is found from the dynamic programming equations

$$J_t(t, x, \gamma, \hat{u}) = \inf_{u \in K}\{F(t, J, x, \gamma, u) - J_x(t, x, \gamma, u)G(t, J, x, \gamma, u) - J_y(t, x, \gamma, u)H(t, J, x, \gamma, u)\}$$

and

$$J_t(t, \gamma, \hat{u}) = \inf_{u \in K}\{F(t, J, x(t, \gamma, u), \gamma, u) - J_y(t, \gamma, u)H(t, J, x(t, \gamma, u), \gamma, u)\}.$$

It is apparent from the results of [7] that this discussion can be carried over immediately to functions on arbitrary Banach spaces which are Fréchet differentiable in J, x, and y and Bochner integrable with respect to t. The main difficulty in applying the theory of dynamic programming to unbounded functions lies in verifying the equivalence between the integral surface and the corresponding characteristic.

3. Linear problems. The dynamic programming equations assume a simple form for linear controlled equations because the solution of the underlying boundary value problem and the choice of the optimal control can be separated. In order to describe this simplification let us consider the system

$$\begin{aligned} J' &= A_{11}(t)J + A_{12}(t)x + F(t, u), \qquad J(t_0) = fx(t_0) + \alpha, \\ x' &= A_{21}(t)J + A_{22}(t)x + G(t, u), \qquad g(J(t_1), x(t_1)) = 0, \end{aligned} \tag{3.1}$$

where the A_{ij} and f are suitably defined linear functions or operators on R or $D \subset X$, and where F and G are a scalar- and a vector-valued function, respectively. g is a (not necessarily linear) vector-valued function denoting the boundary condition at t_1. It was shown in detail [7] that for fixed $u \in K$ the invariant imbedding equation (2.6) corresponding to (3.1) has the solution

$$J(t, x, u) = T(t)x + w(t, u), \tag{3.2}$$

where T and w are solutions of the ordinary differential equations

$$T' = A_{12}(t) + A_{11}(t)T - TA_{22}(t) - TA_{21}(t)T, \qquad T(t_0) = f, \tag{3.3}$$

$$w' = [A_{11}(t) - T(t)A_{21}(t)]w - T(t)G(t, u) + F(t, u), \qquad w(t_0) = \alpha. \tag{3.4}$$

A solution of the Riccati equation (3.3) is a function defined on $[t_0, t_1]$ with values in X', the dual of X, such that $T(t)x$ is differentiable for almost all $t \in (t_0, t_1)$ and all $x \in D$. We note that only the solution w of the scalar equation (3.4) is affected

by the choice of the control $u \in K$; hence Theorem 2.4 can be restated for the linear system (3.1) as follows.

THEOREM 3.1. *Suppose that the Riccati operator equation* (3.3) *and the scalar equation* (3.4) *have solutions* T *and* w *over* $[t_0, t_1]$ *for all* $u \in K$. *Assume further that there exists a unique solution* $\hat{x} \in D$ *of the equation* $g(T(t_1)x + w(t_1u), x) = 0$ *and that the (backward) initial value problem* $x' = [A_{21}(t)T(t) + A_{22}(t)]x + A_{21}(t)w(t, u) + G(t, u)$, $x(t_1) = \hat{x}$ *has a solution* $x(t) \in D$ *for all* $u \in K$. *Then* $\hat{u}$ *is optimal if it satisfies*

$$\varphi(\hat{u}) = \inf_{u \in K} \{-T(t)G(t, u) + F(t, u)\} \tag{3.5}$$

for all $t \in (t_0, t_1)$.

Proof. Under the hypotheses of the theorem, (3.1) and the corresponding invariant imbedding equation are equivalent so that for fixed $u \in K$ the solution of (3.1) is imbedded in $T(t)x + w(t, u)$. The optimality condition follows by observing that $w(t, u)$ has the representation

$$w(t, u) = \varphi(t, t_0) + \int_{t_0}^{t} \varphi(t, r)[F(r, u(r)) - T(r)G(r, u(r))]\, dr,$$

where $\varphi(t, r) = \exp\left[\int_r^t [A_{11}(s) - T(s)A_{21}(s)]\, ds\right] > 0$. Relation (3.5) can also be obtained by substituting the representation (3.2) into the dynamic programming equation and neglecting all terms not depending on u.

Frequently, the so-called state equation for the vector x is subject to an initial rather than a final condition, while $J(0, u)$ is to be minimized. Interchanging t_0 and t_1 and the boundary conditions we see that in this case the Riccati and the linear equation are subject to

$$T(t_1) = f, \qquad w(t_1) = \alpha, \tag{3.6}$$

while $J(t, x, u)$ will decrease maximally along the section $x = \hat{x}$ (the solution of $g(T(0)x + w(0, u), x) = 0$) if $J_t(t, x, \hat{u}) = \sup_{u \in K} J_t(t, \hat{x}, u)$ or $\varphi(\hat{u}) = \sup_{u \in K} [-T(t)G(t, u) + F(t, u)]$. It follows from Theorem 3.1 that necessary and sufficient conditions for the optimum control can be derived in two steps. First we verify that the hypotheses of Theorem 3.1 obtain; then we show that the dynamic programming equation has a solution. Existence of solutions for "square" Riccati equations has been discussed in the literature (see [8] and the references cited there). Since $T(t)$ for fixed t belongs to X' rather than $L(X, X)$, the space of bounded linear operators from X to X, previous discussions do not apply. Under certain conditions, however, a local result is obtainable. For this development we need the following result due to Kato [6].

THEOREM 3.2. *For each* $t \in [0, t_1]$, $A(t)$ *is a densely defined closed linear operator in* X, *and its spectrum is contained in a fixed sector* $S_\theta : |\arg z| \leqq \theta < \pi/2$. *The resolvent of* $A(t)$ *satisfies the inequality* $\|(z - A(t))^{-1}\| \leqq M_0/|z|$ *for* $z \notin S_0$, *where* M_0 *is a constant independent of* t. *Furthermore,* $z = 0$ *also belongs to the resolvent set of* $A(t)$ *and*

$$\|A(t)^{-1}\| \leqq M_1,$$

M_1 being independent of t. Furthermore, for some $h = 1/m$, where m is a positive integer, the domain $D[A(t)^h] \equiv D$ is independent of t, and there exist constants k, M_2, M_3 such that

$$\|A(t)^h A(s)^{-h}\| \leqq M_2, \qquad 0 \leqq t \leqq t_1,$$
$$\|A(t)^h A(s)^{-h} - 1\| \leqq M_3 |t - s|^k, \qquad 0 \leqq s \leqq t_1, \quad 1 - h < k \leqq 1.$$

Then there exists a unique (parabolic) evolution operator $U(t, s) \in L(X, X)$ defined for $0 \leqq s \leqq t \leqq T$, with the following properties. $U(t, s)$ is strongly continuous for $0 \leqq s \leqq t \leqq T$ and

$$U(t, r) = U(t, s)U(s, r), \qquad r \leqq s \leqq t, \quad U(t, t) = I.$$

For $s < t$ the range of $U(t, s)$ is a subset of $D[A(t)]$ and

$$A(t)U(t, s) \in L(X, X), \qquad \|A(t)U(t, s)\| \leqq M|t - s|^{-1},$$

where M is a constant depending only on $\theta, h, k, T, M_0, M_1, M_2$, and M_3. Furthermore, $U(t, s)$ is strongly continuously differentiable in t for $t > s$ and

$$\frac{\partial U}{\partial t}(t, s) + A(t)U(t, x) = 0.$$

If $u \in D$, then $U(t, s)u$ is strongly differentiable in s for $s < t$ and $(\partial U/\partial s)(t, x)u = U(t, x)A(s)u$.

We remark that the special case of $h = 1$ is discussed in detail in Yosida [9, p. 431]. Theorem 3.2 and various estimates derived during its proof in [6] will be used to give a local existence theorem for the Riccati equation (3.3). For ease of comparison with published results we shall assume that the initial conditions (3.6) hold.

THEOREM 3.3. *Assume that*

(i) *$-A_{22}(t)$ is the generator of a parabolic evolution operator $U(t, r)$ on $[t_0, t_1]$ with time-independent domain $D \subset X$, and*

(ii) *A_{11}, A_{12}, and A_{21} are Lipschitz continuous in t on $[t_0, t_1]$.*

Then the Riccati equation (3.3) subject to $T(t_1) = f$ has a unique solution provided $t_1 - t_0$ is sufficiently small.

Proof. If $x \in D$, then it is readily verified that any strongly differentiable solution $T(t)s$ of (3.3) is also a solution of the integral equation on X' for $t_0 \leqq t \leqq t_1$:

$$\begin{aligned} T(t) = {} & \exp\left[\int_{t_1}^{t} A_{11}(s)\,ds\right] f U(t_1 t) \\ & + \int_{t_1}^{t} \exp\left[\int_{r}^{t} A_{11}(s)\,ds\right]\left[A_{12}(r) - T(r)A_{21}(r)T(r)\right] U(r, t)\,dr, \end{aligned} \tag{3.7}$$

where the integral is interpreted as a Riemann integral.

Conversely, since U is uniformly bounded on $[t_0, t_1] \times [t_0, t_1]$ and X' is complete, we can find a strongly continuous solution T of (3.7) by successive substitution provided $t_1 - t_0$ is taken sufficiently small. To show that $T(t)x$ is differentiable

for $t \in (t_0, t_1)$ and $x \in D$ we consider $\lim_{\Delta t \to 0} ((T(t - \Delta t) - T(t))/t)x$. From the chain rule and the properties of U it follows that $T(t)x$ is differentiable provided the integral

$$\int_{t_1}^{t} \exp\left[\int_{r}^{t} A_{11}(s)\,ds\right][A_{12}(r) - T(r)A_{21}(r)T(r)]U(r,t)A(t)x\,ds$$

is continuous. However, since $x \in D$ we can write $x = A^{-1}(s)A(s)x$ for arbitrary $s \in [t_0, t_1]$. From $\|A(s)x\| = \|[A(s) - A(t_0) + A(t_0)]A(t_0)^{-1}A(t_0)x\| \leqq K\|A(t_0)x\|$ and $\|A(t)x - A(s)x\| \leqq \|[A(t) - A(s)]A^{-1}(s)\|\,\|A(s)x\| \leqq M_3|t - s|^k\|A(s)x\|$ it follows that $A(t)x$ is continuous in t. Hence the integrand is a continuous function on $[t_0, t_1]$ for each $x \in D$. Since it also is continuous in r we see that the integral is continuous in t and that $T(t) = x$ is strongly differentiable for each $x \in D$.

Furthermore, standard existence theory for ordinary differential equations yields the following result.

THEOREM 3.4. *Assume that $G(t, u(t))$ and $F(t, u(t))$ are Lipschitz continuous on $[t_0, t_1]$ for all $u \in K$. Then the linear scalar equation* (3.4) *has a unique solution $w(t, u)$ on $[t_0, t_1]$.*

THEOREM 3.5. *The characteristic equation*

$$x' = [A_{21}(t)T(t) + A_{22}(t)]x + A_{21}(t)w(t) + G(t, u(t)), \qquad x(t_0) = \hat{x} \tag{3.8}$$

has a strongly differentiable solution $x(t) \in D$ for $t \in [t_0, t_1]$.

Proof. It can be shown that the equation $z' = [A_{21}(t)T(t) + A_{22}(t)]z$, $z(t_0) = \hat{x}$ has the solution $z(t) = V(t, t_0)x$, where $V(t, r)$ again is a parabolic evolution operator. Indeed, for fixed $t \in (t_0, t_1)$, the operator $P(t) \equiv A_{21}(t)T(t)$ is bounded so that $C(t) \equiv A_{22}(t) + P(t)$ is the generator of a holomorphic semigroup for each t on $[t_0, t_1]$. The condition $\|C(t)^{-1}\| \leqq M$ can always be satisfied by a change of variable $z = e^{\beta t}y$. Finally, we observe that $U(t, s)A_{22}(r)^{-1}$ is Lipschitz continuous (see [6]) so that $P(t)A_{22}(r)^{-1}$ is also Lipschitz continuous. Consequently, $\|[I + P(s)A_{22}(s)^{-1}]^{-1}\| = \|[A_{22}(s) + P(s) - P(s)][A_{22}(s) + P(s)]^{-1}\| \leqq 1 + \|P\|\,\|C\|$ and $\|C(t)C(s)^{-1} - I\| = \|\{A_{22}(t) - A_{22}(s) + P(t) - P(s)\}A_{22}(s)^{-1} \cdot [I + P(s)A_{22}(s)^{-1}]^{-1}\|$ show that $\|C(t)C(s)^{-1} - I\| \leqq M|t - s|^k$ for some constant M. The existence of the evolution operator follows from Theorem 3.2. The existence of a strong solution $x(t) \in D$ follows from the variation of constants representation

$$x(t) = V(t, t_0)\hat{x} + \int_{t_0}^{t} V(t, r)[A_{21}(r)w(r) + G(r, u(r))]\,dr \tag{3.9}$$

and the Lipschitz continuity of $A_{21}(r)w(r) + G(r, u(r))$ (see again [6]).

Under the hypotheses of the preceding theorems the functions $J(t) = T(t)x + w(t, u)$ and $x(t)$ are a solution of (3.1). However, the continuity conditions imposed on F and G are too stringent for controlled systems. They can be relaxed by considering mild solutions of evolution equations. Such solutions are required to be strongly continuous, but not necessarily differentiable, and to satisfy an integral equation equivalent to the differential equation. For example, any strongly continuous function T from $[t_0, t_1]$ to X satisfying (3.7) is a mild solution of the Riccati equation (3.3). Similarly, any strongly continuous solution $x(t)$

of (3.9) is a mild solution of (3.8). The integrals of (3.7) and (3.9) may now be interpreted as Bochner integrals.

THEOREM 3.6. *Let $A_{22}(t)$ generate a parabolic evolution operator on $[t_0, t_1]$. Let $\{J(t,u), x(t,u)\}$ be a mild solution of* (3.1) *and let $T(t)$, $w(t,u)$, and $x(t)$ be mild solutions of* (3.3), (3.4), *and* (3.8). *If there exists a sequence of piecewise Lipschitz continuous functions $A_{11}^n, A_{12}^n, A_{21}^n, F^n$ and G^n which converge to $A_{11}, A_{12}, A_{21}, F$, and G, uniformly in t, then $J(t,u) = T(t)x(t) + w(t,u)$.*

Proof. Let $\{J^n(t,u^n), x^n(t,u^n)\}$ and $T^n(t), w^n(t,u^n), x^n(t)$ be the solutions of (3.1), (3.3), (3.4), and (3.8), where A_{ij} is replaced by A_{ij}^n, F by F^n, and G by G^n. It follows from Theorems 3.2–3.5 that these solutions are continuous and piecewise differentiable on $[t_0, t_1]$ so that $J^n(t,u^n) = T^n(t)x^n(t) + w^n(t,u^n)$. Moreover, from the integral equations it is clear that $J^n \to J$, $x^n \to x$, $T^n \to T$, and $w^n \to w$ uniformly in t as $n \to \infty$ so that $J(t,u) = \lim_{n\to\infty} J^n(t,u^n) = \lim_{n\to\infty} T^n(t)x^n(t) + w^n(t,u^n) = T(t)x(t) + w(t,u)$.

Let us now turn to the existence of optimal controls. We have seen that such controls have to minimize the expression

$$w(t_0,u) = \varphi(t_0,t_1) + \int_{t_1}^{t_0} \varphi(t_0,r)[F(r,u(r)) - T(r)G(r,u(r))]\,dr,$$

where $\varphi(t,r)$ is a positive function. We observe that

$$\psi(u) = \int_{t_0}^{t_1} \varphi(t_0,r)[T(r)G(r,u(r)) - F(r,u(r))]\,dr$$

may be interpreted as a functional defined on K. The existence of a minimizing control for such functionals is the subject of the following theorem due to Balakrishnan [1].

THEOREM 3.7. *Let the set K (of functions on $[t_0, t_1]$) be a closed, bounded convex subset of a reflexive Banach space and assume that the functional ψ is continuous and convex in u. Then there exists an element $\hat{u} \in K$ such that*

$$\psi(\hat{u}) = \inf_{u\in K} \psi(u).$$

We remark that in connection with controlled parabolic equations, the set K is frequently chosen to be a subset of the space of square Bochner integrable functions on $[t_0, t_1]$. If X is reflexive, this function space is also reflexive. Once K has been chosen, the hypotheses of Theorem 3.7 are generally easy to verify from (3.5).

In order to illustrate the applicability of our linear theory let us consider the following example discussed by Friedman [4].

$$\begin{aligned} J' &= -\beta f_0(t)x, & J(t_1) &= \alpha f x(t_1), & &\alpha \text{ and } \beta \text{ constant}, \\ x' &= -A(t)x + u(t), & x(0) &= x_0, \end{aligned} \tag{3.10}$$

where f_0 is a continuous function from $[0, t_1]$ to X', where $A(t)$ is the generator of a parabolic evolution operator, and where K, the set of admissible controls, is the set of all Bochner integrable functions from $[0, t_1]$ with values in a subset $U \subset X$. The equations (3.10) are seen to be a particular form of system (3.1).

It is well known that under these conditions the state equation has the mild solution

$$x(t) = U(t,0)x_0 + \int_0^t U(t,r)u(r)\,dr$$

for every $u \in K$. Since the set of simple functions is dense in the space of Bochner integrable functions (see [5, p. 86]), it follows from Theorem 3.6 that $J(t,u) = T(t)x(t) + w(t,u)$, where $x(t)$ is the above mild solution, and where T and w are mild solutions of

$$T' = -f_0(t) + TA(t), \qquad T(t_1) = f,$$
$$w' = -T(t)u, \qquad w(t_1) = 0.$$

The dynamic programming equation (3.5) requires that u be chosen such that

$$\varphi(\hat{u}) = \sup_{u \in K} (-T(t)u(t)). \tag{3.11}$$

Moreover, since the mild solutions T, x, and w exist for all Bochner integrable functions u, the solution of the boundary value problem (3.10) is necessarily imbedded in $J(t,x,u) = T(t)x + w(t,u)$. Hence (3.11) is both necessary and sufficient for optimality. We remark that (3.11) was derived in [4] as a necessary condition for optimality.

If the existence of an optimal control can be postulated, then our formalism can be used to derive necessary conditions for optimality. Suppose, for example, that the state equation of (3.10) is used to minimize the expression

$$J(u) = \|x(t_1,u) - y\|,$$

where y is a given element in X. Assume that no trajectory exists such that $x(t_1) = y$. Let $\hat{u}$ be the optimal control and $\hat{x}$ the associated trajectory. Then there exists a linear functional f such that $\|f\| = 1$ and $f(\hat{x}(t_1) - y) = \|\hat{x}(t_1) - y\|$. If we set $f_0 = 0$ in (3.10), we obtain the control problem

$$J' = 0, \qquad J(t_1) = f(x - y),$$
$$x' = -A(t)x + u, \qquad x(0) = x_0.$$

We see that in this case the Riccati equation (3.3) reduces to

$$T' = TA(t), \qquad T(t_1) = f,$$

while the dynamic programming equation leads to the following necessary condition for optimality:

$$\varphi(\hat{u}) = \sup_{u \in K} (-T(t)u(t)).$$

This again is exactly the maximum principle derived by Friedman.

Similarly, we may minimize the expression $J(u) = \int_0^{t_1} \|x(t) - y(t)\|\,dt$,

where y is a given function from $[0, t_1]$ to X. If again the optimal trajectory $\hat{x}(t, \hat{u})$ is assumed to exist, then we can define a functional $f_0(t)$ such that

$$f_0(t)(x(t) - y(t)) = -\|x(t) - y(t)\| \quad \text{for all } t \in [0, t_1].$$

If we assume that X is a Hilbert space, then f is continuous in t. We now set $f = 0$ in (3.10) and solve the control problem

$$J' = f_0(t)x - f_0(t)y(t), \qquad J(t_1) = 0,$$
$$x' = -A(t)x + u, \qquad x(0) = x_0.$$

In this case the Riccati equation reduces to

$$T' = f_0(t) + TA(t), \qquad T(t_1) = 0,$$

and the dynamic programming equation becomes

$$\varphi(\hat{u}) = \sup(-T(t)u(t)).$$

This maximum also was derived earlier by Friedman [4].

To conclude our discussion, let us formally derive the dynamic programming equations for linear multipoint control problems. Suppose the equations are as follows:

$$J' = A_{11}J + A_{12}x + A_{13}y + F(u), \qquad J(t_0) = \alpha,$$
$$x' = A_{21}J + A_{22}x + A_{23}y + G(u), \qquad x(t_1) = \beta,$$
$$y' = A_{31}J + A_{32}x + A_{33}y + H(u), \qquad y(t_2) = \gamma,$$

where $J(t_2, u)$ is to be minimized. As described in [8], the invariant imbedding equation (2.11) has the solution

$$J(t, x, y, u) = T_{11}(t)x + T_{12}(t)y + w_1(t, u),$$

where $(T_{11} \quad T_{12})$ and w_1 are solutions of

$$(T_{11} \quad T_{12})' = (A_{12} \quad A_{13}) + A_{11}(T_{11} \quad T_{12})$$
$$- (T_{11} \quad T_{12})\begin{pmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{pmatrix} - (T_{11} \quad T_{12})\begin{pmatrix} A_{21} \\ A_{31} \end{pmatrix}(T_{11} \quad T_{12}),$$
$$(T_{11} \quad T_{12})(0) = (0 \quad 0),$$
$$w_1' = \left[A_{11} - (T_{11} \quad T_{12})\begin{pmatrix} A_{21} \\ A_{31} \end{pmatrix}\right]w_1 - (T_{11} \quad T_{12})\begin{pmatrix} G(u) \\ H(u) \end{pmatrix} + F(u), \qquad w_1(t_0) = \alpha.$$

Hence the maximal decrease of $J(t, x, y, u)$ along the section $y = \gamma$ is assured if $\hat{u}$ is chosen such that

$$\varphi(\hat{u}) = \inf_{u \in K}[-T_{11}(t)G(u) - T_{12}(t)H(u) + F(u)] \quad \text{for } t \in [t_0, t_1].$$

Over $[t_1, t_2]$ we have the representation $\begin{pmatrix} J(t, y, u) \\ x(t, y, u) \end{pmatrix} = \begin{pmatrix} T_{21}(t) \\ T_{22}(t) \end{pmatrix}y + \begin{pmatrix} W_2(t, u) \\ W_3(t, u) \end{pmatrix}$,

where

$$\begin{pmatrix} T_{21} \\ T_{22} \end{pmatrix}' = \begin{pmatrix} A_{13} \\ A_{23} \end{pmatrix} + \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} T_{21} \\ T_{22} \end{pmatrix} - \begin{pmatrix} T_{21} \\ T_{22} \end{pmatrix} A_{33} - \begin{pmatrix} T_{21} \\ T_{22} \end{pmatrix}(A_{31} \quad A_{32})\begin{pmatrix} T_{21} \\ T_{22} \end{pmatrix},$$

$$\begin{pmatrix} T_{21} \\ T_{22} \end{pmatrix}(t_1) = \begin{pmatrix} T_{12}(t_1) \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} W_2 \\ W_3 \end{pmatrix}' = \left[\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} - \begin{pmatrix} T_{21} \\ T_{22} \end{pmatrix}(A_{31} \quad A_{32})\right]\begin{pmatrix} W_2 \\ W_3 \end{pmatrix} - \begin{pmatrix} T_{21}(t) \\ T_{22}(t) \end{pmatrix} H(u) + \begin{pmatrix} F(u) \\ G(u) \end{pmatrix},$$

$$\begin{pmatrix} W_2 \\ W_3 \end{pmatrix}(t_2) = \begin{pmatrix} T_{11}(t_1) & +w_1(t_1) \\ & 0 \end{pmatrix}.$$

Minimization now requires that $\hat{u}$ be chosen such that

$$\varphi(\hat{u}) = \inf_{u \in K} \{U_{11}(t_2, r)[F(u) - T_{21}(r)H(u)] + U_{12}(t_2, r)G(U) - T_{22}(r)H(u)\},$$

where $U(t, r)$ is the fundamental matrix generated by

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} - \begin{pmatrix} T_{21} \\ T_{22} \end{pmatrix}(A_{31} \quad A_{32}).$$

Similarly, we can derive maximum principles for controlled linear N-point boundary value problems.

Acknowledgment. The author wishes to thank Mobil Research and Development Corporation for permission to publish this paper.

REFERENCES

[1] A. V. BALAKRISHNAN, *Optimal control problems in Banach spaces*, this Journal, 3 (1965), pp. 152–180.

[2] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, N.J., 1957.

[3] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. II, Interscience, New York, 1962.

[4] A. FRIEDMAN, *Optimal control in Banach spaces*, J. Math. Anal. Appl., 19 (1967), pp. 35–55.

[5] E. HILLE AND R. E. PHILLIPS, *Functional Analysis and Semigroups*, rev. ed., Colloquium Publications, vol. 31, American Mathematical Society, Providence, 1957.

[6] T. KATO, *Abstract evolution equations of parabolic type in Banach and Hilbert spaces*, Naboya Math. J., 19 (1961), pp. 93–125.

[7] G. H. MEYER, *On a general theory of characteristics and the method of invariant imbedding*, SIAM J. Appl. Math., 16 (1968), pp. 488–509.

[8] ———, *The invariant imbedding equations for multipoint boundary value problems*, Ibid., 18 (1970), pp. 433–453.

[9] K. YOSHIDA, *Functional Analysis*, Springer-Verlag, Berlin, 1965.

SIAM J. CONTROL
Vol. 8, No. 3, August 1970

A NEW ITERATIVE PROCEDURE FOR THE MINIMIZATION OF A QUADRATIC FORM ON A CONVEX SET*

THOMAS PECSVARADI AND KUMPATI S. NARENDRA†

Introduction. An iterative procedure is developed in this paper for computing the minimum of a quadratic form over a convex set R in Euclidean n-space. The problem was originally posed and solved by Gilbert [1] who suggested minimizing the quadratic form on a suitably chosen series of line segments in R. Barr [2] extended Gilbert's procedure so that the minimization is carried out on a sequence of convex polyhedra. The iterative procedure presented in this paper was proposed by the authors in [3] as an alternative to Barr's procedure in [2].

After submitting this paper for publication, it was brought to the authors' attention by the Editor that Barr had, in a subsequent work [4] in this Journal, suggested two new selection rules for use in his iterative procedure. Since there is considerable overlap between Barr's procedure and that outlined here in general approach as well as in some specific details, the main objective of this paper is to indicate the essential differences between the two procedures and to provide numerical data for comparison purposes. At the referee's suggestion the proof of the theorem has not been included and the reader is referred to [4].

1. Statement of the problem. The following notation will be used throughout the paper: E^n denotes the Euclidean n-space; if $x, y \in E^n$, then their inner product is denoted by $\langle x, y\rangle$ and the Euclidean norm of x, $\|x\|$, is given by $\|x\| = \sqrt{\langle x, x\rangle}$; if $y_1, y_2, \cdots, y_p \in E^n$, then their closed convex hull is denoted by $\Delta(y_1, y_2, \cdots, y_p)$.

Let R be a compact convex set and α a fixed point in E^n. Let $y \in R$ and $h = \alpha - y$. Using similar definitions as in [1], we let $\eta(h) = \max_{z\in R}\langle h, z - y\rangle$ denote a support function of R and

$$P(h) = \{x; \langle h, x - y\rangle = \eta(h)\}, \tag{1}$$

the support hyperplane of R with outward normal h for any $h \neq 0$ (see Fig. 1). A contact point $s(h) \in R$ is defined by the relationship $\langle h, s(h) - y\rangle = \eta(h)$, $h \neq 0$. Thus

$$s(h) \in P(h) \cap R. \tag{2}$$

Clearly $s(h)$ is unique if and only if R is strictly convex.

A basic requirement of all the algorithms developed is that there exist a method for evaluating $s(h)$ for any $h \neq 0$. If for some $h \neq 0$ $s(h)$ is not unique, then any value of $s(h)$ may be used.

The problem. Given a compact convex set R and a fixed point α in E^n, determine a point $z^* \in R$ such that

$$\|\alpha - z^*\| = \min_{z\in R} \|\alpha - z\|. \tag{3}$$

* Received by the editors June 24, 1969, and in final revised form December 26, 1969.

† Department of Engineering and Applied Science, Yale University, New Haven, Connecticut 06520.

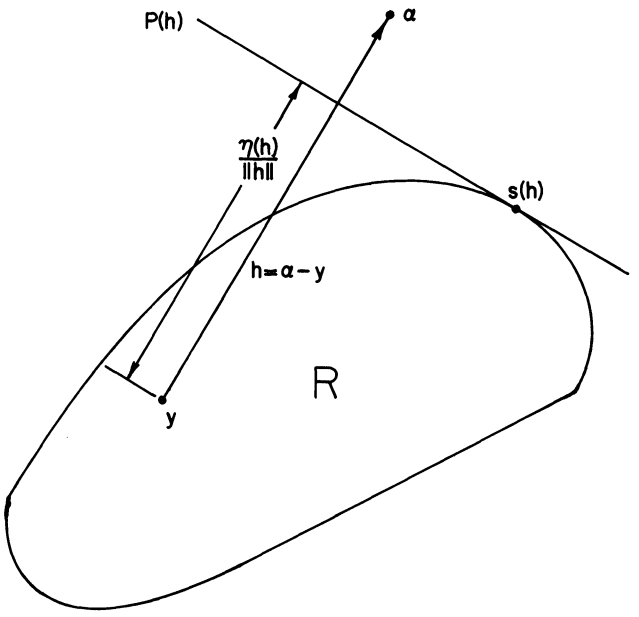


FIG. 1. *Support plane P(h) and contact point s(h)*

A solution z^* exists and is unique because R is compact and convex and $\|\alpha - z\|$ is a continuous, strictly convex function of z. Furthermore $z^* = \alpha$ if and only if $\alpha \in R$; if $\alpha \notin R$, then $z^* \in \partial R$. Finally, if $\alpha \notin R$, then $z = z^*$ if and only if $z \in P(\alpha - z) \cap R$.

2. The iterative procedure. The minimization of the quadratic form over the given convex set R is accomplished using an iterative technique. At any stage k the minimization is carried out over $Q_k \subset R$ $(k < n)$ or $S_k \subset R$ $(k \geqq n)$, where Q_k and S_k are convex hulls of at most $n + 1$ points belonging to R. The Q_k are of dimension less than n, while the S_k are generally n-dimensional simplexes. The principal objective of the procedure is to choose the Q_k and S_k to achieve rapid convergence.

(i) Q_k $(1 \leqq k < n)$: Let $y_0, y_1, \cdots, y_{k-1}$ be k known points in R and denote their convex hull by $Q_k \triangleq \Delta(y_0, y_1, \cdots, y_{k-1})$. Let $z_{k-1} \in Q_k$ be the point such that

$$\|\alpha - z_{k-1}\| = \min_{z \in Q_k} \|\alpha - z\|. \tag{4}$$

Define $h_k \triangleq \alpha - z_{k-1}$ and let

$$y_k = s(h_k), \tag{5}$$

where $s(h_k)$ is a contact point of R corresponding to h_k. Then Q_{k+1} is defined as the convex hull

$$Q_{k+1} \triangleq \Delta(y_k, Q_k). \tag{6}$$

(ii) Q_k, S_k $(k \geqq n)$: Let $y_0, y_1, \cdots, y_{n-1}$ be n known points in R and denote their convex hull by $Q_k \triangleq \Delta(y_0, y_1, \cdots, y_{n-1})$. Define z_{k-1} and h_k as above and let

$$y_n = s(h_k). \tag{7}$$

The convex hull S_k is defined as

$$S_k \triangleq \Delta(y_n, Q_k) \tag{8}$$

and is an n-simplex if $y_0, y_1, \cdots, y_n$ do not lie on a hyperplane. The point $z_k \in S_k$ is determined by

$$\|\alpha - z_k\| = \min_{z \in S_k} \|\alpha - z\|. \tag{9}$$

Finally, define the hyperplane P_{k+1} as

$$P_{k+1} \triangleq \{x; \langle \alpha - z_k, z_k - x \rangle = 0\} \tag{10}$$

and let d_i, $i = 0, 1, \cdots, n-1$, be the Euclidean distance from y_i to P_{k+1}, viz.

$$d_i \triangleq \min_{x \in P_{k+1}} \|x - y_i\|, \qquad i = 0, 1, \cdots, n-1. \tag{11}$$

Assume that $\max_{0 \leqq i < n} d_i$ occurs for $i = m$. Then the point y_m is replaced by y_n, and Q_{k+1} is defined as

$$Q_{k+1} \triangleq \Delta(y_0, y_1, \cdots, y_{n-1}). \tag{12}$$

(If $\max_{0 \leqq i < n} d_i$ occurs for several i, then one of these y_i is chosen arbitrarily and is replaced by y_n.) It is clear that if $y_0, y_1, \cdots, y_n$ do not lie on a hyperplane, then $z_k \neq z_{k-1}$, and $z_k \in Q_{k+1}$. To start the process choose any point $y_0 \in R$. Then $z_0 = Q_1 = y_0$. At every stage k of the iterative process the following bounds exist for $\|\alpha - z^*\|$:

$$\max\left\{0, \frac{\langle h_k, \alpha - y_p \rangle}{\|h_k\|}\right\} \leqq \|\alpha - z^*\| \leqq \|h_k\|, \tag{13}$$

where $p = k$ if $k < n$, and $p = n$ if $k \geqq n$.

THEOREM. *The sequence $\{z_k\}$ generated by the iterative procedure is such that:*

(i) $z_k \in R$ *for* $k = 1, 2, \cdots$;

(ii) $\|\alpha - z_k\| \leqq \|\alpha - z_{k-1}\|$, *where the equality sign holds if and only if* $z_k = z_{k-1} = z^*$;

(iii) $z_k \to z^*$.

3. Comparison with Barr's procedure. In the iterative procedure, IP, suggested by Barr in [2] and [4] the norm $\|\alpha - z\|^2$ is minimized at the kth stage over the convex polyhedron

$$H_k = \Delta(y_1(k), y_2(k), \cdots, y_p(k), s(\alpha - z_k), z_k), \tag{14}$$

where p is an arbitrary positive integer chosen prior to implementing the procedure. The points $y_1(k), y_2(k), \cdots, y_p(k)$ belong to the set R, and the three selection rules A, B and C of [4] indicate how they are to be chosen.

The procedure described in this paper minimizes the norm $\|\alpha - z\|^2$ for $k < n$ over a sequence of successively higher dimensional polyhedra

$$Q_k = \Delta(y_0, y_1, \cdots, y_{k-1}), \tag{15}$$

where y_0 is an arbitrary point in R and $y_i = s(\alpha - z_{i-1})$, $i = 1, 2, \cdots, k-1$. For $k \geqq n$ the minimization at each stage is performed on the convex hull of $n + 1$ points

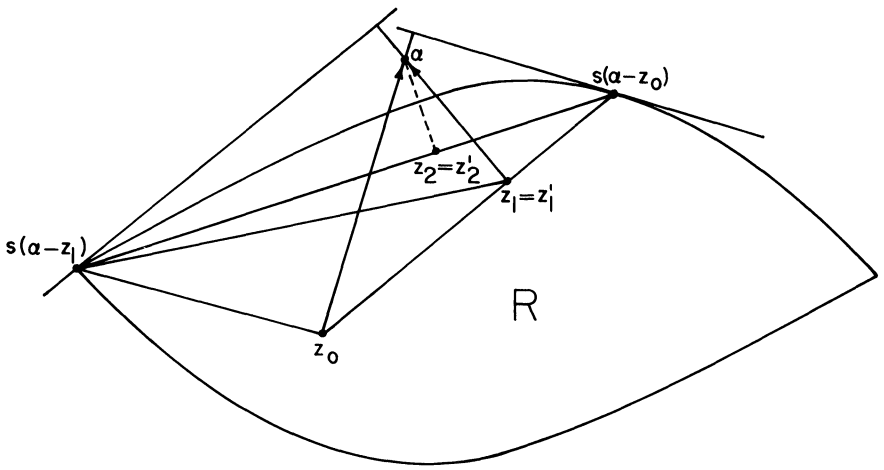
$$S_k = \Delta(y_0, y_1, \cdots, y_n). \tag{16}$$

Once $z_k \in S_k$ is obtained, one of the points $y_0, y_1, \cdots, y_n$ is replaced by $s(\alpha - z_k)$.

The following brief comments indicate the essential similarities and differences between the two procedures.

(a) Since the new procedure and Barr's procedures yield different polyhedra, except possibly at the initial stage, comparisons of the two methods are difficult.

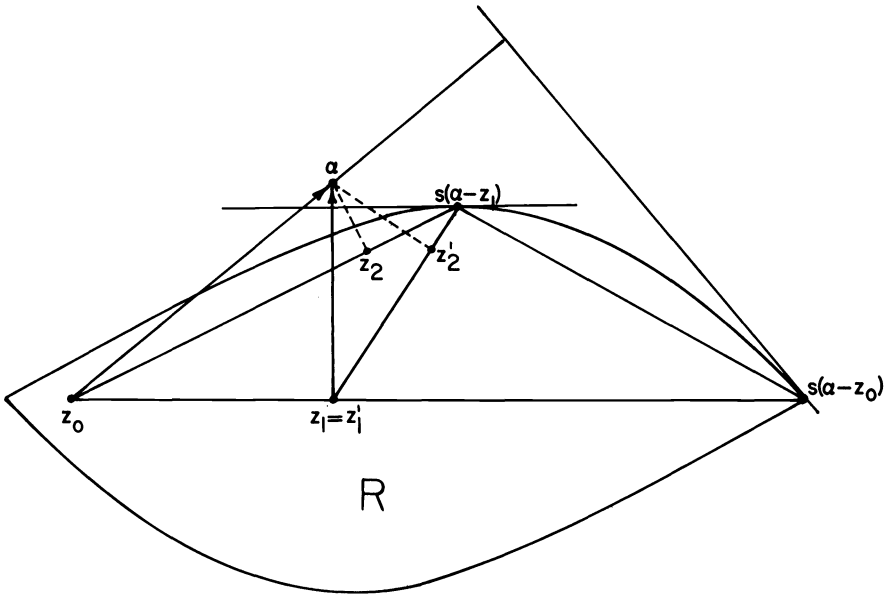
(b) For $k \leqq p$, Barr's procedure uses z_k to form the polyhedron H_k. The use of the initial point z_0 in the new procedure results in the largest possible polyhedron that can be formed at the kth stage. This point is illustrated in Fig. 2(a) and (b) using two 2-dimensional examples. Let $\{z'_k\}$ represent the sequence generated by IP of [4]. In Fig. 2(a) $z_2 = z'_2$, while in Fig. 2(b) $z_2 \neq z'_2$ and $\|\alpha - z_2\| < \|\alpha - z'_2\|$.



(a) $z_2 = z'_2$

FIG. 2. *The effect of* z_0 *on Barr's* IP *and the new procedure for an arbitrary convex set in* E^2

(c) While the minimization using the iterative procedure of [4] is carried out over the convex hull of $p + 2$ points as indicated above, it is pointed out by Barr that computational results indicate good convergence for $p = n$ and little improvement is obtained for $p > n$. Thus, for best results the procedure in [4] recommends the use of $n + 2$ points (although Barr shows that by using selection rule C, and in certain cases selection rule B, the minimization may be carried out on the convex hull of at most $n + 1$ points). The procedure in this paper uses at most $n + 1$ points. When only $n + 1$ points are used instead of $n + 2$, the minimization is simpler to accomplish.



(b) $z_2 \neq z_2'$, $\|\alpha - z_2\| < \|\alpha - z_2'\|$

FIG 2. *The effect of* z_0 *on Barr's* IP *and the new procedure for an arbitrary convex set in* E^2

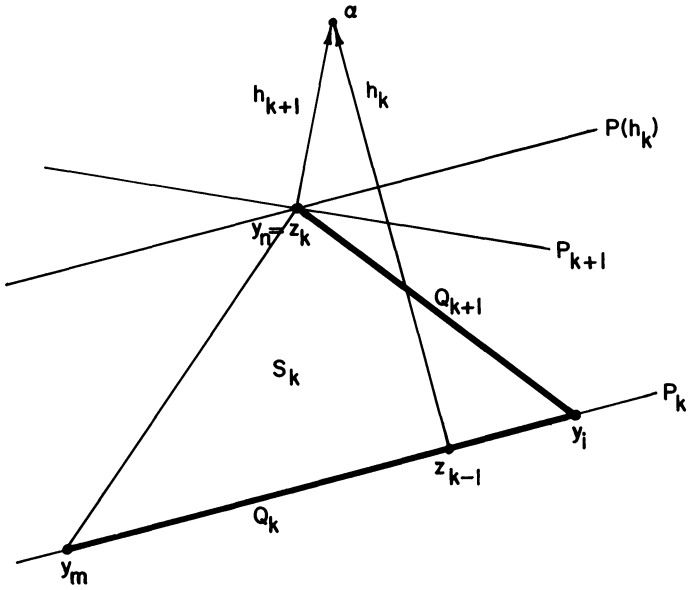


FIG. 3. *The k-th iteration,* $k \geqq n$, $z_k \neq z_{k-1}$

TABLE 1

$n = 3$, $z_0 = (5, 4, 2)$, $\lambda_2 = 1{,}000$, $\lambda_3 = 100$

ε	1	0.1	0.01	10^{-3}	10^{-4}	10^{-5}	10^{-6}
IP Selection Rule A ($p = n$)	9	16	23	26	29	33	36
New Procedure	6	12	16	19	22	26	29

TABLE 2

$n = 4$, $z_0 = (6, 1, 2, 2)$, $\lambda_2 = 1{,}000$, $\lambda_3 = 500$, $\lambda_4 = 100$

ε	1	0.1	0.01	10^{-3}	10^{-4}	10^{-5}	10^{-6}
IP Selection Rule A ($p = n$)	12	18	25	29	37	48	53
New Procedure	8	15	22	30	35	44	52

TABLE 3

For $n = 3$, $z_0 = (6, 2, 2)$; *for* $n = 4$, $z_0 = (6, 2, 2, 1)$; *for* $n = 5$, $z_0 = (5, 3, 1, 1.8, 2.6)$; *for* $n = 6$, $z_0 = (4, 3, 2.6, 2.6, 1.8, 1.8)$

						IP Selection Rule A ($p = n$)			IP Selection Rule B ($p = n$)			New Procedure		
n	λ_2	λ_3	λ_4	λ_5	ε / λ_6	1	10^{-3}	10^{-6}	1	10^{-3}	10^{-6}	1	10^{-3}	10^{-6}
3	10	10				3	7	12				3	7	12
3	100	100				4	9	13				2	9	13
3	1,000	1,000				4	10	12				4	9	12
3	100	10				6	17	32				6	17	31
3	1,000	1				6	11	17				4	10	16
3	1,000	10				7	20	29				5	16	27
3	1,000	100				8	23	32	9	23	32	7	18	30
3	100	50				7	18	30	7	21	31	5	18	30
3	100	90				6	17	31	6	19	29	5	19	32
4	100	50	10			7	29	49	8	30	51	7	29	46
4	100	90	10			7	28	48	7	28	47	6	23	38
4	1,000	500	100			10	31	48	10	35	50	11	28	49
5	100	70	50	10		9	39	66	8	36	63	11	38	71
5	100	90	80	70		10	38	70	9	37	69	8	39	70
5	1,000	900	500	100		12	43	79	15	42	74	9	55	84
6	100	90	70	50	10	12	45	91	11	49	92	9	52	96
6	1,000	90	70	50	10	14	59	94	15	60	94	11	51	94

(d) For $k > n$ both procedures replace a point with a new contact point. In [4] that point $y_i(k)$ is replaced which has the minimum value of $\mu(\hat{z}) = \langle \hat{z}, y_i(k) \rangle / \|\hat{z}\|$ associated with it. The procedure in this paper discards the point y_i whose Euclidean distance to the hyperplane P_{k+1} is largest. The latter discarding criterion has the desirable effect that the remaining n points of S_k form that face of S_k which is nearest to z^* (see Fig. 3).

4. Examples. For purposes of comparison the same problems have been solved using the present iterative procedure that appear in [2] and [4]. In these problems R is defined as

$$R = \left\{ z; 1 + \frac{1}{2} \sum_{i=2}^{n} \frac{(z^i)^2}{\lambda_i} \leqq z^1 \leqq 10; \lambda_i > 0, i = 2, 3, \cdots, n \right\}, \tag{17}$$

$\alpha = 0$, and the stopping criterion is $\|z_k\| - \|z^*\| \leqq \varepsilon$. The optimum is $z^* = (1, 0, \cdots, 0)^T$. The results, including those from [2] and [4] for the same problems with $p = n$, are tabulated in Tables 1, 2 and 3, which show the number of iterations necessary to satisfy the above stopping criterion.

REFERENCES

[1] E. G. GILBERT, *An iterative procedure for computing the minimum of a quadratic form on a convex set*, this Journal, 4 (1966), pp. 61–80.

[2] R. O. BARR, *Computation of optimal controls on convex reachable sets*, Mathematical Theory of Control, Academic Press, New York, 1967, pp. 63–70.

[3] T. PECSVARADI AND K. S. NARENDRA, *A new iterative procedure for the minimization of a quadratic form on a convex set*, Dunham Laboratory Tech. Rep. CT-28, Department of Engineering and Applied Science, Yale University, New Haven, Conn., 1969.

[4] R. O. BARR, *An efficient computational procedure for a generalized quadratic programming problem*, this Journal, 7 (1969), pp. 415–429.

[5] G. HADLEY, *Linear Algebra*, Addison-Wesley, Reading, Mass., 1961.

SIAM J. CONTROL
Vol. 8, No. 3, August 1970

SUFFICIENT CONDITIONS FOR NONNEGATIVITY OF THE SECOND VARIATION IN SINGULAR AND NONSINGULAR CONTROL PROBLEMS*

D. H. JACOBSON†

Abstract. Sufficient conditions for nonnegativity of the second variation in singular and nonsingular control problems are presented; these conditions are in the form of equalities and differential inequalities. Control problem examples illustrate the use of the new conditions. The relationships of the new conditions to existing necessary conditions of optimality for singular and nonsingular problems are discussed. When applied to nonsingular control problems, it is shown that the conditions are sufficient to ensure the boundedness of the solution of the well-known matrix Riccati differential equation.

1. Preliminaries.

1.1. Introduction. Singular control problems occur often in engineering; for example, in the aerospace industry a number of important problems are singular [1], [2]. Mathematical economics is another field in which singular optimal control problems are common [3]. These and other examples have prompted researchers to inquire into the mathematical properties of singular arcs [4]–[20], [28]–[33]. Circa 1964, Kelley [4] discovered a new necessary condition of optimality for singular arcs. This condition was generalized subsequently by Robbins [5], Tait [6], Kelley et al. [7] and Goh [8], and is now commonly known as the generalized Legendre–Clebsch condition (or Kelley's condition). In [9] an additional necessary condition of optimality for singular arcs was derived and was shown to be nonequivalent to the generalized Legendre–Clebsch condition. For want of an alternative, we shall refer to this condition as Jacobson's condition.[1]

In this paper we present sufficient conditions for nonnegativity of the second variation in singular control problems; in strengthened form these conditions (equalities and inequalities) are sufficient for a weak relative minimum. Both Kelley's and Jacobson's necessary conditions of optimality are derived easily from the new conditions. We show that the conditions are applicable to totally singular, partially singular[2] and nonsingular control functions. Moreover, when applied to nonsingular problems, sufficient conditions for the boundedness of the solution of the well-known matrix Riccati differential equation are obtained; these are, in certain cases, less stringent than those known heretofore [21], [26].

Control problems without terminal constraints are considered first; the results are then generalized to the case where constraints on the terminal states

* Received by the editors July 7, 1969, and in revised form February 22, 1970.

† Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts 02138. This work was supported by the U.S. Army Research Office, the U.S. Air Force Office of Scientific Research and the U.S. Office of Naval Research under the Joint Services Electronics Program under Contracts N00014-67-A-0298-0006, 0005 and 0008 and by the National Aeronautics and Space Administration under Grant NGR 22-007-068.

[1] It has been pointed out by the reviewers that Gabasov [29], [32] has obtained, independently, this necessary condition for the case of unconstrained terminal state. The treatment in [9] includes constraints on the terminal state.

[2] Defined in § 1.3.

are present. It turns out that the presence of terminal constraints does not complicate unduly the derivation.

1.2. Problem formulation. We shall consider the class of control problems where the dynamical system is described by the ordinary differential equations:

$$\dot{x} = f(x,u,t), \qquad x(t_0) = x_0, \tag{1}$$

where (except for § 5)

$$f(x,u,t) = f_1(x,t) + f_u(x,t)u. \tag{2}$$

The performance of the system is measured by the cost functional

$$V(x_0,t_0) = \int_{t_0}^{t_f} L(x,t)\,dt + F(x(t_f),t_f) \tag{3}$$

and the terminal states must satisfy

$$\psi(x(t_f),t_f) = 0. \tag{4}$$

The control function $u(\cdot)$ is required to satisfy the following constraint:

$$u(t) \in U, \qquad t \in [t_0,t_f], \tag{5}$$

where the set U is defined by:

$$U \equiv \{u(t): |u_i(t)| \leqq 1, i = 1,\cdots,m\}. \tag{6}$$

Here, x is an n-dimensional state vector and u is an m-dimensional control vector. f_1 is an n-dimensional vector function of x at time t and f_u is an $n \times m$ matrix function of x at time t; the functions L and F are scalar. The terminal constraint function ψ is an s-dimensional column vector function of $x(t_f)$ at t_f. The functions f, L, F and ψ are assumed smooth. The final time t_f is assumed to be given explicitly.

The control problem is: determine the control function $u(\cdot)$ which satisfies (4) and (5) and minimizes $V(x_0,t_0)$.

1.3. Totally and partially singular problems. It can be shown that, along an optimal trajectory, the following necessary conditions (Pontryagin's principle) hold:

$$-\dot{\lambda} = H_x(\bar{x},\bar{u},\lambda,t), \qquad \lambda(t_f) = F_x(\bar{x}(t_f),t_f) + \psi_x^T v, \tag{7}$$

where

$$\bar{u} = \arg\min_{u \in U} H(\bar{x},u,\lambda,t) \tag{8}$$

and

$$H(x,u,\lambda,t) = L(x,t) + \lambda^T f(x,u,t). \tag{9}$$

Here, $\bar{x}(\cdot), \bar{u}(\cdot)$ denote the candidate state and control functions and $\lambda(\cdot)$ denotes an n-vector of Lagrange multiplier functions of time. v is an s-dimensional vector of Lagrange multipliers associated with ψ.

In general the optimal control function (for the class of problems formulated in § 1.2) consists of bang-bang subarcs and singular subarcs.[3] A bang-bang arc is one along which $H_{u_i}(\bar{x}, \lambda, t) \neq 0$, $i = 1, \cdots, m$ (except at a finite number of switch times where the components of $\bar{u}$ change sign).

A singular arc [17] is one along which

$$H_{u_i}(\bar{x}, \lambda, t) = 0, \qquad i = 1, \cdots, m, \tag{10}$$

for a finite time interval.[4] Note that this implies that, on a singular arc, H is explicitly independent of the control u.

In the sequel we shall make use of the following definitions and assumption.

DEFINITION 1. A *totally singular control function* is one along which (10) holds for all $t \in [t_0, t_f]$.

DEFINITION 2. A *partially singular control function* is one along which (10) holds for k subintervals of length T_i, $i = 1, \cdots, k$, and where $\sum_{i=1}^{k} T_i < (t_f - t_0)$.

Assumption. The totally singular control function $\bar{u}(\cdot)$ is continuous in t.

2. Totally singular control functions, unconstrained terminal state.

2.1. Existing necessary conditions of optimality. In [7] Kelley et al. show that the following (generalized Legendre–Clebsch) condition is necessary for the optimality of a singular arc:

$$(-1)^q \frac{\partial}{\partial u}\left[\frac{d^{2q}}{dt^{2q}} H_u(\bar{x}, \lambda, t)\right] \geqq 0, \tag{11}$$

where the $2q$th time derivative of H_u is the first to contain explicitly the control u.[5] Kelley at al. used special control variations in order to derive this result; see [7]. Recently an additional (Jacobson's) necessary condition was discovered [9]. In order for a singular arc to be optimal it is necessary that

$$H_{ux} f_u + f_u^T Q f_u \geqq 0, \tag{12}$$

where

$$-\dot{Q} = H_{xx} + f_x^T Q + Q f_x, \qquad Q(t_f) = F_{xx}(\bar{x}(t_f), t_f). \tag{13}$$

The partial derivatives f_u, H_{xu}, H_{ux} and f_x are all evaluated along the singular arc $\bar{x}(\cdot), \bar{u}(\cdot)$. In [9] the above condition is derived for a scalar control using the technique of *Differential Dynamic Programming* [22]; in that paper, $Q(t)$ is shown to be the second partial derivative of $V(x, t)$ with respect to x obtained whilst *keeping* $u(\cdot) = \bar{u}(\cdot)$.[6] An alternative derivation, using the Lagrange multiplier rule, is given in the Appendix of this paper.

---

[3] "Arc" and "subarc" are used synonymously.

[4] For simplicity, we shall consider all the controls to be singular simultaneously. If this is not the case, no conceptual difficulties arise.

[5] It is possible that no time derivative of H_u contains the control u.

[6] Note that here $V(x, t)$ *is not the optimal value function.*

Of course, in addition to conditions (11) and (12), Pontryagin's principle must be satisfied.

2.2. Second variation ($\delta^2 V$). An expression for the second variation is (see [23])

$$\delta^2 V = \int_{t_0}^{t_f} \{\tfrac{1}{2}\delta x^T H_{xx}\delta x + \delta u^T H_{ux}\delta x\}\, dt + \tfrac{1}{2}\delta x^T F_{xx}\delta x|_{t_f} \tag{14}$$

subject to the differential equation

$$\delta\dot{x} = f_x\delta x + f_u\delta u, \qquad \delta x(t_0) = 0. \tag{15}$$

In order for the singular (stationary) solution to be minimizing it is necessary that

$$\delta^2 V \geqq 0 \tag{16}$$

for all $\delta u(\cdot)$ sufficiently small to justify the second order expansion of V, and such that

$$\bar{u}(t) + \delta u(t) \in U \quad \text{for all } t \in [t_0, t_f]. \tag{17}$$

Both Kelley's and Jacobson's conditions are necessary for (16) to hold; see [7] and the Appendix of this paper. In § 2.4 we present sufficient conditions for (16) to hold. Note that the auxiliary minimization problem (14), (15) cannot be solved routinely because it is singular.

2.3. Adjoining linearized system to $\delta^2 V$. We now adjoin (15) to (14) using a vector Lagrange multiplier function of time $\delta\lambda(t)$:

$$\begin{aligned}\delta^2 \hat{V} = \int_{t_0}^{t_f} \{&\tfrac{1}{2}\delta x^T H_{xx}\delta x + \delta u^T H_{ux}\delta x + \delta\lambda^T[f_x\delta x + f_u\delta u - \delta\dot{x}]\}\, dt \\ &+ \tfrac{1}{2}\delta x^T F_{xx}\delta x|_{t_f}.\end{aligned} \tag{18}$$

Integrating $\delta\lambda^T\delta\dot{x}$ by parts, we obtain

$$\begin{aligned}\delta^2 \hat{V} = \int_{t_0}^{t_f} \{&\tfrac{1}{2}\delta x^T H_{xx}\delta x + \delta u^T H_{ux}\delta x + \delta\lambda^T[f_x\delta x + f_u\delta u] + \delta\dot{\lambda}^T\delta x\}\, dt \\ &+ [\tfrac{1}{2}\delta x^T F_{xx}\delta x - \delta\lambda^T\delta x]|_{t_f}.\end{aligned} \tag{19}$$

Let us now *choose*[7]

$$\delta\lambda(t) = \tfrac{1}{2}P(t)\delta x, \tag{20}$$

where $P(t)$ is an $n \times n$ symmetric,[8] time-varying matrix having continuous time derivative $\dot{P}(t)$. The second variation becomes

$$\begin{aligned}\delta^2 \hat{V} = \int_{t_0}^{t_f} \{&\tfrac{1}{2}\delta x^T(\dot{P} + H_{xx} + f_x^T P + P f_x)\delta x + \delta u^T(H_{ux} + f_u^T P)\delta x\}\, dt \\ &+ [\tfrac{1}{2}\delta x^T F_{xx}\delta x - \tfrac{1}{2}\delta x^T P\delta x]|_{t_f}\end{aligned} \tag{21}$$

---

[7] This choice of $\delta\lambda$ is made in order that $\delta^2\hat{V}$ be a quadratic functional of δx and δu (as $\delta^2 V$ is).

[8] There is no loss of generality in choosing P to be symmetric; this is so since if P were chosen to be unsymmetric, $P + P^T$ would appear in place of P in (21).

subject to

$$\delta\dot{x} = f_x\delta x + f_u\delta u, \qquad \delta x(t_0) = 0. \tag{22}$$

Note that $\delta^2\hat{V} = \delta^2\hat{V}$, if (22) holds.

2.4. Sufficient conditions for nonnegativity of $\delta^2 V$. As remarked in § 2.2, the auxiliary problem (14), (15) or (21), (22) cannot be solved routinely owing to the fact that it is singular. Our new approach to the problem is to *choose* the matrix function $P(\cdot)$ such that

$$H_{ux} + f_u^T P = 0 \qquad \text{for all } t \in [t_0, t_f]. \tag{23}$$

Here, $P(t)$ is an $n \times n$ symmetric matrix function of time and H_{ux} is $m \times n$ so that there are cases where (23) can be solved by choosing appropriate values for some of the elements of P. By choosing P according to (23) we annihilate the coefficients of the mixed $\delta u\delta x$ terms in (21). The remaining terms are quadratic forms in $\delta x(t)$ and $\delta x(t_f)$. Clearly, sufficient conditions for $\delta^2\hat{V} = \delta^2 V \geqq 0$ are that (23) hold and

$$\dot{P} + H_{xx} + f_x^T P + Pf_x = M(t) \geqq 0 \tag{24}$$

and

$$-P(t_f) + F_{xx}(\bar{x}(t_f), t_f) = G(t_f) \geqq 0. \tag{25}$$

Equality (23) together with inequalities (24) and (25) constitute sufficient conditions for $\delta^2 V \geqq 0$ for all $\delta x(\cdot)$.

2.5. Sufficient conditions for optimality. Sufficient conditions for a weak relative minimum are obtained by strengthening (24) and (25):

$$\dot{P} + H_{xx} + f_x^T P + Pf_x = M(t) > 0 \quad \text{for all } t \in [t_0, t_f], \tag{26}$$

$$-P(t_f) + F_{xx}(\bar{x}(t_f), t_f) = G(t_f) > 0. \tag{27}$$

To see this, note that if (23), (26) and (27) hold, then $\delta^2\hat{V} = 0$ if and only if $\delta x(\cdot) = 0$ almost everywhere including t_f. However, if $\delta x(\cdot) = 0$ almost everywhere including t_f, then by our assumptions on L and F (see § 1.2) we have that the *total* change in cost is

$$\begin{aligned}\Delta V &= \int_{t_0}^{t_f} L(\bar{x} + \delta x, t)\,dt - \int_{t_0}^{t_f} L(\bar{x}, t)\,dt + F(\bar{x}(t_f) + \delta x(t_f), t_f) - F(\bar{x}(t_f), t_f) \\ &= 0,\end{aligned} \tag{28}$$

i.e.,

$$\delta^2\hat{V} = 0 \Rightarrow \Delta V = 0. \tag{29}$$

Thus, since u does not appear explicitly in L, we can always choose $\delta x(\cdot) \neq 0$ sufficiently small so that $\delta^2\hat{V}$ is the dominant term in the expansion for ΔV; hence we have sufficiency.

Example. $H_{ux} = 0$, $H_{xx} > 0$, $F_{xx} > 0$. In this case, $\dot{P}(t) = P(t) = 0$ for all $t \in [t_0, t_f]$ satisfies (23), (26), (27).

Note. If the dynamical equations (1) are linear and L and F are quadratic, then (23), (24) and (25) are sufficient conditions for optimality because all variations higher than the second vanish identically.

Example.

$$\begin{aligned} \dot{x}_1 &= x_2, \qquad x_1(0) = 0, \\ \dot{x}_2 &= u, \qquad x_2(0) = 0, \end{aligned} \tag{30}$$

$$V = \int_0^1 (\tfrac{1}{2}x_1^2 + 2x_1x_2 + \tfrac{1}{2}x_2^2)\, dt, \tag{31}$$

$$|u| \leqq 1. \tag{32}$$

Here, $\bar{u}(\cdot) = 0$ is a totally singular control which satisfies Pontryagin's principle. We have that

$$H_{ux} = 0, \tag{33}$$

$$F = 0 \tag{34}$$

and

$$H_{xx} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}. \tag{35}$$

Note that H_{xx} is *not* positive semidefinite. Equation (23) yields

$$P_{12}(t) = P_{22}(t) = 0, \qquad t \in [0, 1], \tag{36}$$

so that the left-hand side of (24) becomes

$$\begin{bmatrix} \dot{P}_{11} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & P_{11} \\ P_{11} & 0 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \tag{37}$$

and the left-hand side of (25) becomes

$$\begin{bmatrix} -P_{11}(t_f) & 0 \\ 0 & 0 \end{bmatrix}. \tag{38}$$

Inequalities (24) and (25) are satisfied if we choose

$$\dot{P}_{11} = 0, \qquad P_{11} = -2; \tag{39}$$

and since the system dynamics are linear, and the cost is quadratic, $\bar{u}(\cdot) = 0$ is optimal.

2.6. Relationship to existing necessary conditions. Both Kelley's [7] and Jacobson's [9] conditions can be derived from (23), (24), (25).

Jacobson's condition. Let

$$Q + \bar{P} = P, \tag{40}$$

where Q and $\bar{P}$ are both $n \times n$, symmetric matrix functions of time; then, from (23),

$$H_{ux} + f_u^T Q + f_u^T \bar{P} = 0 \tag{41}$$

so that

$$\tfrac{1}{2}f_u^T(H_{xu} + Qf_u + \bar{P}f_u) + \tfrac{1}{2}(H_{ux} + f_u^T Q + f_u^T \bar{P})f_u = 0. \tag{42}$$

From (24) and (40),

$$-\dot{\bar{P}} - \dot{Q} = H_{xx} + f_x^T(Q + \bar{P}) + (Q + \bar{P})f_x - M(t); \tag{43}$$

and from (25) and (40),

$$-Q(t_f) - \bar{P}(t_f) + F_{xx}(\bar{x}(t_f), t_f) - G(t_f) = 0. \tag{44}$$

Now set

$$Q(t_f) = F_{xx}(\bar{x}(t_f), t_f) \tag{45}$$

and

$$-\dot{Q} = H_{xx} + f_x^T Q + Qf_x \tag{46}$$

so that

$$\bar{P}(t_f) = -G(t_f) \tag{47}$$

and

$$-\dot{\bar{P}} = -M(t) + f_x^T \bar{P} + \bar{P}f_x. \tag{48}$$

Now, since

$$M(t) \geqq 0 \quad \text{for all } t \in [t_0, t_f] \quad \text{and} \quad G(t_f) \geqq 0, \tag{49}$$

we have that

$$\bar{P}(t) \leqq 0 \quad \text{for all } t \in [t_0, t_f]. \tag{50}$$

Using inequality (50) in (42) and noting that (23) implies that $H_{ux}f_u$ is symmetric, we obtain

$$H_{ux}f_u + f_u^T Q f_u \geqq 0. \tag{51}$$

Inequality (51) together with (45) and (46) comprise Jacobson's necessary condition.

Kelley's condition (*generalized Legendre–Clebsch*). Differentiating (23) with respect to time yields

$$\dot{H}_{ux} + \dot{f}_u^T P + f_u^T \dot{P} = 0 = \dot{H}_{ux} + \dot{f}_u^T P - f_u^T(H_{xx} + f_x^T P + Pf_x - M). \tag{52}$$

Postmultiplying (52) by f_u and adding its transpose, we obtain

$$\begin{aligned} &\dot{H}_{ux}f_u + f_u^T \dot{H}_{ux} + \dot{f}_u^T P f_u + f_u^T P \dot{f}_u - 2f_u^T H_{xx} f_u - 2f_u^T f_x^T P f_u \\ &\qquad - 2f_u^T P f_x f_u + 2f_u^T M f_u = 0. \end{aligned} \tag{53}$$

Using

$$H_{ux} = -f_u^T P \tag{54}$$

in (53) we have

$$\dot{H}_{ux}f_u + f_u^T\dot{H}_{ux} - \dot{f}_u^T H_{xu} - H_{ux}\dot{f}_u - 2f_u^T H_{xx}f_u + 2f_u^T f_x^T H_{xu} + 2H_{ux}f_x f_u + 2f_u^T M f_u = 0. \tag{55}$$

Rearranging (53), we have

$$-2f_u^T H_{xx}f_u + 2H_{ux}(f_x f_u - \dot{f}_u) + 2(f_u^T f_x^T - \dot{f}_u^T)H_{xu} + \frac{d}{dt}(H_{ux}f_u) + \frac{d}{dt}(f_u^T H_{xu}) = -2f_u^T M f_u. \tag{56}$$

However, we have that

$$M(t) \geqq 0 \quad \text{for all } t \in [t_0, t_f] \tag{57}$$

so that

$$-f_u^T H_{xx}f_u + H_{ux}(f_x f_u - \dot{f}_u) + (f_u^T f_x^T - \dot{f}_u^T)H_{xu} + \frac{d}{dt}(H_{ux}f_u) \leqq 0. \tag{58}$$

Now, the left-hand side of (58) is just

$$\frac{\partial}{\partial u}\left[\frac{d^2}{dt^2}H_u\right] \tag{59}$$

so that

$$(-1)\frac{\partial}{\partial u}\left[\frac{d^2}{dt^2}H_u\right] \geqq 0. \tag{60}$$

This is Kelley's first necessary condition. If this is met with equality, i.e.,

$$f_u^T M(t) f_u = 0, \tag{61}$$

then (56) is again differentiated with respect to time and (54) and (56) are substituted in. This yields Kelley's second condition, viz.,

$$\frac{\partial}{\partial u}\left[\frac{d^4}{dt^4}H_u\right] \geqq 0. \tag{62}$$

The generalized condition

$$(-1)^q\frac{\partial}{\partial u}\left[\frac{d^{2q}}{dt^{2q}}H_u\right] \geqq 0 \tag{63}$$

is obtained by further differentiations.

Note. In § 2.5, we gave sufficient conditions for optimality; a requirement was that

$$M(t) > 0 \quad \text{for all } t \in [t_0, t_f]. \tag{64}$$

However, this condition cannot hold unless $q = 1$ (see (56); if $q > 1$, then $f_u^T M f_u = 0$, contradicting (64)).

3. Totally singular control functions, constrained terminal state.

3.1. Second variation ($\delta^2 V^*$). We shall allow the terminal constraint

$$\psi(x(t_f), t_f) = 0, \tag{65}$$

where ψ is an s-dimensional vector function. As before, t_f is assumed to be given explicitly.

If ψ is adjoined to the cost functional by Lagrange multipliers ν (see [23]), the second variation is

$$\delta^2 V^* = \int_{t_0}^{t_f} \{\tfrac{1}{2}\delta x^T H_{xx}\delta x + \delta u^T H_{ux}\delta x\}\, dt + \tfrac{1}{2}\delta x^T(F_{xx} + \nu^T\psi_{xx})\delta x|_{t_f} \tag{66}$$

subject to[9]

$$\delta\dot{x} = f_x\delta x + f_u\delta u, \qquad \delta x(t_0) = 0 \tag{67}$$

and

$$\psi_x\delta x|_{t_f} = 0. \tag{68}$$

3.2. Adjoining linearized system to $\delta^2 V^*$. As in § 2, we adjoin (67) to (66) by a Lagrange multiplier function $\delta\lambda(\cdot)$. We integrate the term $\delta\lambda^T\delta\dot{x}$ by parts and set

$$\delta\lambda = \tfrac{1}{2}P(t)\delta x. \tag{69}$$

We obtain finally

$$\delta^2 V^* = \int_{t_0}^{t_f} \{\tfrac{1}{2}\delta x^T(\dot{P} + H_{xx} + f_x^T P + Pf_x)\delta x + \delta u^T(H_{ux} + f_u^T P)\delta x\}\, dt + \tfrac{1}{2}\delta x^T(F_{xx} + \nu^T\psi_{xx}P)\delta x|_{t_f} \tag{70}$$

subject to $\psi_x\delta x(t_f) = 0$.

If ψ_x has rank s, then s components of $\delta x(t_f)$—referred to as $\delta x^s(t_f)$—can be solved for in terms of the remaining $n - s$ components, $\delta x^{n-s}(t_f)$; for example,[10]

$$\delta x^s(t_f) = -A_1^{-1}A_2\delta x^{n-s}(t_f), \tag{71}$$

where

$$s\Big\{\overbrace{\big[\, A_1 \,\vdots\, A_2 \,\big]}^{n} = \psi_x \tag{72}$$

(A_1: s columns; A_2: $(n-s)$ columns)

so that

$$\delta x(t_f) = \begin{bmatrix} -A_1^{-1}A_2\delta x^{n-s}(t_f) \\ \hline \delta x^{n-s}(t_f) \end{bmatrix} = \begin{bmatrix} -A_1^{-1}A_2 \\ \hline I \end{bmatrix}\delta x^{n-s}(t_f) \triangleq Z\delta x^{n-s}(t_f), \tag{73}$$

where Z is $n \times (n - s)$.

[9] More precisely, we have that $(\bar{x} + \delta x')^{\cdot} = f(\bar{x} + \delta x', \bar{u} + \delta u, t)$ and $\psi(\bar{x}(t_f) + \delta x'(t_f), t_f) = 0$. However, expansions of these which are of higher order than the first do not influence $\delta^2 V^*$.

[10] If A_1 is singular, then differently partitioned ψ_x and $\delta x(t_f)$ must be used.

We now eliminate the constraint

$$\psi_x \delta x|_{t_f} = 0 \tag{74}$$

from (70) by using (73) in the boundary terms of $\delta^2 V^*$:

$$\delta^2 \tilde{V} = \int_{t_0}^{t_f} \{\tfrac{1}{2}\delta x^T(\dot{P} + H_{xx} + f_x^T P + P f_x)\delta x + \delta u^T(H_{ux} + f_u^T P)\delta x\}\, dt + \tfrac{1}{2}(\delta x^{n-s})^T\{Z^T(F_{xx} + v^T\psi_{xx} - P)Z\}\delta x^{n-s}|_{t_f}. \tag{75}$$

3.3. Sufficient conditions for nonnegativity of $\delta^2 \tilde{V}$. Sufficient conditions for $\delta^2 \tilde{V} \geqq 0$ are (by analogy with § 2.4)

$$H_{ux} + f_u^T P = 0 \quad \text{for all } t \in [t_0, t_f], \tag{76}$$

$$\dot{P} + H_{xx} + f_x^T P + P f_x = M'(t) \geqq 0 \qquad \text{for all } t \in [t_0, t_f] \tag{77}$$

and

$$Z^T(F_{xx} + v^T\psi_{xx} - P)Z|_{t_f}) = G'(tf) \geqq 0. \tag{78}$$

Note that if $s = 0$ (no terminal constraints),

$$Z = {}^{n}\updownarrow^{\overset{n}{\longleftrightarrow}}[I] \tag{79}$$

and (76)–(78) reduce to (23)–(25).

3.4. Sufficient conditions for optimality. By strengthening the inequalities in (77) and (78), we obtain

$$\delta^2 \tilde{V} > 0 \quad \text{for all } \delta x(\cdot) \neq 0 \tag{80}$$

with

$$\delta^2 \tilde{V} = 0 \quad \text{if and only if} \quad \delta x(\cdot) = 0 \tag{81}$$

almost everywhere, including t_f. The argument of § 2.5 can be used here to show that (80), (81) imply optimality (weak relative minimum).

Note. As in the case of unconstrained terminal states, these strengthened conditions can hold only if the singular arc is first order (i.e., the generalized Legendre–Clebsch condition holds with strict inequality for $q = 1$).

3.5. Relationship to existing necessary conditions. As in § 2.6 it is easy to show that satisfaction of (76)–(78) implies that Kelley's condition is satisfied. Jacobson's condition for problems with constrained terminal state is more complex than for the unconstrained case; see [9]. We shall not derive this condition here, from (76)–(78).

3.6. Comment on problems with constrained terminal state. When deriving necessary conditions of optimality for problems with terminal constraints by constructing variations of the control function, one is faced with the task of showing that the chosen variation is indeed admissible [7], [9]. This is a formidable task even if the linearized dynamical system is assumed to be completely con-

trollable and ψ_x is assumed to have rank s.[11] We remark that the approach taken in this paper does not require arguments of the type referred to above. We need only *assume* that it is possible to satisfy $\psi(x(t_f), t_f) = 0$, and that ψ_x has rank s at $\bar{x}(t_f), t_f$. We do *not* have to construct explicitly admissible control variations; our conclusions follow directly by requiring that certain time-varying matrices be positive semidefinite (see § 2.3).

4. Partially singular control functions.[12]

4.1. First and second variation ($\delta V^* + \delta^2 V^*$). As defined in § 1.3, a partially singular control function may have both singular and nonsingular portions (i.e., subintervals of singular and bang-bang control). Along nonsingular arcs, $H_u \neq 0$, and the condition (Pontryagin's)

$$\min_{u \in U} H(\bar{x}, u, \lambda, t) \tag{82}$$

must hold (this is trivially satisfied along a singular arc). In this case the sum of the first and second variations is

$$\begin{aligned} \delta V^* + \delta^2 V^* = \int_{t_0}^{t_f} \{H_u^T \delta u + \tfrac{1}{2}\delta x^T H_{xx} \delta x + \delta u^T H_{ux} \delta x\}\, dt \\ + \tfrac{1}{2}\delta x^T (F_{xx} + v^T \psi_{xx}) \delta x|_{t_f} \end{aligned} \tag{83}$$

subject to

$$\delta \dot{x} = f_x \delta x + f_u \delta u, \qquad \delta x(t_0) = 0 \tag{84}$$

and

$$\psi_x \delta x|_{t_f} = 0. \tag{85}$$

In order to enforce (85) (and $\psi(x(t_f), t_f) = 0$), we have

$$\bar{u}(\cdot) + \delta u(\cdot) \in U_3, \tag{86}$$

where

$$U_3 = U_2 \cap U_1 \tag{87}$$

and

$$\begin{aligned} U_1 &= \{u(\cdot) : |u_i(t)| \leqq 1, t \in [t_0, t_f], i = 1, \cdots, m\}, \\ U_2 &\equiv \{u(\cdot) : \psi(x(t_f), t_f) = 0, \dot{x} = f(x, u, t); x(t_0) = x_0\}. \end{aligned} \tag{88}$$

Note that by (82),

$$H_u^T \delta u \geqq 0, \qquad \bar{u}(\cdot) + \delta u(\cdot) \in U_3 \tag{89}$$

with equality holding along singular arcs and at switch times of the bang-bang control arcs. If there are no singular arcs and no switchings of the control (i.e., $|H_u| \neq 0$ for all $t \in [t_0, t_f]$ so that $\bar{u}(\cdot) = \text{const.} = +1$ or -1), then Pontryagin's

[11] These are common assumptions [23].

[12] In this section we treat the constrained terminal state problem; the unconstrained problem is a special case.

principle is sufficient for optimality because the second order terms in (83) can be made insignificant (i.e., dominated by $H_u^T \delta u$) for $\|\delta u(\cdot)\|$ sufficiently small. In the case where bang-bang arcs are present (i.e., where $\bar{u}(t)$ switches between its upper and lower bounds) one can, by placing a control variation in the immediate vicinity of a switch point, cause $H_u^T \delta u$ to contribute *less* to the change in cost $\delta V^* + \delta^2 V^*$ than the second variation terms.

Clearly, sufficient conditions for $\delta^2 V^* \geqq 0$ are (76)–(78) and sufficient conditions of optimality (weak minimum) are these in strengthened form. Less restrictive sufficient conditions for purely bang-bang control functions have been given previously [24], [25]. However, in this section, we allow partially singular (i.e., "partially bang-bang") control functions and thus embrace a wider class of problems than in [24] and [25].

5. Problems nonlinear in control.

5.1. Introduction. In the last section we indicated that our approach to sufficiency is independent of whether the control function is totally singular or partially singular or, in the purely bang-bang case, nonsingular. In this section we study the more general nonsingular problem where the control u appears nonlinearly in f and L. We show that our approach is indeed applicable and give examples to illustrate our results. As a byproduct of the analysis, we obtain sufficient conditions for the boundedness of the solution of a certain matrix Riccati differential equation.

We shall consider the following nonlinear optimal control problem:

$$\dot{x} = f(x, u, t), \qquad x(t_0) = x_0, \tag{90}$$

$$V(x_0, t_0) = \int_{t_0}^{t_f} L(x, u, t)\, dt + F(x(t_f), t_f). \tag{91}$$

Here it is assumed, for simplicity, that there are no constraints on the control u or on the terminal state $x(t_f)$, though this in no way limits the wider applicability of the analysis (see § 5.5 for constrained terminal state). In this case the second variation is

$$\delta^2 V = \int_{t_0}^{t_f} \{\tfrac{1}{2}\delta x^T H_{xx} \delta x + \delta u^T H_{ux} \delta x + \tfrac{1}{2}\delta u^T H_{uu} \delta u\}\, dt + \tfrac{1}{2}\delta x^T F_{xx} \delta x|_{t_f} \tag{92}$$

subject to

$$\delta \dot{x} = f_x \delta x + f_u \delta u, \qquad \delta x(t_0) = 0. \tag{93}$$

Here,

$$H_{uu}(t) \geqq 0 \quad \text{for all } t \in [t_0, t_f] \tag{94}$$

is a well-known necessary condition (Legendre–Clebsch) of optimality. For the problem to be nonsingular, strict inequality must hold, i.e.,

$$H_{uu}(t) > 0 \quad \text{for all } t \in [t_0, t_f]. \tag{95}$$

A known necessary condition of optimality[13] [26] (which together with (95) and Pontryagin's principle forms sufficient conditions of optimality) is that the solution to the following matrix Riccati differential equation be bounded for $t \in [t_0, t_f]$:

$$\begin{aligned} -\dot{S} &= H_{xx} + f_x^T S + S f_x - (H_{ux} + f_u^T S)^T H_{uu}^{-1}(H_{ux} + f_u^T S); \\ S(t_f) &= F_{xx}|_{t_f}. \end{aligned} \tag{96}$$

Sufficient conditions for the boundedness of $S(\cdot)$ are known to be (see [21], [26])

$$\begin{aligned} &H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} \geqq 0 \quad \text{for all } t \in [t_0, t_f], \\ &F_{xx}(\bar{x}(t_f), t_f) \geqq 0, \\ &H_{uu}^{-1}(t) > 0 \quad \text{for all } t \in [t_0, t_f]. \end{aligned} \tag{97}$$

5.2. Sufficient conditions for optimality. Equation (93) can be adjoined to (92) using a vector Lagrange multiplier function of time $\delta\lambda(\cdot)$. If, as before, we let

$$\delta\lambda = \tfrac{1}{2}P(t)\delta x, \tag{98}$$

then, the second variation becomes

$$\begin{aligned} \delta^2 \hat{V} = \int_{t_0}^{t_f} \{&\tfrac{1}{2}\delta x^T(\dot{P} + H_{xx} + f_x^T P + P f_x)\delta x + \delta u^T(H_{ux} + f_u^T P)\delta x \\ &+ \tfrac{1}{2}\delta u^T H_{uu}\delta u\}\, dt + \tfrac{1}{2}\delta x^T(F_{xx} - P)\delta x|_{t_f}. \end{aligned} \tag{99}$$

Clearly, $\delta^2 \hat{V} \geqq 0$ if we can choose $P(t)$ so that

$$H_{ux} + f_u^T P = 0 \quad \text{for all } t \in [t_0, t_f], \tag{100}$$

$$\dot{P} + H_{xx} + f_x^T P + P f_x = M(t) \geqq 0 \quad \text{for all } t \in [t_0, t_f], \tag{101}$$

$$-P(t_f) + F_{xx}(\bar{x}(t_f), t_f) = G(t_f) \geqq 0. \tag{102}$$

Moreover, because of (95), we have that

$$\delta^2 \hat{V} \geqq kN^2[\delta u(\cdot)] \quad \text{for all } \delta u(\cdot), \tag{103}$$

where N is a suitable norm on $\delta u(\cdot)$ and $k > 0$. Inequality (103) indicates that $\delta^2 \hat{V}$ is *strongly positive* and, by a theorem of Gel'fand and Fomin (27, p. 100], this is sufficient for $\bar{u}(\cdot)$ to be a minimizing control function (weak relative minimum). Thus conditions (100)–(102) are sufficient for optimality in this nonsingular problem. As an immediate consequence we have the following result: Conditions (100)–(102) imply that the matrix Riccati equation (96) has a bounded solution in the interval $[t_0, t_f]$ (because the boundedness of $S(\cdot)$ is a necessary condition of optimality). These conditions are, in certain cases, considerably weaker than (97), as the following example illustrates.

[13] Classically known as the "no-conjugate-point condition" [27, p. 100].

Example.

$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = u, \tag{104}$$

$$V = \int_0^1 (\tfrac{1}{2}x_1^2 + 2x_1x_2 + \tfrac{1}{2}x_2^2 + \tfrac{1}{2}u^2)\,dt. \tag{105}$$

Here,

$$H_{ux} = 0, \quad H_{xx} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}, \quad F = 0. \tag{106}$$

These values do not satisfy conditions (97). However, from § 2.5, we have that

$$P_{12}(t) = P_{22}(t) = 0 = \dot{P}_{11}(t) \quad \text{for all } t \in [t_0, t_f] \tag{107}$$

and

$$P_{11} \leqq -1 \tag{108}$$

satisfy (100)–(102), so that the stationary solution to this problem, obtained from Pontryagin's principle, is optimal. Note that in this particular case the checking of (100)–(102) is considerably easier than integrating the matrix Riccati differential equation to see whether or not its solution is bounded in the interval [0, 1].[14]

5.3. Derivation of Riccati equation. The Riccati differential equation (96) can be derived directly from (100)–(102) as follows: From (100) and (101),

$$-\dot{P} = H_{xx} + f_x^T P + P f_x - M(t) - (H_{ux} + f_u^T P)^T H_{uu}^{-1}(H_{ux} + f_u^T P). \tag{109}$$

Let

$$P = \bar{P} + S; \tag{110}$$

then

$$\begin{aligned} -\dot{\bar{P}} - \dot{S} &= H_{xx} + f_x^T(\bar{P} + S) + (\bar{P} + S)f_x - M(t) \\ &\quad - [H_{ux} + f_u^T(\bar{P} + S)]^T H_{uu}^{-1}[H_{ux} + f_u^T(\bar{P} + S)] \end{aligned} \tag{111}$$

$$\begin{aligned} &= H_{xx} + f_x^T(\bar{P} + S) + (\bar{P} + S)f_x - M(t) \\ &\quad - (H_{ux} + f_u^T S)^T H_{uu}^{-1}(H_{ux} + f_u^T S) \\ &\quad - (H_{ux} + f_u^T S)^T H_{uu}^{-1} f_u^T \bar{P} - \bar{P} f_u H_{uu}^{-1}(H_{ux} + f_u^T S) \\ &\quad - \bar{P} f_u H_{uu}^{-1} f_u^T \bar{P}. \end{aligned} \tag{112}$$

Using (100) and (110) in the last three terms of (112), we obtain

$$\begin{aligned} -\dot{\bar{P}} - \dot{S} &= H_{xx} + f_x^T(S + \bar{P}) + (\bar{P} + S)f_x - (H_{ux} + f_u^T S)^T H_{uu}^{-1}(H_{ux} + f_u^T S) \\ &\quad + \bar{P} f_u H_{uu}^{-1} f_u^T \bar{P} - M(t). \end{aligned} \tag{113}$$

Now *choose*

$$-\dot{\bar{P}} = -M(t) + f_x^T \bar{P} + \bar{P} f_x + \bar{P} f_u H_{uu}^{-1} f_u^T \bar{P}. \tag{114}$$

[14] Of course, in more complicated examples, checking of (100)–(102) may be less trivial.

From (102) and (110), we have that

$$-\bar{P}(t_f) - S(t_f) + F_{xx} = G(t_f) \geqq 0. \tag{115}$$

Choose

$$\bar{P}(t_f) = -G(t_f). \tag{116}$$

Now we have that $\bar{P}(t)$ is bounded in the interval $[t_0, t_f]$. This follows from the fact that $(-\bar{P})$ satisfies a Riccati equation for which conditions (97) hold, viz.,

$$\begin{aligned} &M(t) \geqq 0, \qquad H_{uu}^{-1}(t) > 0 \qquad \text{for all } t \in [t_0, t_f], \\ &G(t_f) \geqq 0. \end{aligned} \tag{117}$$

Using these results in (113) and (115), we obtain finally that

$$\begin{aligned} -\dot{S} &= H_{xx} + f_x^T S + S f_x - (H_{ux} + f_u^T S)^T H_{uu}^{-1}(H_{ux} + f_u^T S), \\ S(t_f) &= F_{xx}(\bar{x}(t_f), t_f) \end{aligned} \tag{118}$$

which is the Riccati equation (96). Now since (100)–(102) are satisfied by a matrix function $P(\cdot)$ which has bounded elements, and since, by (117), $\bar{P}(\cdot)$ is bounded, we have from (110) the result that $S(\cdot)$ is bounded.

5.4. Another example.

$$\begin{aligned} \dot{x}_1 &= x_2, \qquad x_1(0) = x_{10}, \\ \dot{x}_2 &= u, \qquad x_2(0) = x_{20}, \end{aligned} \tag{119}$$

$$V = \int_0^1 (-\tfrac{1}{2}x_1^2 + 2x_2^2 + \tfrac{1}{2}u^2)\, dt. \tag{120}$$

Here,

$$H_{ux} = 0, \quad H_{xx} = \begin{bmatrix} -1 & 0 \\ 0 & 4 \end{bmatrix}, \quad F = 0. \tag{121}$$

These values do not satisfy conditions (97). Conditions (100)–(102) become

$$[P_{12} \quad P_{22}] = 0, \tag{122}$$

$$\begin{bmatrix} \dot{P}_{11} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & P_{11} \\ P_{11} & 0 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 0 & 4 \end{bmatrix} \geqq 0 \tag{123}$$

and

$$\begin{bmatrix} -P_{11}(1) & 0 \\ 0 & 0 \end{bmatrix} \geqq 0. \tag{124}$$

Let us choose $P_{11}(t_f) = 0$; this satisfies (124). From (122), $P_{12}(t) = P_{22}(t) = 0$ for all $t \in [t_0, t_f]$. If we choose

$$\dot{P}_{11} = 2 \quad \text{then} \quad P_{11}(0) = -2, \tag{125}$$

and (123) becomes

$$\begin{bmatrix} 1 & -2+2t \\ -2+2t & 4 \end{bmatrix} \geqq 0. \tag{126}$$

Inequality (126) holds for all $t \in [0, 1]$. Thus the solution obtained from Pontryagin's principle is optimal, and the Riccati equation associated with the above control problem has a bounded solution

5.5. Constrained terminal state. From § 3.3 and § 5.2, sufficient conditions for optimality are

$$H_{ux} + f_u^T P = 0 \quad \text{for all } t \in [t_0, t_f], \tag{127}$$

$$\dot{P} + H_{xx} + f_x^T P + P f_x = M'(t) \geqq 0 \quad \text{for all } t \in [t_0, t_f], \tag{128}$$

$$Z^T(F_{xx} + v^T \psi_{xx} - P)Z|_{t_f} = G'(t_f) \geqq 0 \tag{129}$$

and (Legendre–Clebsch)

$$H_{uu}(t) > 0 \quad \text{for all } t \in [t_0, t_f]. \tag{130}$$

Example.

$$\begin{aligned} \dot{x}_1 &= x_2, \quad x_1(0) = x_{10}, \quad x_1(1) = 0, \\ \dot{x}_2 &= u, \quad x_2(0) = x_{20}, \quad x_2(1) = 0, \end{aligned} \tag{131}$$

$$V = \int_0^1 (-\tfrac{1}{2}x_1^2 + \tfrac{1}{2}x_2^2 + \tfrac{1}{2}u^2)\, dt. \tag{132}$$

Here,

$$H_{ux} = 0, \tag{133}$$

$$H_{xx} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad F = 0, \quad H_{uu} = 1. \tag{134}$$

In this case, because $n = s = 2$, condition (129) disappears. As before, we have that

$$H_{ux} + f_u^T P = [P_{12} \quad P_{22}] = 0. \tag{135}$$

Condition (128) becomes

$$\begin{bmatrix} \dot{P}_{11} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & P_{11} \\ P_{11} & 0 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \geqq 0. \tag{136}$$

Choosing $P_{11}(0) = -1$ and $\dot{P}_{11} = 2$, the left-hand side of (136) becomes

$$\begin{bmatrix} 1 & -1+2t \\ -1+2t & 1 \end{bmatrix} \tag{137}$$

which is $\geqq 0$ for all $t \in [0, 1]$, so that the stationary solution obtained from Pontryagin's principle is optimal. Note that the above sufficiency conditions are,

in this case, easier to check than the usual sufficiency conditions for nonsingular, constrained terminal state problems [23]. Moreover, the presence of the terminal constraints actually makes the choice of $\dot{P}$ and $P(t_f)$ easier (if in this example there were no terminal constraints, inequality (129) would be violated by our above choice of $P_{11}(0)$ and $\dot{P}_{11}$).

6. Applicability of the new conditions. If the conditions

$$H_{ux} + f_u^T P = 0 \quad \text{for all } t \in [t_0, t_f], \tag{138}$$

$$\dot{P} + H_{xx} + f_x^T P + P f_x = M'(t) \geqq 0 \quad \text{for all } t \in [t_0, t_f], \tag{139}$$

$$Z^T(F_{xx} + v^T \psi_{xx} - P)Z|_{t_f} = G'(t_f) \geqq 0 \tag{140}$$

cannot be satisfied, then no conclusion can be drawn regarding the nature (optimality or nonoptimality) of the stationary control function. This is because the above conditions are *sufficient* (but probably not necessary).[15]

Example.

$$\dot{x} = u, \qquad x(t_0) = 0, \tag{141}$$

$$|u| \leqq 1, \tag{142}$$

$$V = \int_0^1 \tfrac{1}{2}x^2 \, dt - \tfrac{1}{2}\alpha(t_f)x^2(t_f). \tag{143}$$

Clearly,

$$\bar{u}(\cdot) = 0 \tag{144}$$

is a stationary solution for the above problem. Here, $H_{ux} = 0$, $H_{xx} = 1$, $F_{xx} = -\alpha(t_f)$ and P is scalar so that (100) determines

$$P(t) = 0 \quad \text{for all } t \in [0, 1]. \tag{145}$$

Condition (101) becomes

$$1 \geqq 0 \tag{146}$$

and condition (102) becomes

$$-\alpha(t_f) \geqq 0. \tag{147}$$

Clearly, (101) is satisfied and (102) is satisfied if

$$\alpha(t_f) \leqq 0 \tag{148}$$

but is violated if

$$\alpha(t_f) > 0. \tag{149}$$

However, application of Jacobson's necessary condition [9] to this problem shows that if $\alpha(t_f) > 0$, the stationary solution (144) is *not* minimizing.

The above example suggests the following sufficient condition for nonoptimality of a singular control function.

[15] Current research now indicates that these conditions *are* necessary.

7. Sufficient conditions for nonoptimality of a singular control function. The second variation for the unconstrained terminal state problem is

$$\delta^2 \hat{V} = \int_{t_0}^{t_f} \{\tfrac{1}{2}\delta x^T(\dot{P} + H_{xx} + f_x^T P + P f_x)\delta x + \delta u^T(H_{ux} + f_u^T P)\delta x\}\, dt + \tfrac{1}{2}\delta x^T(F_{xx} - P)\delta x|_{t_f}. \tag{150}$$

If it is possible to choose $P(t)$, $t \in [t_0, t_f]$, such that

$$\dot{P} + H_{xx} + f_x^T P + P f_x = M''(t) \leqq 0 \tag{151}$$

and

$$-P(t_f) + F_{xx}(\bar{x}(t_f), t_f) = G''(t_f) \leqq 0 \tag{152}$$

and

$$\tfrac{1}{2} f_u^T H_{xu} + \tfrac{1}{2} H_{ux} f_u + f_u^T P f_u < 0, \tag{153}$$

then the singular control is nonoptimal.

The first two conditions cause the quadratic forms in δx and $\delta x(t_f)$ in (150) to be nonpositive. If a rectangular pulse variation $\delta u(\cdot)$ of height η and duration ΔT is introduced, then the dominant term (for η and ΔT sufficiently small) of

$$\int_{t_0}^{t_f} \delta u^T(H_{ux} + f_u^T P)\delta x\, dt \tag{154}$$

is

$$\tfrac{1}{2}\eta^T[\tfrac{1}{2} f_u^T H_{xu} + \tfrac{1}{2} H_{ux} f_u + f_u^T P f_u]\eta(\Delta T)^2; \tag{155}$$

so that if

$$\tfrac{1}{2} f_u^T H_{xu} + \tfrac{1}{2} H_{ux} f_u + f_u^T P f_u < 0, \tag{156}$$

then

$$\delta^2 \hat{V} < 0, \tag{157}$$

and the singular control is not minimizing.

Example.

$$\dot{x} = u, \qquad x(0) = 0, \tag{158}$$

$$V = \int_0^1 \frac{1}{2} x^2\, dt - x^2(t_f). \tag{159}$$

In this case, conditions (151) and (152) become

$$\dot{P} + 1 \leqq 0 \Rightarrow \dot{P} \leqq -1, \tag{160}$$

$$-P(t_f) - 2 \leqq 0 \Rightarrow P(t_f) \geqq -2 \tag{161}$$

and

$$\tfrac{1}{2} f_u^T H_{xu} + \tfrac{1}{2} H_{ux} f_u + f_u^T P f_u = P. \tag{162}$$

Choose

$$\dot{P} = -1 \quad \text{and} \quad P(t_f) = -2; \tag{163}$$

then conditions (151)–(153) are satisfied and the singular arc is nonoptimal.

8. Conclusion. In this paper sufficient conditions are presented for the second variation to be nonnegative in both singular and nonsingular control problems. It is demonstrated that known necessary conditions of optimality for singular problems and the no-conjugate-point condition for nonsingular problems are implied by the new conditions. Simple illustrative examples demonstrate the usefulness of the new conditions. A sufficient condition of optimality for singular problems is obtained by strengthening the inequality conditions; it is shown that these strengthened conditions can only be satisfied by first order singular problems.

When applied to the nonsingular control problem, the new conditions yield sufficient conditions for the boundedness of the solution of the matrix Riccati differential equation; this result appears to be useful in its own right.

The derivations presented are carried out for the case of u an n-vector, and s-vector constraints on the terminal state are permitted. Throughout, the final time t_f is assumed to be given explicitly; the generalization of the conditions to the case where t_f is given implicitly is straightforward but tedious.

The Appendix contains a Lagrange multiplier derivation of a necessary condition of optimality for singular control problems which was derived previously using differential dynamic programming [9].

The derivations in this paper are formal. In order to make the proofs rigorous it is necessary to justify the integrations by parts. Indeed, in view of the assumptions made in § 1.2, § 1.3 these integrations are valid.

Appendix. Lagrange multiplier derivation of Jacobson's necessary condition of optimality for singular problems (no terminal constraints). The second variation is

$$\delta^2 V = \int_{t_0}^{t_f} \{\tfrac{1}{2}\delta x^T H_{xx}\delta x + \delta u^T H_{ux}\delta x\}\, dt + \tfrac{1}{2}\delta x^T F_{xx}\delta x|_{t_f} \tag{A.1}$$

subject to

$$\delta\dot{x} = f_x\delta x + f_u\delta u, \quad \delta x(t_0) = 0. \tag{A.2}$$

Adjoining (A.2) to (A.1) with Lagrange multiplier

$$\delta\lambda = \tfrac{1}{2}Q(t)\delta x \tag{A.3}$$

(where Q is an $n \times n$ symmetric matrix function of time) and integrating by parts, we obtain

$$\begin{aligned}\delta^2 \hat{V} = \int_{t_0}^{t_f} \{\tfrac{1}{2}\delta x^T(\dot{Q} + H_{xx} + f_x^T Q + Q f_x)\delta x + \delta u^T(H_{ux} + f_u^T Q)\delta x\}\, dt \\ + \tfrac{1}{2}\delta x^T(F_{xx} - Q)\delta x|_{t_f}.\end{aligned} \tag{A.4}$$

Now, *choose*

$$-\dot{Q} = H_{xx} + f_x^T Q + Q f_x, \qquad Q(t_f) = F_{xx}(\bar{x}(t_f), t_f); \tag{A.5}$$

then

$$\delta^2 \hat{V} = \int_{t_0}^{t_f} \delta u^T (H_{ux} + f_u^T Q)\delta x \, dt. \tag{A.6}$$

Introduce a variation $\delta u(\cdot)$ which is zero everywhere except, say, in the interval $[t_1, t_1 + \Delta T]$ where

$$t_1 \quad \text{and} \quad t_1 + \Delta T \in [t_0, t_f], \tag{A.7}$$

and which has constant magnitude η (note that $\bar{u}(\cdot) + \delta u(\cdot) \in U$).

The dominant term of (A.6) produced by this variation is seen easily to be

$$\tfrac{1}{2}\eta^T[\tfrac{1}{2}f_u^T H_{xu} + \tfrac{1}{2}H_{ux}f_u + f_u^T Q f_u]\eta(\Delta T)^2. \tag{A.8}$$

From (A.8), for nonnegative $\delta^2 V$, we must have

$$\tfrac{1}{2}f_u^T H_{xu} + \tfrac{1}{2}H_{ux}f_u + f_u^T Q f_u \geqq 0. \tag{A.9}$$

A known necessary condition of optimality [5], [8], [28] is that $H_{ux}f_u$ be symmetric. Using this in (A.9) yields

$$H_{ux}f_u + f_u^T Q f_u \geqq 0. \tag{A.10}$$

This inequality, together with (A.5), comprises the necessary condition of optimality obtained (for the case of scalar control), using differential dynamic programming, in [9].

REFERENCES

[1] H. J. KELLEY, *Singular extremals in Lawden's problem of optimal rocket flight*, J. AIAA, 1 (1963), pp. 1578–1580.

[2] H. M. ROBBINS, *Optimality of intermediate-thrust arcs of rocket trajectories*, Ibid., 3 (1965), pp. 1094–1098.

[3] A. R. DOBELL AND Y. C. HO, *Optimal investment policy: An example of a control problem in economic theory*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 4–14.

[4] H. J. KELLEY, *A second variation test for singular extremals*, J. AIAA, 2 (1964), pp. 1380–1382.

[5] H. M. ROBBINS, *A generalized Legendre–Clebsch condition for the singular cases of optimal control*, Rep. 66-825 P 2043, IBM, Federal Systems Division, Owego, N.Y., 1966.

[6] K. S. TAIT, *Singular problems in optimal control*, Doctoral thesis, Harvard University, Cambridge, Mass., 1965.

[7] H. J. KELLEY, R. E. KOPP AND H. G. MOYER, *Singular extremals*, Topics in Optimization, G. Leitmann, ed., Academic Press, New York, 1967.

[8] B. S. GOH, *Necessary conditions for singular extremals involving multiple control variables*, this Journal, 4 (1966), pp. 716–731.

[9] D. H. JACOBSON, *A new necessary condition of optimality for singular control problems*, this Journal, 7 (1969), pp. 578–595.

[10] C. D. JOHNSON AND J. E. GIBSON, *Singular solutions in problems of optimal control*, IEEE Trans. Automatic Control, AC-8 (1963), pp. 4–15.

[11] W. M. WONHAM AND C. D. JOHNSON, *Optimal bang-bang control with quadratic performance index*, ASME Trans. J. Basic Engrg., 86 (1964), pp. 107–115.

[12] D. R. SNOW, *Singular optimal controls for a class of minimum effort problems*, this Journal, 2 (1964), pp. 203–219.

[13] C. D. JOHNSON, *Singular solutions in optimal control problems*, Advances in Control Systems, vol. 2, C. T. Leondes, ed., Academic Press, New York, 1965.

[14] M. ATHANS AND M. D. CANNON, *On the fuel optimal singular control of nonlinear second-order systems*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 360–370.

[15] R. W. BASS AND R. F. WEBER, *On synthesis of optimal bang-bang feedback control systems with quadratic performance criterion*, Proc. 6th Joint Automatic Control Conference, Troy, N.Y., 1965, pp. 213–219.

[16] H. HERMES, *Controllability and the singular problem*, this Journal, 2 (1964), pp. 241–260.

[17] H. HERMES AND G. W. HAYNES, *On the nonlinear control problem with control appearing linearly*, this Journal, 1 (1963), pp. 85–107.

[18] H. J. KELLEY, *A transformation approach to singular subarcs in optimal trajectory and control problems*, this Journal, 2 (1964), pp. 234–240.

[19] C. G. PFEIFFER, *Some new results in optimal final value control theory*, J. Franklin Inst., 283 (1967), pp. 102–119.

[20] B. S. GOH, *The second variation for the singular Bolza problem*, this Journal, 4 (1966), pp. 309–325.

[21] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp.102–119.

[22] D. H. JACOBSON AND D. Q. MAYNE, *Differential Dynamic Programming*, American Elsevier, New York, 1970.

[23] A. E. BRYSON AND Y. C. HO, *Applied Optimal Control*, Blaisdell, Waltham, Mass., 1969.

[24] D. H. JACOBSON, *Differential dynamic programming methods for solving bang-bang control problems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 661–675.

[25] P. DYER AND S. R. MCREYNOLDS, *On optimal control problems with discontinuities*, J. Math. Anal. Appl., 23 (1968), p. 585.

[26] J. V. BREAKWELL AND Y. C. HO, *On the conjugate point condition for the control problem*, Internat. J. Engrg. Sci., 2 (1965), pp. 565–579.

[27] I. M. GEL'FAND and S. V. FOMIN, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, N.J., 1963.

[28] I. B. VAPNYARSKII, *An existence theorem for optimal control in the Bolza problem; some of its applications and the necessary conditions for the optimality of moving and singular systems*, U.S.S.R. Comput. Math. and Math. Phys., 7 (1947), pp. 22–53; English transl., 1969.

[29] R. GABASOV, *Necessary conditions for optimality of singular control*, Engrg. Cybernetics, 1968, no. 5, pp. 28–37; English transl., 1969.

[30] I. B. VAPNYARSKII, *Solution of some optimal control problems*, Ibid., 1966, no. 4, pp. 37–43.

[31] V. I. GURMAN, *Methods of multiple maxima and the conditions of relative optimality of degenerate regimes*, Automat. Remote Control, 1967, no. 12, pp. 38–45.

[32] R. GABASOV, *On the theory of necessary optimality conditions governing special controls*, Soviet Physics Dokl., 13 (1969), no. 11, pp. 1094–1095.

[33] A. A. BOLONKIN, *Special extremals in the problem of optimal controls*, Engrg. Cybernetics, 1969, no. 2, pp. 187–198. (In Russian.)

SIAM J. CONTROL
Vol. 8, No. 3, August 1970

ABSTRACT CONTROL SYSTEMS: CONTROLLABILITY AND OBSERVABILITY*

VELIMIR JURDJEVIC†

Introduction. Conceptual ideas of control theory and dynamical systems have been of profound influence in motivating a general mathematical systems theory. Because of the somewhat special nature of the field of differential equations, within which control theory has been developed, it appears desirable to provide an axiomatic setting for the study of these notions; for instance, an axiomatic approach might, in addition to providing a link between dynamical systems and control theory, be of some value in the study of control systems described by partial differential equations. This paper is an attempt to formulate a set of axioms for dynamical control systems which are abstract generalizations of global solutions of control differential equations. In this context, I have made a study of controllability and observability of linear dynamical control systems.

Among the many generalizations of the concept of dynamical system (see Bushaw [4]) we will mention only a few. Probably one of the weakest structures was the concept of "processes" given by Hajek [10]. Roxin [18] considered "generalized dynamical systems" to be extensions of the classical dynamical systems (for instance Bhatia and Hajek [2]) where essentially the assumption of single-valuedness of solutions has been dropped. In regard to optimal control theory, formalization attempts were made by Bushaw [3] and Halkin [12]; both investigated implications of their "dynamical polysystems" axioms to optimization. Many other formalisms were, to quote Bushaw [4], "little more than preludes to discussions that are conducted in considerably less abstract terms." My axioms have essentially abstracted the global existence and uniqueness of solutions and are of sufficient generality to include a class of partial differential equations.

Section 1, in addition to providing the definitions and basic concepts, deals primarily with controllability of linear dynamical control systems. Its main result is the existence of an invariant approximately controllable linear subspace of a Hilbert state space. The notion of approximate controllability is weaker than that of Antosiewicz [1] or Fattorini [7]. It is shown that the approximately controllable space coincides with the controllable space (in the classical sense) whenever the state space is of finite dimension. From these results the analogue of Kalman's canonical decomposition theorem [14] follows immediately. It is an interesting consequence that all of these results are obtained in the absence of any topological structure on the class of controls.

Section 2 deals with linear observed systems. The notion of observability of a linear system is introduced via the concept of "indistinguishable states" and is due to M. Arbib. It is then proved that, corresponding to any linear observed system, there exists a unique invariant maximal subspace of the state space on which the system is unobservable. It is also proved that the definition of ob-

* Received by the editors July 15, 1969.

† IBM Corporation, Poughkeepsie, New York 12603, and Systems Research Center, Case Western Reserve University, Cleveland, Ohio 44106. Now a Fellow in Applied Mathematics at Harvard University, Cambridge, Massachusetts.

servability used here is equivalent to that used by Kalman [14] and Markus [17] whenever the state space has the structure of a Hilbert space.

1. Dynamical control systems.

1.1. Basic definitions. X is a nonempty set, to be referred to as the *state space*. Elements of X will be called *states*. Ω is a nonempty set, to be called the *restraint* set.

The set $F \subseteq \{f | f : R \to \Omega\}$ is called the set of *controls* and its elements will be called *controls*. R denotes the set of real numbers. F is said to be *closed under translations* if and only if, for any $f \in F$ and any $s \in R$, the map f_s defined by $f_s(t) = f(t - s)$ for all $t \in R$ is a member of F. F is *closed under concatenations* if and only if, for any f and g in F and any $s \in R$, the map h defined by

$$h(t) = \begin{cases} f(t) & \text{for } t < s, \\ g(t) & \text{for } t \geqq s \end{cases}$$

is a member of F.

$X \times R$ will be called the *phase space* and its elements will be called *phases*.

DEFINITION 1.1.1. Π is a *dynamical control system* if and only if Π satisfies the following axioms:

(A1) $\Pi : X \times F \times R \times R \to X$

(A2) $\Pi(x, f, t, t) = x$ for all $x \in X, f \in F$ and $t \in R$.

(A3) $\Pi(x, f, r, t) = \Pi(\Pi(x, f, r, s), f, s, t)$ for all $x \in X$, $f \in F$ and all r, s, t in R.

(A4) If, for some x and y in X, f and g in F, r, s, t, v in R, we have

$$\Pi(x, f, r, t) = \Pi(y, g, s, t),$$

and

$$f(u) = g(u) \text{ for } u \in (t, v),$$

then

$$\Pi(x, f, r, v) = \Pi(y, g, s, v).$$

Axioms (A2)–(A4) will be called the identity, homomorphism and nonanticipation axiom respectively.

A control $f \in F$ is said to *steer* a phase (x, s) into a phase (y, t) if and only if $s \leqq t$ and $\Pi(x, f, s, t) = y$. If U and V are subsets of X then f steers (U, s) into (V, t) if and only if f steers every $(x, s) \in (U, s)$ into some phase in (V, t).

The set $\mathscr{A}(x, s, t) = \{y | \Pi(x, f, s, t) = y \text{ for some } f \in F\}$ is called the *set of attainability* from (x, s) at t. Similarly, if $U \subseteq X$ then the set of attainability from (U, s) at t is defined as follows:

$$\mathscr{A}(U, s, t) = \bigcup_{x \in U} \mathscr{A}(x, s, t).$$

A dynamical control system Π is said to be *controllable* on a set $U \subseteq X$ at time $s \in R$ if and only if for any x and y in U there exists a control $f \in F$ which steers (x, s) into (y, t) for some $t \in R$.

A set $U \subseteq X$ is said to be *positively invariant* (under Π) if and only if

$$\bigcup_{t \geqq 0} \mathscr{A}(U, 0, t) \subseteq U.$$

Similarly, U is negatively invariant (under Π) if and only if

$$\bigcup_{t \leqq 0} \mathscr{A}(U, 0, t) \subseteq U.$$

U is *invariant* (under Π) if and only if U is both positively and negatively invariant (under Π). It follows directly from the definitions that any intersection of positively (negatively) invariant sets is positively (negatively) invariant.

DEFINITION 1.1.2. Φ is a *dynamical system* if and only if Φ satisfies the following axioms:

(B1) $\Phi: X \times R \times R \to X$.

(B2) $\Phi(x, t, t) = x$ for all $x \in X$ and $t \in R$.

(B3) $\Phi(x, r, t) = \Phi(\Phi(x, r, s), s, t)$ for all $x \in X$ and all r, s, t in R.

Let Π be a dynamical control system. For each $f \in F$ let $\Phi_f: X \times R \times R \to X$ be defined as follows:

$$\Phi_f(x, s, t) = \Pi(x, f, s, t) \qquad \text{for all } x \in X \quad \text{and} \quad \text{all } s, t \text{ in } R.$$

It follows directly from Definition 1.1.1 that Φ_f satisfies the conditions of Definition 1.1.2 and is, therefore, a dynamical system. Thus Π can be viewed as a collection of dynamical systems parameterized by the elements of F.

It is evident that a dynamical control system Π is a dynamical system if and only if

$$\Pi(x, f, s, t) = \Pi(x, g, s, t)$$

for all f, g in F, s, t in R and all $x \in X$.

DEFINITION 1.1.3. Let the set of controls F be closed under translations. A dynamical control system Π is said to be *autonomous* if and only if

$$\Pi(x, f, r, t) = \Pi(x, f_s, r - s, t - s)$$

for all $x \in X, f \in F$ and r, s, t in R.

It follows immediately from the above remarks that a dynamical system Φ is autonomous if and only if

$$\Phi(x, r, t) = \Phi(x, r - s, t - s)$$

for all $x \in X$ and r, s, t in R; or equivalently, Φ is autonomous if and only if there exists a function

$$\Phi': X \times R \to X$$

with

$$\Phi'(x, 0) = x, \qquad \Phi'(x, t) = \Phi'(\Phi'(x, s), t - s)$$

for all $x \in X$, s, t in R and $\Phi(x, s, t) = \Phi'(x, t - s)$.

Definition 1.1.1 is similar to Kalman's definition of a dynamical system [14]; it is more general in that it does not include any continuity requirements. Almost all of the terminology and basic concepts associated with a dynamical control system have been taken from control theory with the exception that in a few instances they have been defined more precisely.

The notion of a dynamical system in Definition 1.1.2 is different from the usual definition given in Bhatia and Hajek [2]. The latter, except for the continuity requirement, is what is here called an autonomous dynamical system. I hope that this choice of terminology adds to the clarity of these notions.

1.2. Elementary properties of affine and linear dynamical control systems. Let the state space X and the set of controls F be linear spaces defined over the same ring of scalars $\mathscr{R}$. It will be assumed that $\mathscr{R}$ contains a unit which will be denoted by 1.

DEFINITION 1.2.1. A dynamical control system Π is said to be *affine* if and only if

$$\alpha\Pi(x,f,s,t) + (1-\alpha)\Pi(y,g,s,t) = \Pi(\alpha x + (1-\alpha)y, \alpha f + (1-\alpha)g, s, t)$$

for all $\alpha \in \mathscr{R}$, x, y in X, f, g in F and s, t in R.

DEFINITION 1.2.2. A dynamical control system Π is *linear* if and only if

$$\alpha\Pi(x,f,s,t) + \beta\Pi(y,g,s,t) = \Pi(\alpha x + \beta y, \alpha f + \beta g, s, t)$$

for all α, β in $\mathscr{R}$, x, y in X, f, g in F and s, t in R.

THEOREM 1.2.3. *A dynamical control system* Π *is affine if and only if there exist* φ *and* ψ *such that*

$$\Pi(x,f,s,t) = \varphi(x,f,s,t) + \psi(s,t).$$

for all $x \in X, f \in F$ *and* s, t *in* R, *and that*:

(i) φ *is a linear dynamical control system*;

(ii) $\psi : R \times R \to R$ *is such that* $\psi(r,t) = \varphi(\psi(r,s), 0, r, t) + \psi(s,t)$ *for all* r, s, t in R.

Proof. If $\Pi(x,f,s,t) = \varphi(x,f,s,t) + \psi(s,t)$ as given in the conditions of the theorem, then it follows by direct verification that Π is an affine dynamical control system.

To prove the converse statement of the theorem assume that Π is an affine dynamical control system. Let $\psi : R \times R \to X$ be defined by $\psi(s,t) = \Pi(0,0,s,t)$ for all $(s,t) \in R \times R$, and let $\varphi : X \times F \times R \times R \to X$ be defined by $\varphi(x,f,s,t) = \Pi(x,f,s,t) - \psi(s,t)$ for all $x \in X, f \in F$ and s, t in R.

It is a laborious but straightforward matter to show that φ is a linear dynamical control system, and its proof will be omitted. Since it follows directly from the homomorphism axiom and linearity of φ that ψ satisfies (ii), the proof of the theorem is complete.

Incidentally, since $\psi(s,t)$ must be equal to $\Pi(0,0,s,t)$ it follows that φ and ψ are uniquely determined by Π.

This remark essentially proves the following corollary.

COROLLARY 1.2.4. *An affine dynamical control system* Π *is linear if and only if* $\Pi(0,0,s,t) = 0$ *for all* s, t *in* R.

THEOREM 1.2.5. *A dynamical control system* Π *is linear if and only if there exist* Φ *and* Λ *such that*

$$\Pi(x,f,s,t) = \Phi(x,s,t) + \Lambda(f,s,t)$$

for all $x \in X, f \in F$ *and* s, t *in* R, *and that*:

(i) Φ *is a linear dynamical system*;

(ii) $\Lambda : F \times R \times R \to X$ *is such that, for any* f, g *in* F *and any* α, β *in* $\mathscr{R}$,

$$\alpha\Lambda(f,s,t) + \beta\Lambda(g,s,t) = \Lambda(\alpha f + \beta g, s, t) \quad \text{for all } s, t \text{ in } R;$$

(iii) $\Lambda(f,t,t) = 0$ *for all* $f \in F$, $t \in R$;

(iv) $\Lambda(f,r,t) = \Phi(\Lambda(f,r,s),s,t) + \Lambda(f,s,t)$ *for all* $f \in F$ *and* r,s,t *in* R;

(v) $\Lambda(f,r,t) = \Lambda(g,r,t)$ *for all* f *and* g *in* F *with* $g(s) = f(s)$ *for all* $s \in (r,t)$, $t > r$.

Proof. Let $\Pi(x,f,s,t) = \Phi(x,s,t) + \Lambda(f,s,t)$ for all $x \in X, f \in F$ and s,t in R where Φ and Λ satisfy (i)–(v) respectively. It follows by direct verification that Π satisfies all of the axioms of a linear dynamical control system. To prove the other implication of the theorem, assume that Π is a linear dynamic control system. Define $\Phi: X \times R \times R \to X$ as follows:

$$\Phi(x,s,t) = \Pi(x,0,s,t) \qquad \text{for all } x \in X \quad \text{and} \quad s,t \text{ in } R.$$

Φ is a dynamical system as was observed in § 1.1 and its linearity follows directly from the linearity of Π.

Define $\Lambda: F \times R \times R \to X$ by

$$\Lambda(f,s,t) = \Pi(0,f,s,t) \qquad \text{for all } f \in F \quad \text{and} \quad s,t \text{ in } R.$$

Property (ii) is a direct consequence of the linearity of Π, and properties (iii) and (v) follow immediately from the identity and nonanticipation properties of Π respectively.

To verify property (iv), let $f \in F$ and r,s,t be in R. Then

$$\begin{aligned}\Lambda(f,r,t) &= \Pi(0,f,r,t) = \Pi(\Pi(0,f,r,s),f,s,t)\\ &= \Pi(\Lambda(f,r,s),f,s,t) = \Phi(\Lambda(f,r,s),s,t) + \Lambda(f,s,t).\end{aligned}$$

This proves (iv) and now the proof of the theorem is completed.

COROLLARY 1.2.6. *A dynamical control system* Π *is affine if and only if there exist* Φ, Λ, ψ *such that*

$$\Pi(x,f,s,t) = \Phi(x,s,t) + \Lambda(f,s,t) + \psi(s,t)$$

for all $x \in X, f \in F$ *and* s,t *in* R, *where* Φ *and* Λ *satisfy the conditions of Theorem* 1.2.4, *and* ψ *satisfies condition* (ii) *of Theorem* 1.2.3.

Proof. This is a direct consequence of Theorems 1.2.3 and 1.2.4.

The following corollary is immediate.

COROLLARY 1.2.7. *An autonomous dynamical control system* Π *is affine if and only if there exist* Φ, Λ, ψ *such that*

$$\Pi(x,f,s,t) = \Phi(x,t-s) + \Lambda(f,s,t) + \psi(t-s)$$

for all $x \in X, f \in F$ *and* s,t *in* R, *where*:

(i) Φ *is a linear autonomous dynamical system*;

(ii) Λ *satisfies conditions* (ii), (iii), (iv) *and* (v) *of Theorem* 1.2.5 *and in addition* $\Lambda(f,r,t) = \Lambda(f_s, r-s, t-s)$ *for all* $f \in F$ *and* r,s,t *in* R;

(iii) $\psi: R \to X$ *with* $\psi(0) = 0$ *and* $\psi(t-r) = \Phi(\psi(s-r), t-s) + \psi(t-s)$ *for all* r,s *and* t *in* R.

1.3. Controllability of linear dynamical control systems. Let the state space X and the set of controls F be linear spaces over the same ring of scalars $\mathscr{R}$. Even though a few results of this section do not require F to be closed under translations and concatenations, we shall assume this property throughout the section.

Corresponding to any linear dynamical control system Π, the symbols Φ and Λ will have the meaning from Theorem 1.2.5 (and will not be necessarily defined whenever used). Φ will be referred to as the *zero response* of Π.

DEFINITION 1.3.1. Let Π be a linear dynamical control system. For each $s \in R$, the set

$$\mathcal{N}(s) = \{x | \Pi(x, f, s, t) = 0 \text{ for some } f \in F \text{ and some } t \geqq s\}$$

is called the *domain of null controllability* of Π at s.

Clearly $x \in \mathcal{N}(s)$ if and only if $\Phi(x, s, t) + \Lambda(f, s, t) = 0$ for some $f \in F$ and $t \geqq s$, or equivalently,

$$x = -\Phi(\Lambda(f, s, t), t, s) \quad \text{for some } f \in F \quad \text{and} \quad t \geqq s.$$

If $f \in F$ steers a phase (x, r) into $(0, s)$ for some $s \geqq r$, then $g: R \to \Omega$, defined by $g(u) = f(u)$ for $u < s$ and $g(u) = 0$ for $u \geqq s$, steers (x, r) into $(0, t)$ for any $t \geqq s$. This remark along with linearity of Π essentially proves the next theorem.

THEOREM 1.3.2. *Let Π be a linear dynamical control system. Then for each $s \in R$, the domain of null controllability $\mathcal{N}(s)$ of Π at s is a linear subspace of X.*

If Π is an autonomous dynamical control system, then it follows immediately that for any s and t in R, $\mathcal{N}(s) = \mathcal{N}(t)$; thus in the sequel no reference to time will be made in the description of the domain of null controllability $\mathcal{N}$ of a linear autonomous dynamical control system.

THEOREM 1.3.3. *If Π is an autonomous linear dynamical control system, then its domain of null controllability $\mathcal{N}$ is negatively invariant.*

Proof. Let $y = \Pi(x, f, 0, -t)$ with $x \in \mathcal{N}, f \in F$ and $t \in R$, $t > 0$. We want to show that $y \in \mathcal{N}$. Let $g \in F$ steer $(x, 0)$ into $(0, u)$. Define $h: R \to \Omega$ by letting

$$h(s) = \begin{cases} f(s) & \text{for } s < 0, \\ g(s) & \text{for } s \geqq 0. \end{cases}$$

Thus $h \in F$, and

$$\begin{aligned} \Pi(y, h, -t, u) &= \Pi(\Pi(\Pi(x, f, 0, -t), f, -t, 0), g, 0, u) \\ &= \Pi(x, g, 0, u) = 0. \end{aligned}$$

This shows that h steers $(y, -t)$ into $(0, u)$, and therefore $y \in \mathcal{N}$. This concludes the proof of the theorem.

DEFINITION 1.3.4. Let Π be a dynamical control system. For any $s \in R$,

$$\mathcal{A}(0, s) = \{x : \Pi(0, f, s, t) = x \text{ for some } f \in F \text{ and } t \geqq s\}$$

is called the set of attainability from $(0, s)$.

Equivalently, $x \in \mathcal{A}(0, s)$ if and only if there exists a control $f \in F$ which steers $(0, s)$ into (x, t) for some $t \in R$. If Π is a linear dynamical control system, then $x \in \mathcal{A}(0, s)$ if and only if

$$x = \Lambda(f, s, t)$$

for some $f \in F$ and some $t \geqq s$.

If $f \in F$ steers $(0, r)$ into (x, s) for some $s \geqq r$, then $g: R \to \Omega$, defined by $g(u) = 0$ for $u < r + t - s$ and $g(u) = f(u + t - s)$ for $u \geqq r + t - s$, steers $(0, r)$

into (x, t) for any $t \geqq s$. Thus in a manner completely analogous to Theorems 1.3.2 and 1.3.3 we get the following theorems.

THEOREM 1.3.5. *Let Π be a linear autonomous dynamical control system. For each $s \in R$, the set of attainability $\mathscr{A}(0, s)$ is a linear subspace of X.*

THEOREM 1.3.6. *If Π is a linear autonomous dynamical control system, then the set $\mathscr{A}$ of attainability from zero is positively invariant.*

Clearly, if Π is autonomous then $\mathscr{A}(0, t) = \mathscr{A}(0, s)$ for any s, t in R and therefore in the sequel, corresponding to any linear autonomous dynamical control system the set of attainability from $(0, s)$ will be simply called the set of attainability from zero and it will be denoted by $\mathscr{A}$.

All of the previous notions along with their properties have been developed solely through the structural and algebraic properties of X, F and Π. The following concept of approximate controllability will, however, be of topological nature. For this purpose it will be assumed that X is a linear topological space.

DEFINITION 1.3.7. A linear dynamical control system Π is said to be *approximately controllable* on a set $A \subseteq X$ at time $s \in R$ if and only if there exists a set $B \subseteq A$, such that $\bar{B} = \bar{A}$ and that, for any $x \in A$, $y \in B$ and any neighborhood U of x, there exists a control $f \in F$ which steers (y, s) into (U, t) for some $t \in R$.

If Π is approximately controllable on A (at time s), then Π is approximately controllable on any subset of A which contains B. If Π is an autonomous dynamical control system, and if Π is approximately controllable on A at time s, then Π is approximately controllable on A at every time $t \in R$. In such cases no reference to time will be made, and Π will be said to be approximately controllable on A.

LEMMA 1.3.8. *Let Π be an autonomous linear dynamical control system. If for each $t \in R$, the zero response Φ of Π is a continuous function of x, then the closure of the set of attainability $\bar{\mathscr{A}}$ is positively invariant, and the closure of the domain of null controllability $\bar{\mathscr{N}}$ is negatively invariant.*

Proof. We first show that for each $t \geqq 0$ in R,

$$\Phi(\bar{\mathscr{A}}, t) \subseteq \bar{\mathscr{A}}.$$

Let $x \in \bar{\mathscr{A}}$, and let U be any open neighborhood of $\Phi(x, t)$. By continuity of Φ, $V = \{y \in X : \Phi(y, t) \in U\}$ is an open neighborhood of x. Let $y \in \mathscr{A} \cap V$. By positive invariance of $\mathscr{A}$, $\Phi(y, t) \in \mathscr{A} \cap U$. This shows that any neighborhood U of $\Phi(x, t)$ contains a point of $\mathscr{A}$, which implies that $\Phi(x, t) \in \bar{\mathscr{A}}$ or that $\Phi(\bar{\mathscr{A}}, t) \subseteq \bar{\mathscr{A}}$.

Now let $x \in \bar{\mathscr{A}}$, $f \in F$ and $t \geqq 0$. Then $\Lambda(f, 0, t) \in \mathscr{A}$ by the definition of $\mathscr{A}$. Since $\mathscr{A}$, and hence also $\bar{\mathscr{A}}$, is a linear subspace of X, we obtain that

$$\Pi(x, f, 0, t) = \Phi(x, t) + \Lambda(f, 0, t) \in \bar{\mathscr{A}}.$$

This proves positive invariance of $\bar{\mathscr{A}}$.

The proof of negative invariance of $\bar{\mathscr{N}}$ is completely analogous, and therefore, the proof of the lemma is complete.

LEMMA 1.3.9. *Let Φ be a linear autonomous dynamical system defined on a Banach space X. Assume that Φ satisfies the following continuity requirement: Given $\varepsilon > 0$ and $s \in R$, there exists $\delta > 0$ such that, for all $t \in R$ with $|t - s| < \delta$,*

$$\sup_{\|x\| \leqq 1} \|\Phi(x, t) - \Phi(x, s)\| < \varepsilon.$$

Then for each $x \in X$, $\Phi(x,t)$ is an analytic function of t; furthermore, there exists a bounded transformation A on X such that for all $x \in X$ and all $t \in R$

$$\Phi(x,t) = \sum_{i=0}^{\infty} \frac{t^i}{i!} A^i(x) \equiv e^{tA}(x).$$

(For proof see Hille and Phillips [13, pp. 338–342].)

Any autonomous linear dynamical system having the above continuity requirement will be termed *uniformly continuous.* This is in accordance with the terminology used in the context of semigroups by Hille and Phillips [13] and Yosida [20].

Since $\|\Phi(x,t) - y\| \leqq \|\Phi(x,t) - x\| + \|y - x\|$, we have that uniform continuity of Φ implies joint continuity of Φ; i.e., whenever $x_n \to x$ and $t_n \to t$,

$$\|\Phi(x_n, t_n) - \Phi(x,t)\| \to 0.$$

THEOREM 1.3.10. *Let Π be a linear autonomous dynamical control system defined on a Hilbert space X. Let the zero response Φ of Π be uniformly continuous. Then there exists an invariant closed linear subspace C of X such that Π is approximately controllable on C.*

Proof. Let $C = \mathscr{A}$. By Theorems 1.3.5, 1.3.6 and Lemma 1.3.8, C is a positively invariant closed subspace of X. Let $C^\perp$ denote the orthogonal complement of C. Then C is a closed subspace of X, and $X = C \oplus C^\perp$. Let $P_1 : X \to C$ and $P_2 : X \to C^\perp$ be the projection maps. For $k = 1, 2$ set $\Pi_k = P_k \cdot \Pi$, $\Phi_k = P_k \cdot \Phi$ and $\Lambda_K = P_K \cdot \Lambda$. Let Φ_{k1}, Φ_{k2} be the restrictions of Φ_K to C and $C^\perp$ respectively; from linearity of Φ,

$$\Pi_1(x,f,s,t) = \Phi_{11}(y, t-s) + \Phi_{12}(z, t-s) + \Lambda_1(f,s,t),$$

$$\Pi_2(x,f,s,t) = \Phi_{21}(y, t-s) + \Phi_{22}(z, t-s) + \Lambda_2(f,s,t)$$

whenever $x = y + z$ with $y \in C$, $z \in C^\perp$, and $f \in F$, s, t in R. Since C is positively invariant, $\Pi_2(y, f, 0, t) = 0$ for $y \in C$, $f \in F$ and $t \geqq 0$; hence $\Phi_{21}(y,t) = 0$ and $\Lambda_2(f, 0, t) = 0$. By Lemma 1.3.9, $\Phi(y, \cdot)$ is an analytic function of t, and hence so is $\Phi_{21}(y, \cdot)$. Therefore, $\Phi_{21}(y,t) \equiv 0$ for all $t \in R$.

For $z \in C^\perp$ and s, t in R we have

$$\begin{aligned}\Phi_{12}(z, s+t) + \Phi_{22}(z, s+t) &= \Phi(z, s+t) \\ &= \Phi(\Phi(z,s), t) \\ &= \Phi_{11}(\Phi_{12}(z,s), t) + \Phi_{12}(\Phi_{22}(z,s), t) \\ &\quad + \Phi_{22}(\Phi_{22}(z,s), t)\end{aligned}$$

so that

$$\Phi_{22}(z, s+t) = \Phi_{22}(\Phi_{22}(z,s), t);$$

thus, Φ_{22} is a dynamical system on $C^\perp$. In particular, for any $f \in F$ and $t > 0$,

$$\begin{aligned}\Lambda_2(f, 0, -t) &= -\Phi_{22}(\Lambda_2(f_t, 0, t), -t) \\ &= -\Phi_{22}(0, -t) = 0;\end{aligned}$$

thus, $\Lambda_2(f, 0, t) = 0$ for all $f \in F$ and $t \in R$. Thus we have shown that for any $y \in C$, $\Pi_2(y, f, 0, t) = 0$ for all $f \in F$ and $t \in R$. Thus $\Pi(y, f, 0, t) = \Pi_1(y, f, 0, t) \in C$,

and hence C is invariant under Π. In the course of proving the above assertion we have also proved that Π has the following representation:

$$\Pi_1(x,f,s,t) = \Phi_{11}(y,t-s) + \Phi_{12}(z,t-s) + \Lambda_1(f,s,t),$$
$$\Pi_2(x,f,s,t) = \Phi_{22}(z,t-s)$$

for all $x \in X$ with $x = y + z$, $y \in C$, $z \in C^\perp$ and all $f \in F, s, t$ in R. For such z, if $f \in F$ steers $(x,0)$ into $(0,t)$ for some $t \in R$, then $\Phi_{22}(z,t) = 0$; since Φ_{22} is a linear dynamical system, necessarily $z = 0$. This shows that $\mathscr{N} \subseteq C = \bar{\mathscr{A}}$, and therefore, $\bar{\mathscr{N}} \subseteq C$.

$\bar{\mathscr{N}}$ is a negatively invariant closed subspace of X (Theorems 1.3.2, 1.3.3, and Lemma 1.3.8); analogous arguments yield $C = \bar{\mathscr{A}} \subseteq \bar{\mathscr{N}}$, and therefore,

$$C = \bar{\mathscr{A}} = \bar{\mathscr{N}}.$$

This will be used now to show that Π restricted to C is approximately controllable.

First, $\mathscr{N}$ is dense in C. Let $x \in C$, let $y \in \mathscr{N}$, and let U be any neighborhood of x. We must show that there exists a control $h \in F$ which steers $(y,0)$ into (U,t) for some $t \in R$. Since $\mathscr{A}$ is dense in C, there exists

$$w \in \mathscr{A} \cap U.$$

Let $f \in F$ steer $(y,0)$ into $(0,s)$ for some $s \in R$ and let $g \in F$ steer $(0,0)$ into (w,t) for some $t \in R$.

Define $h: R \to \Omega$ by

$$h(u) = \begin{cases} f(u) & \text{for } u < s, \\ g(u+s) & \text{for } u \geqq s. \end{cases}$$

Then $h \in F$ and h steers $(y,0)$ into $(w, s+t)$. This shows that Π restricted to C is approximately controllable and the proof of the theorem is completed.

In the course of this proof we have also obtained the following corollary.

COROLLARY 1.3.11. *If Π is an autonomous linear dynamical control system defined on a Hilbert space X and if its zero response is uniformly continuous, then the closure of the domain of null controllability is equal to the closure of the set of attainability from zero.*

THEOREM 1.3.12. *Let Π be an autonomous linear dynamical control system defined on a Hilbert state space X. Let the zero response Φ of Π be uniformly continuous.*

If either $\Phi(\cdot,t)$ is a self-adjoint operator on X for each $t \in R$, or $X = R^n, n \geqq 1$, then a space C satisfying the conditions of Theorem 1.3.10 *is unique.*

Proof. Let Π be a linear autonomous dynamical control system defined on a Hilbert space X with Φ uniformly continuous.

Let C be the closure of the set of attainability from zero. By Theorem 1.3.10 C is invariant under Π, Π restricted to C is approximately controllable, and furthermore, Π has the following representation:

$$\Pi_1(x,f,s,t) = \Phi_{11}(y,t-s) + \Phi_{12}(z,t-s) + \Lambda_1(f,s,t),$$
$$\Pi_2(x,f,s,t) = \Phi_{22}(z,t-s),$$

where $x = y + z, y \in C, z \in C^\perp$.

Let D be any other closed, invariant subspace of X on which Π is approximately controllable.

Let B be a set dense in D with the property that for any $x \in D$, $y \in B$ and any neighborhood U of x, there exists a control $f \in F$ which steers $(y, 0)$ into (U, t) for some $t \in R$.

By invariance of D, it immediately follows that $\mathscr{A} \subseteq D$, so that $C \subseteq D$. Let $y \in B, y = z + w$ with $z \in C$ and $w \in C^{\perp}$. Since $B \subseteq D$, we have $y \in D$ and therefore $-y \in D$.

From the controllability properties of D it follows that, for some sequences $\{f(n)\} \subseteq F, t_n \geqq 0$,

$$\lim_{n \to \infty} \Pi(y, f(n), 0, t_n) = -y.$$

This implies that $\lim_{n\to\infty} \Pi_2(y, f(n), 0, t_n) = \lim_{n\to\infty} \Phi_{22}(w, t_n) = -w$.

Let $(\cdot, \cdot)$ denote the inner product on X. Observe that if $\Phi(\cdot, t)$ is a self-adjoint operator in X, then so is $\Phi_{22}(\cdot, t)$. For any $t \in R$,

$$\begin{aligned}(\Phi_{22}(w, t), w) &= (\Phi_{22}(w, t/2 + t/2), w) = (\Phi_{22}(w, t/2), \Phi_{22}^*(w, t/2)) \\ &= (\Phi_{22}(w, t/2), \Phi_{22}(w, t/2)) = \|\Phi_{22}(w, t/2)\|^2 \geqq 0.\end{aligned}$$

But,

$$\lim_{n \to \infty} (\Phi_{22}(w, t_n), w) = (\lim \Phi_{22}(w, t_n), w) = (-w, w) = -\|w\|^2 \leqq 0.$$

This shows that $w = 0$, and therefore, $B \subseteq C$. Since $\bar{B} = D$ it follows that $D \subseteq C$, and hence $D = C$. This completes the proof of the first case.

For the second let $X = R^n, n \geqq 1$. Set

$$E = D \cap C.$$

E is a closed linear subspace of X with the property that for any $x \in E$ and any $y \in B$ with $y = z + w, z \in C, w \in C^{\perp}$, there exists a sequence $t_n \geqq 0$ with $\lim_{n\to\infty} \Phi_{22}(w, t_n) = 0$. If $t_n \leqq M < \infty$ for all n, then $w = 0$, and therefore, $B \subseteq C$. Since $\bar{B} = D$ we have that $D \subseteq C$, and therefore, $D = C$. So assume $t_n \to \infty$. From our assumption on $X = R^n$, $\Phi_{22}(w, t) = e^{A_{22}t}(w)$ where A_{22} is an $m \times m$ matrix, m being the dimension of $C^{\perp}$. Let S be the nonsingular matrix of dimension m such that

$$e^{A_{22}t}w = e^{S(D+N)tS^{-1}}(w) = Se^{(D+N)t}S^{-1}(w),$$

where D and N are diagonal and nilpotent matrices respectively. This implies that $\lim_{n\to\infty} \Phi_{22}(w, t_n) = 0$ if and only if, for each eigenvalue λ_i of A_{22}, $\operatorname{Re} \lambda_i < 0$ in which case

$$\lim_{t \to \infty} \Phi_{22}(w, t) = 0$$

and therefore $E = \{0\}$. This implies that $D \subseteq C$, and hence $C = D$. This concludes the proof of the theorem.

The space C of Theorems 1.3.10 and 1.3.12 will be called the *space of approximate controllability* of Π.

COROLLARY 1.3.13. *Let* Π *be a linear autonomous dynamical control system defined on* $X = R^n$, $n \geqq 1$. *If the zero response* Φ *of* Π *is uniformly continuous in* t, *then* Π *restricted to the space of approximate controllability is controllable.*

Proof. Let Π, X and Φ satisfy the conditions.

Every subspace of a finite-dimensional state space is closed so that from Corollary 1.3.11

$$\mathscr{N} = \bar{\mathscr{N}} = C = \bar{\mathscr{A}} = \mathscr{A}.$$

Let x and y be in C. Let $f \in F$ steer $(x, 0)$ into $(0, s)$ for some $s \in R$ and let $g \in F$ steer $(0, 0)$ into (y, t) for some $t \in R$.

Define $h : R \to \Omega$ by

$$h(u) = \begin{cases} f(u) & \text{for } u \leqq s, \\ g(u + s) & \text{for } u \geqq s. \end{cases}$$

Then $h \in f$, and h steers $(x, 0)$ into $(y, s + t)$. Therefore Π restricted to C is indeed controllable.

COROLLARY 1.3.14. *Let* Π *be a linear autonomous dynamical control system defined on* R^n. *If the zero response* Φ *of* Π *is uniformly continuous in* t, *then there exists a time* $T > 0$ *such that any two states in the controllable space can be steered one to the other in time at most* T.

Proof. Let Π, X and Φ satisfy the above conditions. Let $\{x_1, x_2, \cdots, x_m\} \subseteq X$, $m \leqq n$, be a basis for $\mathscr{A}$. Let $t_i \in R$ be such that $(0, 0)$ can be steered to (x_i, t_i) for $i = 1, 2, \cdots, m$, and set $T = \max t_i$, $1 \leqq i \leqq m$. Then, $(0, 0)$ can be steered to (x_i, T) for $i = 1, 2, \cdots, m$; let $f_i \in F$ steer $(0, 0)$ into (x_i, T), i.e.,

$$\Lambda(f_i, 0, T) = x_i.$$

Every $x \in \mathscr{A}$ has the form $x = \sum_{i=1}^m \lambda_i x_i$ for some λ_i; set $f = \sum_{i=1}^m \lambda_i f_i$. Then $\Lambda(f, 0, T) = x$, and therefore f steers $(0, 0)$ into (x, T).

Let x and y be any states in C. From invariance of C, $y - \Phi(x, T) \in C$ and therefore there exists $f \in F$ which steers $(0, 0)$ into $(y - \Phi(x, T), T)$, i.e.,

$$y - \Phi(x, T) = \Lambda(f, 0, T).$$

This shows that f steers $(x, 0)$ into (y, T); the proof of the corollary is completed.

Remark. Theorem 1.3.12 gives some sufficient conditions for the uniqueness of an approximately controllable space. I conjecture that, even under the hypothesis of Theorem 1.3.10, an approximately controllable space is unique.

1.4. Examples. In addition to the linear control differential equation in R^n (whose controllability properties are well known), the vibrating string equation provides an interesting example of a linear dynamical control system in infinite-dimensional spaces.

Example 1.4.1 (The vibrating string equation). Let the space X be the space of all pairs (x_1, x_2) where $x_k : R \to R$ is bounded and has a bounded and absolutely continuous derivative $(k = 1, 2)$. For $x \in X$, let the norm of x be defined by the following:

$$\|x\| = \max \left\{ \sup_{\xi \in R} |x_1(\xi)|, \sup_{\xi \in R} |x_2(\xi)|, \sup_{\xi \in R} \left| \frac{dx_1(\xi)}{d\xi} \right|, \sup_{\xi \in R} \left| \frac{dx_2(\xi)}{d\xi} \right| \right\}.$$

With this norm (and the obvious linear space structure) X becomes a Banach space.

Let the restraint set Ω be the space of all real, bounded and differentiable functions on R and let the set of controls F consist of all piecewise constant $f: R \to \Omega$; F is closed under translations and concatenations.

Let $\Phi: X \times R \to X$ be defined by

$$\Phi(t)x = \begin{pmatrix} \mu'(t) & \mu(t) \\ \mu''(t) & \mu'(t) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

for all $t \in R$ and $x \in X$, where

$$[\mu(t)x](\xi) = \frac{1}{2} \int_{\xi - t}^{\xi + t} x(\tau)\, d\tau$$

for all $\xi \in R$ and all bounded and measurable $x: R \to R$ and $\mu'(t)$ stands for the time derivative of μ.

After a somewhat laborious but straightforward computation, it can be shown that Φ is a linear autonomous dynamical system on X.

Using Φ define a dynamical control system Π on $X \times F \times R \times R$ as follows:

$$\Pi(x, f, s, t) = \Phi(t - s)x + \int_s^t \Phi(t - s - \tau)\tilde{f}(\tau)\, d\tau$$

for all $x \in X, s, t$ in R and $f \in F$ with $\tilde{f} = \begin{pmatrix} 0 \\ f \end{pmatrix}$. Π is an autonomous linear dynamical control system.

If Π_1 and Π_2 are the components of Π, then it can be verified directly that:

(i) $(\partial \Pi_2 / \partial t)(x, f, s, t) = \Pi_1(x, f, s, t)$ for all $t \in R$;

(ii) for each $x \in X$ and $f \in F$, Π_1 satisfies the partial differential equation

$$\frac{\partial^2 y(\xi, t)}{\partial t^2} - \frac{\partial^2 y(\xi, t)}{\partial \xi^2} = f(t)(\xi)$$

with $y(\xi, s) = x_1(\xi)$ and $(\partial y / \partial t)(\xi, s) = x_2(\xi)$ for all $\xi \in R$.

This is a well-known inhomogeneous vibrating string equation; it belongs to the class of partial differential equations of hyperbolic type. Actually it can be shown that solutions of hyperbolic equations in general are examples of dynamical systems.

The next example gives a dynamical control system defined on an infinite-dimensional Hilbert space but where the system is not induced by a partial differential equation.

Example 1.4.2 (Translation example). Let the state space $X = L_2(-\infty, \infty)$; let the set of controls F consist of all $f \in L_2(-\infty, \infty)$ which are bounded and have compact support.

Define $\Pi: X \times F \times R \times R \to X$ as follows:

$$\Pi(x, f, s, t)(\xi) = x(\xi - (t - s)) + \int_s^t f(\xi - (t - \tau))\, d\tau$$

for all $x \in X, f \in F, s, t$ in R and all $\xi \in R$.

For every $f \in F$ there exists an $M < \infty$ and $a < b$ such that $|f(\xi)| < M$ for all $\xi \in R$ and $\{\xi : f(\xi) \neq 0\} \subseteq [a, b]$. This implies that for any $s \in t$ in R

$$\int_{-\infty}^{\infty} \left| \int_s^t f(\xi - (t - \tau))\, d\tau \right|^2 d\xi \leqq \int_{-\infty}^{\infty} \left[\int_{\xi-(t-s)}^{\xi} |f(\tau)|\, d\tau \right]^2 d\xi$$

$$\leqq M^2(t - s)^2(b + t - s - a) < \infty.$$

Thus

$$\|\Pi(x, f, s, t)\| \leqq \|x\|^2 + M^2(t - s)^2(b + t - s - a) < \infty,$$

and therefore Π is indeed a mapping from $X \times F \times R \times R$ into X.

It can be directly verified that Π satisfies all the axioms of a linear autonomous dynamical control system. A point x is in the set of attainability from zero if and only if for some $f \in F$ and $t > 0$

$$x(\xi) = \int_0^t f(\xi - (t - \mu))\, d\mu = \int_{\xi-t}^{\xi} f(\mu)\, d\mu$$

for all $\xi \in R$.

Let

$$B = \left\{ x \in X : \frac{dx}{d\xi} \in L_2(-\infty, \infty), \frac{dx}{d\xi} \text{ has compact support} \right\}.$$

For any $x \in B$ and any $t > 0$ there exists an $f \in F$ which steers $(0, 0)$ into (x, t); thus $B \subseteq \mathscr{A}$, and since $\bar{B} = X$, it follows that

$$\bar{\mathscr{A}} = X.$$

Similarly one can show that the domain of null controllability $\mathscr{N}$ satisfies $\bar{\mathscr{N}} = X$. Therefore Π is approximately controllable on X.

2. Observed linear systems.

2.1. Basic definitions. Let Z be a linear topological space over the ring of scalars $\mathscr{R}$. Z will be called the *output space*. $W = \{f : R \to Z\}$ will be called the *space of output functions*. We assume that the set of controls $F \subseteq \Omega^R$ is a linear space over $\mathscr{R}$ and that it is closed under translations.

DEFINITION 2.1.1. A relation $S \subseteq F \times W$ is an *observed linear system* if and only if there exists a triple (X, Π, H) such that:

(C1) (i) X is a linear topological space (the state space).

(ii) Π is an autonomous linear dynamical control system on $X \times F \times R \times R$.

(iii) $H : X \to Z$ is linear and continuous.

(C2) Given any $(f, w) \in F \times W$, $(f, w) \in S$ if and only if $w(t) = H\Pi(x, f, 0, t)$ for some $x \in X$ and all $t \in R$.

Any triple (X, Π, H) with (C1) and (C2) will be called a *state representation* for S. Evidently a linear observed system has many state representations.

DEFINITION 2.1.2. Let (X, Π, H) be a state representation for an observed linear system S.

Then states x and y are said to be *indistinguishable* under Π and H (and this will be denoted by $x \sim y$) if and only if

$$H\Pi(x, f, 0, t) = H\Pi(y, f, 0, t)$$

for all $f \in F$ and $t \in R$. It follows immediately that $x \sim y$ if and only if

$$H\Phi(x - y, t) = 0$$

for all $t \in R$, where Φ is the zero response of Π. Clearly, $\sim$ is an equivalence relation on X. For each $x \in X$ let $[x]$ denote the equivalence class of x under $\sim$ and let $\tilde{X} = X/\sim$.

S is said to be *completely observable* in X if and only if $[x] = \{x\}$ for all $x \in X$. S is called *completely unobservable* in X if and only if $X = \{X\}$; or, equivalently, $[x] = [y]$ for all x and y in X. Since in that case $[0] = [x]$ for all $x \in X$ (and Φ maps onto), it follows that S is completely unobservable in X if and only if $H \equiv 0$.

2.2. Observability; decomposition theorem.

THEOREM 2.2.1. *Let (X, Π, H) be a state representation for an observed linear system S. Then there exists a unique subspace U of X with the following properties:*

(i) *U is invariant under Φ, i.e., $\Phi(U, t) \subseteq U$ for all $t \in R$;*

(ii) *$H(x) = 0$ for all $x \in U$;*

(iii) *U is a maximal subspace of X with properties* (i) *and* (ii).

Proof. Let $U = \{x \in X : H \cdot \Phi(x, t) = 0 \text{ for all } t \in R\}$.

It follows immediately from linearity that U is a linear subspace of X. To show (i) let $x \in U$ and let $s \in R$. Since

$$H\Phi(\Phi(x, s), t) = H\Phi(x, t + s) = 0$$

for all $t \in R$, we get that $\Phi(x, s) \in U$. This shows that indeed $\Phi(U, s) \subseteq U$. (ii) follows immediately from the fact that, for all $x \in U$,

$$Hx = H\Phi(x, 0) = 0.$$

To show (iii), let V be any subspace of X with properties (i) and (ii). Let $v \in V$. Then $\Phi(v, t) \in V$ by (i), and $H \cdot V \equiv 0$ by (ii); therefore, $H\Phi(v, t) = 0$ for all $t \in R$. This shows that $v \in U$ and hence $V \subseteq U$. This proves (iii). Of course, uniqueness is an immediate consequence of maximality. This concludes the proof.

THEOREM 2.2.2. *Let (X, Π, H) be a state representation for an observed linear system S, where X is a Hilbert space and the zero response Φ of Π is continuous on $X \times R$.*

Then S is completely observable in X if and only if there do not exist subspaces U and V of X such that

(i) $X = U \oplus V, \qquad U \neq 0;$

(ii) $$\Pi_1(x, f, s, t) = \Phi_{11}(u, t - s) + \Phi_{12}(v, t - s) + \Lambda_1(f, s, t),$$
$$\Pi_2(x, f, s, t) = \Phi_{22}(v, t - s) + \Lambda_2(f, s, t)$$

for all $x \in X$ with $x = u + v$, $u \in U$, $v \in V$, $f \in F$ and s, t in R; here Π_1, Π_2 and Λ_1, Λ_2 are the projections of Π and Λ onto U and V respectively; Φ_{11}, Φ_{12} and Φ_{22} denote

the appropriate restrictions, to U and V, of the projections of Φ onto U and V;

(iii) $H \cdot U \equiv 0$.

Proof. Let S be a linear observed system and let (X, Π, H) be a state representation for S satisfying the additional requirements of the theorem.

Assume that there exist subspaces U and V of X satisfying (i), (ii) and (iii). Let $u \in U$, $u \neq 0$. Then, because of (iii),

$$H\Pi(0,f,0,t) - H\Pi(u,f,0,t) = H \cdot \Lambda(f,0,t) - H\Phi_{11}(u,t) - H\Lambda(f,0,t) = 0$$

for all $f \in F$ and $t \in R$. This shows that u and 0 are indistinguishable and hence S is not completely observable in X.

To show the opposite implication, assume that S is not completely observable in S. Let distinct states x and y be indistinguishable under Π and H. Then $0 \neq x - y \sim 0$, so that $U = \{x : H\Phi(x,t) = 0 \text{ for all } t \in R\}$ is a nontrivial linear subspace of X: evidently U is closed by continuity of H and Φ. Let V be the orthogonal complement of U, so that $X = U \oplus V$. Let $\Pi_1, \Pi_2, \Phi_1, \Phi_2$ and Λ_1, Λ_2 be the projections of Π, Φ and Λ onto U and V respectively. Similarly let Φ_{21} and Φ_{22} be the analogous restrictions of Φ_2. By (i) of Theorem 2.1.1, $\Phi_{21}(u,t) \equiv 0$ for all $u \in U$ and $t \in R$. Therefore Π has the following representation:

$$\Pi_1(x,f,s,t) = \Phi_{11}(u, t-s) + \Phi_{12}(v, t-s) + \Lambda_1(f,s,t),$$
$$\Pi_2(x,f,s,t) = \Phi_{22}(v, t-s) + \Lambda_2(f,s,t)$$

for all $x \in X$ with $x = u + v$, $u \in U$, and $v \in V$, $f \in F$ and s, t in R. By (ii) of Theorem 2.2.1 it follows that $H \cdot U \equiv 0$. The proof of this theorem is completed.

COROLLARY 2.2.3. *Let (X, Π, H) be a state representation for an observed linear system S, where X is a Hilbert space and the zero response Φ of Π is continuous on $X \times R$. Then there exists a unique closed subspace U of X with the following properties:*

(i) *U is invariant under Φ;*

(ii) *S restricted to U is completely unobservable;*

(iii) *U is a maximal subspace of X having* (i) *and* (ii);

(iv) *S is completely observable in the orthogonal complement V of U.*

Proof. Let $U = \{x : H\Phi(x,t) \equiv 0 \text{ for all } t \in R\}$.

By Theorem 2.2.1, U satisfies (i), (ii) and (iii). To prove (iv), assume that S is not completely observable on V. Let $x \neq y$ in X be such that

$$H\Phi(x,t) = H\Phi(y,t) \quad \text{for all } t \in R.$$

Then $H\Phi(x - y, t) \equiv 0$ for all $t \in R$, i.e., $x - y \in U$. Since also $x - y \in V$, we get that $x - y = 0$; this is in contradiction to the assumption that S is not completely observable on V. This completes the proof.

Remarks. The notion of indistinguishable states and its role in the definition of observability is apparently due to M. Arbib; it was communicated to me by L. Kerschberg. The alternate condition for observability given by Theorem 2.2.2 was used as the working definition by both Kalman in [14] and Markus in [17]. The definition adopted in the present paper is more general in that it depends

less on the structure of the state space, and, furthermore, it gives a very constructive way of generating the completely unobservable subspace of the state space (Theorem 2.2.1).

Acknowledgments. The author expresses his warmest gratitude to Professors O. Hajek and S. K. Mitter for their valuable help and encouragement.

REFERENCES

[1] H. A. ANTOSIEWICZ, *Linear control theory*, Arch. Rational Mech. Anal., 12 (1963), pp. 314–324.
[2] N. BHATIA AND O. HAJEK, *Theory of dynamical systems*, to appear.
[3] D. BUSHAW, *Dynamical polysystems and optimization*, Contributions to Differential Equations, 2 (1963), pp. 351–365.
[4] ———, *Dynamical polysystems: A survey*, Proc. U.S.-Japan Seminar on Differential and Functional Equations, Minneapolis, Minn., 1967, W. A. Benjamin, New York, 1967, pp. 13–24.
[5] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1965.
[6] H. O. FATTORINI, *Some remarks on complete controllability*, this Journal, 4 (1966), pp. 686–694.
[7] ———, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
[8] ———, *Boundary control systems*, this Journal, 6 (1968), pp. 349–385.
[9] I. GEL'FAND, *On one parametrical group of operators in a normed space*, C.R. Acad. Sci. USSR, 25 (1939), pp. 713–718.
[10] O. HAJEK, *Theory of processes. I*, Czechoslovak Math. J. (92), 17 (1967), pp. 159–199.
[11] ———, *Linear semi-dynamical systems*, Math. Systems Theor., 2 (1968), pp. 195–202.
[12] H. HALKIN, *Topological aspects of optimal control to dynamical polysystems*, Contributions to Differential Equations, 3 (1964), pp. 377–385.
[13] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, Colloquium Publications, 2nd ed., American Mathematical Society, Providence, 1957.
[14] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.
[15] R. E. KALMAN, Y. C. HO AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1963), pp. 189–213.
[16] L. MARKUS, *Controllability and observability*, Functional Analysis and Optimization, E. R. Cainiello, ed., Academic Press, New York, 1966, pp. 133–143.
[17] L. MARKUS AND B. LEE, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[18] E. ROXIN, *On generalized dynamical systems defined by contingent equations*, J. Differential Equations, 1 (1965), pp. 188–205.
[19] T. G. WINDEKNECHT, *Mathematical systems theory; causality*, Math. Systems Theor., 1 (1967), pp. 279–289.
[20] K. YOSIDA, *Functional Analysis*, Academic Press, New York, 1965.

SIAM J. CONTROL
Vol. 8, No. 3, August 1970

ERRATUM: OPTIMAL CONTROL OF PROCESSES DESCRIBED BY INTEGRAL EQUATIONS. I*

V. R. VINOKUROV

The line immediately following (1.1) was printed incorrectly and should be replaced by: "$\cdots$ where x, f and K are n-dimensional column vectors and u is an r-dimensional $\cdots$."

* This Journal, 7 (1969), pp. 324–336.

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

N-PERSON NONZERO SUM DIFFERENTIAL GAMES WITH LINEAR DYNAMICS*

PRAVIN VARAIYA†

Abstract. There exist equilibrium strategies for nonzero sum differential games with N players if the dynamics are linear and if the cost to each player is convex.

1. Introduction. Consider the following linear differential system:

$$\dot{x}(t) = A(t)x(t) + B_1(t)u_1(t) + \cdots + B_N(t)u_N(t), \qquad t \in [0, T] \text{ a.e.}, \tag{1}$$

with state $x(t)$ in R^n and controls $u_j(t)$ in R^{m_j}, $1 \leqq j \leqq N$. The matrices A, B_j have appropriate dimension and their coefficients are bounded, measurable functions of t. We impose once and for all the following boundary values:

$$x(0) = x_0. \tag{2}$$

The function $u_j(t)$, $0 \leqq t \leqq T$, is the *control* of player j. For a given fixed M, $0 \leqq M \leqq \infty$, player j is allowed to choose any measurable control u_j which satisfies the constraint condition (C_M), where

$$\int_0^T |u_j(t)|^2 \, dt \leqq M \qquad \text{if } M < \infty, \tag{C_M}$$

$$\int_0^T |u_j(t)|^2 \, dt < \infty \qquad \text{if } M = \infty. \tag{C_∞}$$

Here $|u_j|^2$ = square of the Euclidean norm in R^{m_j}. We also write

$$\|u_j\|^2 = \int_0^T |u_j(t)|^2 \, dt$$

so that (C_M) reads $\|u_j\|^2 \leqq M$ and (C_∞) reads $\|u_j\|^2 < \infty$.

Controls which satisfy the constraint condition are called *admissible controls.* The set of admissible controls of player j is denoted $\mathcal{U}_j(M)$.

Suppose for each $i = 1, \cdots, N$ player i chooses an admissible control u_i. This choice determines a unique function $x(t)$, $0 \leqq t \leqq T$, which satisfies (1) and (2). This function is called the *trajectory* corresponding to the controls $u_1, \cdots, u_N$. Also corresponding to this choice of controls, player i incurs a cost $J_i = J_i(u_1, \cdots, u_N)$. We consider cost functions of two types and we denote the cost J_i by J_i^{I} or J_i^{II} according to whether it is of type I or type II.

$$\text{(Type I)} \qquad J_i^{\mathrm{I}}(u_1, \cdots, u_N) = g_i(x(T)) + \int_0^T f_i(t, u_i(t)) \, dt,$$

$$\text{(Type II)} \quad J_i^{\mathrm{II}}(u_1, \cdots, u_N) = g_i(x(T)) + \int_0^T f_i(t, u_i(t)) \, dt + \int_0^T h_i(t, x(t)) \, dt.$$

* Received by the editors September 17, 1969, and in revised form February 27, 1970.

† Electronics Research Laboratory, College of Engineering, University of California, Berkeley, California 94720. This research was supported by the National Aeronautics and Space Administration under Grant NGL-05-003-016 (Sup 6).

In the above $x(t), 0 \leqq t \leqq T$, is the trajectory corresponding to $(u_1, \cdots, u_N)$.

It is the objective of each player to minimize his own cost.

DEFINITION 1. Let $0 \leqq M \leqq \infty$ be fixed. For a fixed k equal to I or II let J_i^k be fixed cost functions. Controls $u_i \in \mathcal{U}_i(M)$ are said to form an *equilibrium strategy* for the game $G = (J_1^k, \cdots, J_N^k; M)$ if for each $i = 1, \cdots, N$,

$$J_i^k(u_1, \cdots, u_N) \leqq J_i^k(u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_N) \tag{3}$$

for all $v_i \in \mathcal{U}_i(M)$.

We assume throughout that the functions g_i, f_i, h_i which constitute the cost functions J_i satisfy the following conditions:

(i) $g_i(x), f_i(t, u_i), h_i(t, x)$ are continuous in all variables, bounded from below, and for each fixed t, they are convex in the remaining variables.

(ii) h_i, f_i are C^2 in x, u_i respectively and $g_i'(x), g_i''(x), h_i'(t, x), h_i''(t, x), f_i'(t, u_i), f_i''(t, u_i)$ are continuous in all variables and, furthermore, there exist positive numbers $\varepsilon_1, \varepsilon_2$ such that for all t, x, u_i,

$$f_i''(t, u_i) \geqq \varepsilon_1 I, \quad \varepsilon_2 I \geqq h_i''(t, x), \quad \varepsilon_2 I \geqq g_i''(x).$$

In (ii) the prime denotes differentiation with respect to x or u_i and I denotes the identity matrix.

We can now state our principal result.

THEOREM. (I) *Let $J_1^{\mathrm{I}}, \cdots, J_N^{\mathrm{I}}$ be cost functions of type* I *and let $0 \leqq M < \infty$. There exists an equilibrium strategy for the game $G = (J_1^{\mathrm{I}}, \cdots, J_N^{\mathrm{I}}; M)$.*

(II) *Let $J_1^{\mathrm{II}}, \cdots, J_N^{\mathrm{II}}$ be cost functions of type* II *and let $0 \leqq M \leqq \infty$. There is a $T_0 > 0$ such that if the duration T in* (1) *satisfies $T < T_0$ then there exists an equilibrium strategy for the game $G = (J_1^{\mathrm{II}}, \cdots, J_N^{\mathrm{II}}; M)$.*

In § 2 we give some preliminary results required for the proof of the theorem in § 3. In § 4 we give some extensions to the case of nonlinear dynamics and compare our results with those reported in the literature.[1]

2. Preliminary results. The following notation is helpful to the exposition. Recall the notation $\mathcal{U}_i(M)$. We denote $\mathcal{U}(M) = \{u = (u_1, \cdots, u_N) | u_i \in \mathcal{U}_i(M), 1 \leqq i \leqq N\}$.

Evidently $\mathcal{U}(M) \subset \mathcal{U}(\infty)$ for all M. Now $\mathcal{U}(\infty)$ is a Hilbert space under the norm

$$\|u\|^2 = \sum_{i=1}^{N} \|u_i\|^2 = \sum_{i=1}^{N} \int_0^T |u_i(t)|^2 \, dt.$$

As such, this norm induces a topology on $\mathcal{U}(\infty)$; this topology will be called the *strong* topology. Also associated with $\mathcal{U}(\infty)$ in a natural manner is the weak topology [2]. For $M < \infty$, by the strong (weak) topology of $\mathcal{U}(M)$ we mean the corresponding relative topology. Finally, a function $\sigma: \mathcal{U}(M) \to R^p$ is said to be *strongly* (*weakly*) continuous if σ is continuous when $\mathcal{U}(M)$ has the strong (weak) topology. Evidently if σ is weakly continuous then it is strongly continuous.

[1] The proof of this theorem was suggested by an account of a theorem of Nikado–Isoda in [1, pp. 30–33]. The author is very grateful to the reviewer for pointing out a serious error in the previous version of the proof.

In the remainder of this section let $0 \leqq M \leqq \infty$ and $k = \text{I}$ or II be fixed. Let $J_1^k, \cdots, J_N^k$ be fixed cost functions of type k. Finally let G denote the game $(J_1^k, \cdots, J_N^k; M)$.

DEFINITION 2. Define $\psi : \mathcal{U}(M) \times \mathcal{U}(M) \to R$ by

$$\psi((u_1, \cdots, u_N), (v_1, \cdots, v_N)) = \sum_{i=1}^{N} J_i^k(u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_N).$$

LEMMA 1. *$u^* = (u_1^*, \cdots, u_N^*)$ is an equilibrium strategy for G if and only if*

$$\psi(u^*, u^*) \leqq \psi(u^*, v) \quad \text{for all } v \in \mathcal{U}(M). \tag{4}$$

Proof. Suppose $u^* = (u_1^*, \cdots, u_N^*)$ is an equilibrium strategy. Let $v = (v_1, \cdots, v_N) \in \mathcal{U}(M)$. Then

$$J_i^k(u_1^*, \cdots, u_N^*) \leqq J_i^k(u_1^*, \cdots, u_{i-1}^*, v_i, u_{i+1}^*, \cdots, u_N^*), \quad 1 \leqq i \leqq N. \tag{5}$$

Adding these inequalities yields (4). Conversely, suppose (4) holds. Substitution of $v = (u_1^*, \cdots, u_{i-1}^*, v_i, u_{i+1}^*, \cdots, u_N^*)$ in (4) gives (5).

PROPOSITION 1. *The function $x : \mathcal{U}(M) \to R^n$ given by $x(u_1, \cdots, u_N) = x(T)$ is weakly continuous. (Here $x(T)$ is the state of (1) corresponding to the controls $(u_1, \cdots, u_N)$.)*

Proof.

$$x(T) = \Phi(T, 0)x_0 + \sum_{i=1}^{N} \int_0^T \Phi(T, t)B_i(t)u_i(t)\, dt,$$

where Φ is the transition matrix function associated with (1). Now the function

$$u_i \to \int_0^T \Phi(T, t)B_i(t)u_i(t)\, dt$$

is a strongly continuous *linear* function of $\mathcal{U}_i(\infty)$ into the *finite*-dimensional space R^n. Hence it is weakly continuous and the assertion follows.

COROLLARY 1. *The functions $\tilde{g}_i : \mathcal{U}(M) \to R$, where $\tilde{g}_i(u) = g_i(x(T))$, are weakly continuous.*

Proof. $\tilde{g}_i = g_i \circ x$ is the composition of continuous functions.

DEFINITION 3. A function $\sigma : \mathcal{U}(M) \to R$ is said to be *weakly* (*strongly*) *lower semicontinuous* if for each α in R the set $\{u | u \in \mathcal{U}(M), \sigma(u) \leqq \alpha\}$ is weakly (strongly) closed.

It is well known that if $\sigma_1, \cdots, \sigma_N$ are weakly (strongly) lower semicontinuous then $\sigma = \sum_{i=1}^{N} \sigma_i$ is also weakly (strongly) lower semicontinuous.

PROPOSITION 2. *Let $\sigma : \mathcal{U}(M) \to R$ be convex and strongly lower semicontinuous. Then σ is also weakly lower semicontinuous.*

Proof. Let $\alpha \in R$. Then the set $\{u | \sigma(u) \leqq \alpha\}$ is strongly closed. But since σ is convex this set is also convex, and then it is enough to remember that a strongly closed convex set is also weakly closed.

PROPOSITION 3. *The functions* $\tilde{f}_i: \mathcal{U}(M) \to R$, *where*

$$\tilde{f}_i(u_i) = \int_0^T f_i(t, u_i(t))\, dt,$$

is convex and strongly lower semicontinuous.

Proof. The convexity follows from the fact that f_i is assumed convex (in u_i) and integration is a linear operation.

Next let $u_{i,j}, j = 1, 2, \cdots$, be a sequence in $\mathcal{U}_i(M)$ and let u_i in $\mathcal{U}_i(M)$ be such that

$$\lim_{j \to \infty} \int_0^T |u_i(t) - u_{i,j}(t)|^2\, dt = 0 \tag{6}$$

and

$$\int_0^T f_i(t, u_{i,j}(t))\, dt \leqq \alpha. \tag{7}$$

Because of (6), taking subsequences if necessary, we can assume that $\lim_j u_{i,j}(t) = u_i(t)$ a.e. Now the function f_i is bounded from below by assumption. By Fatou's lemma we conclude (using (7)) that

$$\begin{aligned} \alpha \geqq \liminf \int_0^T f_i(t, u_{i,j}(t))\, dt &\geqq \int_0^T \liminf f_i(t, u_{i,j}(t))\, dt \\ &= \int_0^T f_i(t, u_i(t))\, dt \end{aligned}$$

so that $\tilde{f}_i$ is strongly lower semicontinuous.

COROLLARY 2. *The functions* $\tilde{f}_i$ *as defined above are weakly lower semicontinuous.*

Proof. The result follows from Propositions 2 and 3.

Recall Definitions 1 and 2.

DEFINITION 4. Let $v \in \mathcal{U}(M)$ be fixed. Let $\psi_v: \mathcal{U}(M) \to R$ be the function $\psi_v(u) = \psi(u, u) - \psi(u, v)$. Also let $U_v = \{u | \psi_v(u) > 0\}$.

LEMMA 2. (I) *Suppose that the cost functions* J_i^k *are of type* I. *Then* U_v *is weakly open.*

(II) *Suppose that the cost functions* J_i^k *are of type* II. *Then there is a* $T_0 > 0$ *such that if* $T < T_0$ *then* U_v *is weakly open and* $\psi_v(u) \to \infty$ *as* $\|u\| \to \infty$.

Proof. (I) Suppose the cost functions are of type I. Then ψ_v is also given by

$$\begin{aligned} \psi_v(u_1, \cdots, u_N) = &\sum_{i=1}^N (\tilde{g}_i(u_1, \cdots, u_N) - \tilde{g}_i(u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_N)) \\ &+ \sum_{i=1}^N f_i(u_i) - \sum_{i=1}^N \tilde{f}_i(v_i). \end{aligned} \tag{8}$$

By Corollary 1, the first sum is weakly continuous; the second term is weakly lower semicontinuous by Corollary 2, whereas the third sum is constant. Hence ψ_v is weakly lower semicontinuous, so that the set $\overline{U}_v = \{u | \psi_v(u) \leqq 0\}$ is weakly closed and its complement U_v is weakly open.

(II) Suppose the cost functions are of type II. Then ψ_v is also given by

$$\begin{aligned}\psi_v(u_1, \cdots, u_N) = &\sum_{i=1}^{N} (\tilde{g}_i(u_1, \cdots, u_N) - \tilde{g}_i(u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_N)) \\ &+ \sum_{i=1}^{N} (\tilde{f}_i(u_i) + \tilde{h}_i(u_1, \cdots, u_N) \\ &\qquad - \tilde{h}_i(u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_N)) \\ &- \sum_{i=1}^{N} \tilde{f}_i(v_i).\end{aligned} \tag{9}$$

Once again the first sum is weakly continuous and the third sum is a constant. By Proposition A.1 of the Appendix, there is a $T_0 > 0$ such that if $T < T_0$ then

(i) the first sum plus the second sum is strongly lower semicontinuous and convex (hence weakly lower semicontinuous) and

(ii) the first sum plus the second sum grows indefinitely as $\|u\|$ grows indefinitely.

Repeating the argument of the first part proves assertion (II).

3. Proof of theorem.

Step 1. If the cost functions are of type II, choose T_0 to satisfy the conditions of Lemma 2. Suppose the theorem is false. Then by Lemma 1, for each $u \in \mathcal{U}(M)$ there is a $v \in \mathcal{U}(M)$ such that $\psi(u, u) - \psi(u, v) > 0$, i.e., $u \in U_v$. Hence $\mathcal{U}(M)$ has the following weakly open cover:

$$\mathcal{U}(M) \subset \bigcup_{v \in \mathcal{U}(M)} U_v. \tag{10}$$

Next we show that there is a finite subset $\{v_1, \cdots, v_p\}$ of $\mathcal{U}(M)$ such that

$$\mathcal{U}(M) \subset \bigcup_{i=1}^{p} U_{v_i}. \tag{11}$$

Case i. Suppose $M < \infty$. Then $\mathcal{U}(M)$ is a convex, strongly bounded and closed set so that it is weakly compact. Then (10) must have a finite subcover (11).

Case ii. Suppose $M = \infty$. Let $v_1 \in \mathcal{U}(M)$. By Lemma 2, $\psi_{v_1}(u) \to \infty$ as $\|u\| \to \infty$. Hence there is $M_1 < \infty$ such that $\psi_{v_1}(u) > 0$ whenever $\|u\| > M_1$. That is, $\{u | \|u\|^2 > M_1\} \subset U_{v_1}$. Now since $\mathcal{U}(M_1)$ is weakly compact there exist $v_2, \cdots, v_p$ such that $\mathcal{U}(M_1) \subset \bigcup_{i=2}^{p} U_{v_i}$. Hence

$$\mathcal{U}(\infty) = \mathcal{U}(M_1) \cup \{u | \|u\|^2 > M_1\} \subset \bigcup_{i=1}^{p} U_{v_i},$$

so that once again (11) holds. Note that the assertion says that for each $u \in \mathcal{U}(M)$ there is $1 \leqq j \leqq p$ such that $\psi_{v_j}(u) > 0$.

Step 2. Let V be the convex hull of $\{v_1, \cdots, v_p\}$, i.e.,

$$V = \left\{ \sum_{i=1}^{p} \lambda_i v_i | \lambda_i \geqq 0, \sum_{i=1}^{p} \lambda_i = 1 \right\}.$$

Define the functions $\gamma_j: V \to R$ by

$$\gamma_j(v) = \max(\psi_{v_j}(v), 0).$$

First of all, since V is finite-dimensional the weak and strong topology coincide. Next, γ_{v_j} is strongly continuous on $\mathcal{U}(M)$, hence it is continuous on V, and so γ_j is continuous on V. Finally, by Step 1, the continuous function

$$\gamma = \sum_{j=1}^{p} \gamma_j$$

satisfies

$$\gamma(v) > 0 \qquad \text{for all } v \in V.$$

Step 3. Define the function $\eta: V \to V$ by

$$\eta(v) = \sum_{j=1}^{p} \frac{\gamma_j(v)}{\gamma(v)} v_j.$$

Then η is continuous and so by the Brouwer fixed-point theorem there is v^* in V such that $\eta(v^*) = v^*$. Suppose $\gamma_j(v^*) > 0, j = 1, \cdots, l$, and $\gamma_j(v^*) = 0, j > l$. Then

$$\gamma(v^*) = \sum_{j=1}^{l} \gamma_j(v^*)$$

and the fixed-point condition becomes

$$v^* = \sum_{j=1}^{l} \frac{\gamma_j(v^*)}{\gamma(v^*)} v_j. \tag{12}$$

Also, $\gamma_j(v^*) > 0$ is equivalent to $\psi(v^*, v^*) > \psi(v^*, v_j)$ so that we get

$$\psi(v^*, v^*) > \sum_{j=1}^{l} \frac{\gamma_j(v^*)}{\gamma(v^*)} \psi(v^*, v_j). \tag{13}$$

Step 4. Finally, for cost functions of type I or type II, $\psi(v^*, v)$ is convex in v so that we obtain

$$\psi(v^*, v^*) = \psi\left(v^*, \sum_{j=1}^{l} \frac{\gamma_j(v^*)}{\gamma(v^*)} v_j\right) \leqq \sum_{j=1}^{l} \frac{\gamma_j(v^*)}{\gamma(v^*)} \psi(v^*, v_j)$$

which contradicts (13), and the theorem is proved.

4. Extensions.

(i) It should be easy to see that all the propositions of this paper are true if (1) is replaced by a linear differential-difference equation. Also in part (I) of the theorem the differentiability conditions on g_i are unnecessary.

(ii) A careful study of the Appendix shows that its results remain valid if (1) is replaced by a nonlinear differential equation of the form

$$\dot{x} = f(t, x, u_1, \cdots, u_N)$$

provided f is C^2. In turn this implies that part (II) of the theorem remains valid.

(iii) Part (II) of this theorem was proved in [3] for the special case where the functions $g_i \equiv 0$, and f_i, h_i are quadratic. A detailed study including stability and

synthesis of the solution for the quadratic case appears in [4] where it is also shown that, in general, part (II) of the theorem is false for arbitrary duration T.

Appendix. Let $v = (v_1, \cdots, v_N) \in \mathcal{U}(M)$ be fixed. Let $g_i, f_i, h_i, 1 \leqq i \leqq N$, satisfy the assumptions (i) and (ii) in § 1, and consider the function $\sigma : \mathcal{U}(M) \to R$ defined by

$$\begin{aligned}\sigma(u_1, \cdots, u_N) = \sum_{i=1}^{N} \Big\{ & \tilde{f}_i(u_i) + \tilde{g}_i(u_1, \cdots, u_N) + \tilde{h}_i(u_1, \cdots, u_N) \\ & - \tilde{g}_i(u_1, \cdots, v_i, \cdots, u_N) - \tilde{h}_i(u_1, \cdots, v_i, \cdots, u_N) \Big\} \\ = \sum_{i=1}^{N} \Big\{ & \int_0^T [f_i(t, u_i(t)) + h_i(t, x(t)) - h_i(t, x_i(t))]\, dt \\ & + g_i(x(t)) - g_i(x_i(t)) \Big\},\end{aligned}$$

where $x(t)$, $0 \leqq t \leqq T$, and $x_i(t)$, $0 \leqq t \leqq T$, are the trajectories of (1) corresponding to the controls $(u_1, \cdots, u_N)$ and $(u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_N)$ respectively.

PROPOPOSITION A.1. *There exists $T_0 > 0$ such that for all $0 \leqq T < T_0$,*

(i) *σ is strongly lower semicontinuous,*

(ii) *σ is convex,*

(iii) *$\sigma(u) \to \infty$ as $\|u\| \to \infty$.*

Proof. We already know that $\tilde{f}_i$ and $\tilde{g}_i$ are strongly lower semicontinuous so that it is enough to show that $\tilde{h}_i$ is strongly lower semicontinuous. Now let $u^1, u^2, \cdots$ be a sequence of controls in $\mathcal{U}(M)$ converging strongly to u in $\mathcal{U}(M)$, and let $x^1, x^2, \cdots$ and x be the corresponding trajectories. By well-known arguments we can show that

$$\lim_{j \to \infty} \sup_{0 \leqq t \leqq T} |x(t) - x^j(t)| = 0,$$

i.e., x^j converges uniformly to x. It follows by Fatou's lemma that

$$\lim_{j \to \infty} \int_0^T h_i(t, x^j(t))\, dt \geqq \int_0^T h_i(t, x(t))\, dt.$$

The remaining two assertions follow if we prove there are $T_0 > 0$, $\varepsilon_0 > 0$ such that for all $0 \leqq T < T_0$ and for all u and w in $\mathcal{U}(\infty)$,

$$\frac{\partial^2}{\partial \xi^2} \sigma(u + \xi w)\Big|_{\xi = 0} \geqq \varepsilon_0 \|w\|^2. \tag{A.1}$$

To this end let u, w in $\mathcal{U}(\infty)$ be fixed. Let x_ξ and $x_{i,\xi}$ be the trajectories of (1) corresponding to controls $(u + \xi w)$ and $(u_1 + \xi w_1, \cdots, u_{i-1} + \xi w_{i-1}, v_i, u_{i+1} + \xi w_{i+1}, \cdots, u_N + \xi w_N)$ respectively. Then

$$\begin{aligned}\sigma(u + \xi w) = \sum_{i=1}^{N} \Big\{ & \int_0^T [f_i(t, u_i(t) + \xi w_i(t)) + h_i(t, x_\xi(t)) - h_i(t, x_{i,\xi}(t))]\, dt \\ & + g_i(x_\xi(t)) - g_i(x_{i,\xi}(t)) \Big\},\end{aligned}$$

where

$$x_\xi(t) = x(t) + \xi \sum_j z_j(t); \qquad x_{i,\xi}(t) = x_i(t) + \xi \sum_{j \neq i} z_j(t),$$

$x(t), x_i(t)$ are trajectories corresponding to u and $(u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_N)$ respectively and

$$z_j(t) = \int_0^t \Phi(t, \tau) B_j(\tau) w_j(\tau)\, d\tau. \tag{A.2}$$

It follows that

$$\begin{aligned}\frac{\partial^2 \sigma}{\partial \xi^2}(u + \xi w)\Big|_{\xi=0} = \sum_{i=1}^N \Bigg\{ \int_0^T \Bigg[&\langle w_i(t), f_i''(t, u_i(t)) w_i(t) \rangle \\ &+ \left\langle \sum_j z_j(t), h_i''(t, x(t)) \sum_j z_j(t) \right\rangle \\ &- \left\langle \sum_{j \neq i} z_j(t), h_i''(t, x_i(t)) \sum_{j \neq i} z_j(t) \right\rangle \Bigg] dt \\ &+ \left\langle \sum_j z_j(T), g_i''(x(T)) \sum_j z_j(T) \right\rangle \\ &- \left\langle \sum_{j \neq i} z_j(T), g_i''(x_i(T)) \sum_{j \neq i} z_j(T) \right\rangle \Bigg\}.\end{aligned}$$

By assumptions (i) and (ii) in § 1 we get the estimate

$$\frac{\partial^2 \sigma}{\partial \xi^2}(u + \xi w)\Big|_{\xi=0} \geqq \sum_{i=1}^N \left\{ \int_0^T \left[\varepsilon_1 |w_i(t)|^2 - \varepsilon_2 \Big| \sum_{j \neq i} z_j(t) \Big|^2 \right] dt - \varepsilon_2 \Big| \sum_{j \neq i} z_j(T) \Big|^2 \right\}. \tag{A.3}$$

From (A.2) and the assumptions that the coefficients of the matrices in (A.1) are bounded it is easy to see that there is a constant m (depending only on the matrices $A(t)$, $B_j(t)$) such that

$$|z_j(t)|^2 \leqq mt \|w_j\|^2.$$

Hence

$$\Big| \sum_{j \neq i} z_j(t) \Big|^2 \leqq mNt \left(\sum_j \|w_j\|^2 \right).$$

Combining with (A.3) we obtain

$$\begin{aligned}\frac{\partial^2 \sigma}{\partial \xi^2}(u + \xi w)\Big|_{\xi=0} &\geqq \varepsilon_1 \|w\|^2 + \sum_{i=1}^N - \tfrac{1}{2}\varepsilon_2 mNT^2 \left(\sum_j \|w_j\|^2 \right) - \varepsilon_2 mNT \left(\sum_j \|w_j\|^2 \right) \\ &= (\varepsilon_1 - \tfrac{1}{2}\varepsilon_2 mN^2 T^2 - \varepsilon_2 mN^2 T) \|w\|^2. \end{aligned} \tag{A.4}$$

Obviously one can choose $\varepsilon_0 > 0$, $T_0 > 0$ such that (A.1) is true.

REFERENCES

[1] E. BURGER, *Introduction to the Theory of Games*, Prentice-Hall, Englewood Cliffs, N.J., 1963.
[2] N. DUNFORD AND J. SCHWARTZ, *Linear Operators. Part I: General Theory*, Interscience, New York, 1964.
[3] A. FRIEDMAN, *Linear-quadratic differential games with non-zero sum and with N players*, to appear.
[4] D. L. LUKES AND D. L. RUSSELL, *A global theory for linear quadratic differential games*, MRC Tech. Summary Rep. 915, Mathematics Research Center, University of Wisconsin, Madison, 1968.

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

NONLINEAR CONTROLLABILITY VIA LIE THEORY*

G. W. HAYNES† AND H. HERMES‡

1. Introduction. The form of the control system studied throughout most of this paper is

$$\dot{x}(t) = B(x(t))u(t), \qquad \dot{x} = dx/dt, \tag{1}$$

where x is an n-vector and $B(x)$ an $n \times r$ matrix with columns denoted $b^1(x), \cdots, b^r(x)$. We shall assume that the components of $B(\cdot)$ are C^∞ (infinitely differentiable) functions, although this condition could often be relaxed. The control vector u will always be assumed Lebesgue measurable; of particular interest will be the case where its values lie in a bounded set of Euclidean r-dimensional space, R^r.

We begin by interpreting the work of Chow [1] and Hermann [2], [3], [4] to the system (1). Following Chow, we shall say the *system* (1) *has rank r at x* if B has maximum rank r in every neighborhood of x. A point x is *regular* for (1), or for B, if when system (1) has rank r at x, rank $B(x) = r$. Since, in our formulation coordinates are assumed, we may define the *Jacobi bracket* of two C^∞, n-vector-valued functions a, b as

$$[a, b](x) = a_x(x)b(x) - b_x(x)a(x),$$

where $a_x(x)$ denotes the $n \times n$ matrix of partial derivatives $(\partial a_i(x)/\partial x_j)$, $a_i(x)$ being the ith component of $a(x)$.

Define $D^0(B)$ to be the set $b^1, \cdots, b^r$; $D^1(B)$ to be the set $b^1, \cdots, b^r$ together with all elements of the form $[b^i, b^j]$, $i, j = 1, \cdots, r$; $D^2(B) = D^1(D^1(B))$ etc. We define the *derived set* to be $D(B) = \bigcup_{j \geq 0} D^j(B)$.

Remark. Although our notation is similar to that used by Hermann [2] our meaning is different. Indeed, Hermann considers the set of all C^∞ maps of R^n to R^n, which we denote $C^\infty(R^n, R^n)$, as a module with ring of multipliers $C^\infty(R^n, R^1)$. Consider the columns of B as spanning a subspace $\mathscr{B}$, in this (infinite-dimensional) module. Hermann then defines $D^0(\mathscr{B}) = \mathscr{B}$; $D^1(\mathscr{B}) = \mathscr{B} + [\mathscr{B}, \mathscr{B}]$, i.e., the sum of linear combinations of elements of $\mathscr{B}$ and their Jacobi brackets with coefficients in $C^\infty(R^n, R^1)$; $D^2(\mathscr{B}) = D^1(D^1(\mathscr{B}))$, etc.; and finally the derived system $D(\mathscr{B}) = \bigcup_{j \geq 0} D^j(\mathscr{B})$. Thus here, $D(\mathscr{B})$ is a subspace of the module $C^\infty(R^n, R^n)$; on the other hand, our derived set $D(B)$ is merely a collection of elements of $C^\infty(R^n, R^n)$.

We shall use the notation $D(B)_x$ to denote the elements of $D(B)$ evaluated at x, and we view $D(B)_x$ as a collection of vectors in R^n. Let $\dim D(B)_x$ denote the dimension of the subspace of R^n spanned by $D(B)_x$. Suppose that in every neighborhood of x^0, $B(x)$ has maximal rank r (i.e., system (1) has rank r at x^0) and the maximal value of $\dim D(B)_x$ is s. Following Chow [1] we call s the *rank of*

* Received by the editors July 15, 1969, and in final revised form March 15, 1970. This research was supported by the National Aeronautics and Space Administration under Contract NAS2-4898 at the Ames Research Center, Ames, Iowa.

† Martin Marietta Corporation, Denver, Colorado 80201.

‡ Department of Mathematics, University of Colorado, Boulder, Colorado 80302.

the completion of (1) *at* x^0. Clearly $r \leqq s \leqq n$; the integer $s - r$ is called the *index* of the system (1) at x^0. Assume there exist elements $b^{r+1}, \cdots, b^s$ of $D(B)$ such that if $\tilde{B}$ is the matrix with columns $b^1, \cdots, b^r, b^{r+1}, \cdots, b^s$, then rank $\tilde{B}(x) = s$ in every neighborhood of x^0. The system

$$\dot{x} = \tilde{B}(x)\tilde{u}, \tag{2}$$

where $\tilde{u}$ is now an s-dimensional control vector, is called a *completed system associated with* (1).

Note that we do not associate a unique completed system with (1). In fact, x^0 may be a regular point for some completed system and not for another; or we may have the case where no completed system has x^0 as a regular point. We illustrate these possibilities in the following example, suggested by C. Lobry.

Example 1. Consider $B(x)$ to be the 3×2 matrix with columns

$$b^1(x) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad b^2(x) = \begin{pmatrix} 0 \\ 1 \\ x_1^2 \end{pmatrix}.$$

Then

$$b^3(x) = [b^2, b^1](x) = \begin{pmatrix} 0 \\ 0 \\ 2x_1 \end{pmatrix},$$

while if $\tilde{B}$ has columns b^1, b^2, b^3, we see that $\tilde{B}$ has rank three in every neighborhood of 0 but rank $\tilde{B}(0) = 2$; hence 0 is not a regular point. However

$$b^4(x) = [[b^2, b^1], b^1](x) = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix};$$

hence if we choose $\tilde{B}(x)$ to have columns b^1, b^2, b^4, then rank $\tilde{B}(0) = 3$. For either choice of $\tilde{B}$, the system $\dot{x} = \tilde{B}(x)\tilde{u}$ is a completed system of rank three at the origin associated with $\dot{x} = B(x)u$; in the second case the origin is a regular point.

We may obtain an example where no completed system associated with $\dot{x} = B(x)u$ has the origin as a regular point as follows. Let $B(x)$ be the 3×2 matrix with columns

$$b^1(x) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

while

$$b^2(x) = \begin{cases} \begin{pmatrix} 0 \\ 1 \\ \exp(-1/x_1^2) \end{pmatrix} & \text{if } x_1 \neq 0, \\ \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} & \text{if } x_1 = 0. \end{cases}$$

One may readily verify that the rank of the completion of our system is three, at the origin, but that any element of $D(B)$ has third component zero at the origin.

When interpreted in this setting, Chow's results give sufficient conditions that the set of points attainable by solutions of (2), starting from initial data $x(0) = x^0$, form an s-dimensional manifold. Let the vectors e^i, $i = 1, 2, \cdots, r$, form a basis for R^r. It also follows that all points on this manifold can be attained by solutions of (1), starting at x^0, and having controls with values at time t in the set $\{\pm e^1, \cdots, \pm e^r\}$. This generalizes recent work of Kuchera [5].

Our main purpose will be to study the uniform approximation of trajectories of (2) by trajectories of (1). If, in each case, the controls are restricted, say to have unit, or less, length, one may approximate the orbit (point set in R^n) of any solution of (2) by the orbit of a solution of (1). However, as will be seen, to approximate trajectories of solutions of (2) corresponding to unit controls, by solutions of (1), the magnitude of the control values in (1) needed to do this increases with the index of the system (1).

Another result which arises naturally is that of local–local controllability (see Definition 1), i.e., the ability to reach all points in some neighborhood of an initial point x^0 by solutions of (1) without leaving some other preassigned neighborhood. This is a stronger (in some sense) property than complete controllability and clearly of practical importance. Sufficient conditions for (1) to have this property are given.

2. Interpretation of Chow's results to (1). Let B^T denote the transpose of B. With B one may associate the system of partial differential equations $B^T(x)\partial f/\partial x = 0$; the ith equation has the form $\sum_{j=1}^n b_j^i(x)\partial f/\partial x_j = 0$ while the ordinary vector differential equation $\dot{x} = b^i(x)$ is called its associated *characteristic equation.* One should note that the ith characteristic equation of $B^T(x)\partial f/\partial x = 0$ may be obtained from the control system (1) by placing the ith component of the control vector u equal to 1 and all other components equal to zero.

The results of Chow [1] pertain to points attainable by "piecing together" characteristic solutions; i.e., if $\varphi^i(\cdot, y)$ denotes the solution of the ith characteristic equation, satisfying data $\varphi(0, y) = y$, then a point z attained from y by two characteristic solutions pieced together has the form $z = \varphi^j(t_2, \varphi^i(t_1, y))$. It is of fundamental importance to note that the Chow formulation allows the characteristic solution to be considered with *decreasing*, as well as increasing time. Thus if φ is a piecing together of characteristic solutions such that in some time interval I_i, φ is a solution of the ith characteristic equation, we only know that

$\dot{\varphi}(t) = \pm b^i(\varphi(t))$, $t \in I_i$. This presents no problem in the system (1), since the minus sign may be obtained by merely taking a control with -1 as its ith component and all others zero.

For $1 \leqq i \leqq r$, let $e^i \in R^r$ (real r-dimensional space) have a one in its ith component and all other components zero. Define

$$V = \{\pm e^1, \cdots, \pm e^r\} \subset E^r$$

and

$$U = \{u \text{ measurable}; u(\tau) \in V, \tau \geqq 0\}.$$

Then a solution φ of (1) corresponding to a control $u \in U$ is a piecing together of characteristic solutions in the sense of Chow. With the above in mind, we may combine satz B and C (Chow [1]) (see also Hermann [2]) as follows.

THEOREM 1. *Let system* (1) *have rank r, and its completion have rank s, at* x^0. *Assume* (2) *is a completed system associated with* (1) *such that* x^0 *is a regular point for both B and* $\tilde{B}$. *Then there exists an s-dimensional manifold* M^s *through* x^0 *such that all points on this manifold are attainable by solutions of* (1) *with initial data* $x(0) = x^0$ *and control* $u \in U$. *Furthermore, given a sufficiently small neighborhood of* x^0, *the only points attainable by such solutions of* (1), *which remain in the neighborhood, are points of* M^s.

We next give two examples, the first (following Chow) to illustrate the necessity that x^0 be regular for both B and $\tilde{B}$, the second example to illustrate the "local nature" stressed in the last sentence of the theorem.

Example 2. We consider the three-dimensional system $\dot{x} = B(x)u$, where

$$B(x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & x_1 x_3 \end{pmatrix}.$$

Here the point $x^0 = (0, 0, 0)$ is regular for B; i.e., rank $B(x^0) = 2$. Computing, we have $[b^1, b^2](x) = \begin{pmatrix} 0 \\ 0 \\ x_3 \end{pmatrix} = b^3(x)$. If we let $\tilde{B}$ have columns b^1, b^2, b^3, then the rank of the system $\dot{x} = \tilde{B}(x)\tilde{u}$ is three, at the origin; however, the origin is not a regular point since rank $\tilde{B}(0) = 2$. It is easy to see that all points attainable by solutions of $\dot{x} = B(x)u$ (or $\dot{x} = \tilde{B}(x)\tilde{u}$), initiating from the origin, lie in the plane $x_3 = 0$. Thus the manifold of attainability has dimension two, and we can conclude that 0 is not a regular point for any completed system associated with $\dot{x} = B(x)u$, or equivalently, every element of $D(B)_0$ has third component 0 at the origin.

Throughout, for $y \in R^r$, we shall use the notation $|y|$ to denote the Euclidean length of y.

Example 3. The purpose of this example is to illustrate that if one does not restrict solutions to lie in a small neighborhood of x^0, the last statement of Theorem 1 need no longer be valid. We consider, again, a three-dimensional system

$\dot{x} = B(x)u$ where B is a 3×2 matrix with

$$b_1^1(x) = \begin{cases} \exp[-1/(|x|^2 - 1)] & \text{if } |x| > 1, \\ 0 & \text{if } |x| \leqq 1, \end{cases} \qquad b_1^2(x) = 0,$$

$$b_2^1(x) = 0, \qquad b_2^2(x) = 1,$$

$$b_3^1(x) = 1, \qquad b_3^2(x) = 0.$$

Then, for any x, rank $B(x) = 2$. If $|x| < 1$, $[b^1(x), b^2(x)] = 0$. If $|x| > 1$, $[b^1(x), b^2(x)] = ([2x_2/(|x|^2 - 1)^2] \exp[-1/(|x^2| - 1)], 0, 0)$. Let $\tilde{B}$ have columns $b^1, b^2, [b^1, b^2]$. If $|x| > 1$ and $x_2 \neq 0$, $\tilde{B}$ has rank 3 at x and x is regular for both B and $\tilde{B}$.

Now consider $x_2^0 \neq 0$ and $|x^0| < 1$. Then $\tilde{B} = B$ and x^0 is regular for B and $\tilde{B}$. In this case the integral manifold, M^2, of Theorem 1, is the intersection of the unit ball (origin centered) with the plane $x_1 = x_1^0$. If we choose a neighborhood of x^0, contained in the unit ball, the only points attainable by trajectories of the original system which remain in this neighborhood, are points on this plane. However, without this restriction, all points in some neighborhood of x^0 may be attained by trajectories of the system with controls $u \in U$. This occurs even though the unit ball is foliated by leaves $\{(x_1, x_2, x_3): x_1 = \text{const.}\}$ since we may exit the ball on the leaf $x_1 = x_1^0$, then move on an arbitrary path in the half-space $x_2 > 0$ and reenter the ball on a different leaf to reach points near x^0.

Motivated mainly by the last example, we introduce another concept of controllability for a general control system:

$$\dot{x}(t) = f(x(t), u(t)). \tag{3}$$

DEFINITION 1. The system (3) is *locally-locally controllable at* x^0 if given any $\varepsilon > 0$ there exists a $\delta > 0$ such that all points of the δ-neighborhood of x^0 can be attained by trajectories which do not leave the ε-neighborhood. (Clearly $\delta \leqq \varepsilon$.)

DEFINITION 2. The system (3) is *globally-locally controllable at* x^0 if all points in some neighborhood of x^0 can be attained by trajectories through x^0.

In terms of these definitions, we note that if, in Example 3, $|x^0| < 1$, system (1) is not locally-locally controllable at x^0. However, with $|x^0| < 1$ and $x_2^0 \neq 0$, the system is globally-locally controllable at x^0.

It is interesting to compare these notions with that of complete controllability; i.e., any two points can be joined by a solution. For example, the "Bushaw problem"

$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = -x_1 + u$$

is completely controllable; yet it is easily noted that if $x^0 \neq 0$, then the system is not locally-locally controllable at x^0. On the other hand, complete controllability certainly implies global-local controllability.

Example 2 illustrated that one of the possibilities which may occur when the completion of (1) has rank n at x^0 but x^0 is not regular for any completed system associated with (1) is that all solutions of (1) remain on an $(n-1)$-manifold. The next example, also suggested by C. Lobry, shows that another phenomenon, i.e., local-local controllability, may occur in this case.

Example 4. Let $B(x)$ be the 3×2 matrix with columns

$$b^1(x) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

$$b^2(x) = \begin{pmatrix} 0 \\ 1 \\ \exp(-1/x_1^2) \end{pmatrix} \quad \text{if } x_1 \neq 0, \qquad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{if } x_1 = 0.$$

Then

$$b^3(x) = [b^2, b^1](x) = \begin{pmatrix} 0 \\ 0 \\ (-2/x_1^3)\exp(-1/x_1^2) \end{pmatrix} \quad \text{if } x_1 \neq 0,$$

$$b^3(x) = 0 \quad \text{if } x_1 = 0.$$

As commented in Example 1, $\dim D(B)_0 = 2$; however if $\tilde{B}$ has columns b^1, b^2, b^3, then the completed system $\dot{x} = \tilde{B}(x)\tilde{u}$ has rank 3 at the origin since clearly $\tilde{B}$ has rank 3 in every neighborhood of the origin. However, the origin is not a regular point for any completed system associated with $\dot{x} = B(x)u$. It is quite clear that the completed system $\dot{x} = \tilde{B}(x)\tilde{u}$ is locally-locally controllable at the origin. We shall next show that this is also the case for the original system $\dot{x} = B(x)u$.

Let $y = (y_1, y_2, y_3)$ be any point in some neighborhood of the origin. The case $y_3 = 0$ is trivial, so assume $y_3 \neq 0$. We will show how to attain y by a solution of $\dot{x} = B(x)u$, $x(0) = 0$. First choose $u_1 = 1$, $u_2 = 0$ so that at some time $t_1 > 0$, $x(t_1) \neq 0$. Let $\alpha = \exp(-1/x_1(t_1)^2)$; now choose $u_1(t) = 0$, $u_2(t) = \operatorname{sgn} y_3$ for $t \in (t_1, t_2]$ where t_2 is such that $\alpha(t_2 - t_1) = y_3$. We then get $x_3(t_2) = y_3$, and say $x_2(t_2) = \beta$ while $x_1(t_2) = x_1(t_1)$. Let $t_3 = t_2 + t_1$, and now choose $u_2(t) = 0$ and $u_1(t) = -1$ for $t \in (t_2, t_3]$. This leaves us with $x_1(t_3) = 0$, $x_2(t_3) = \beta$, $x_3(t_3) = y_3$. Now choose $u_1(t) = 0$ and $u_2(t) = \operatorname{sgn}(y_2 - \beta)$ if $(y_2 - \beta) \neq 0$ for $t \in (t_3, t_4]$, where $(t_4 - t_3) = (y_2 - \beta)$ if $(y_2 - \beta) \neq 0$ and otherwise $t_4 = t_3$. We now have $x_1(t_4) = 0$, $x_2(t_4) = y_2$, $x_3(t_4) = y_3$; switch u_2 to zero and we can use u_1 alone to attain the desired final point. It is also easy to see that given any $\varepsilon > 0$, we can reach all points in an $\alpha\varepsilon$-neighborhood of the origin without leaving the ε-neighborhood, hence our system *is* locally-locally controllable at the origin.

Suppose that system (1) has rank r at x^0 while (2) is an associated completed system of rank s at x^0 *and* x^0 is regular for both B and $\tilde{B}$. Then it is a consequence of Theorem 1 that a necessary condition for (1) to be locally-locally controllable at x^0 is that $s = n$. To show that this is also a sufficient condition requires some further analysis and will be a consequence of the next section.

3. Uniform approximation of trajectories of (2) by trajectories of (1). If system (1) has rank r and the completed system (2) has rank s at x^0 and if x^0 is regular for both B and $\tilde{B}$, the tangent space to the manifold M^s of points attainable from x^0 is spanned by $b^1(x), \cdots, b^s(x)$ for x in a neighborhood of x^0. Thus if ψ is a smooth function satisfying $\psi(0) = x^0$, $\dot{\psi}(t) \in \operatorname{span}\{b^1(\psi(t)), \cdots, b^s(\psi(t))\}$ for

$t \geqq 0$, then ψ describes a curve on M^s. Let

$$\tilde{U} = \{\tilde{u} \text{ measurable} : \tilde{u}(t) \in R^s, |\tilde{u}(t)| \leqq 1, t \geqq 0\}.$$

Then clearly a solution of (2), with control $\tilde{u} \in \tilde{U}$ and initial data $x(0) = x^0$, describes a curve on M^s. The goal of this section will be to show that such a solution may be uniformly approximated (on a compact time interval) by a solution of (1); however the magnitude of the control needed in (1) to do this may be large. *Throughout the remainder of this section we will assume* (1) *has rank r at* x^0, *while system* (2) *is an associated completed system of rank s at* x^0 *and* x^0 *is regular for both B and* $\tilde{B}$.

Theorem 1 shows that all points on M^s are attainable by solutions of (1), even with controls $u \in U$. It is natural, then, to try to approximate a solution ψ of (2) on a compact time interval $[0, T]$ by finding a solution φ of (1) which agrees with ψ at many points; i.e., say $\psi(kT/m) = \varphi(kT/m)$ for m a large integer and $k = 0, 1, \cdots, m$.

The major difficulty that occurs in doing this is to show that the time it takes to reach an arbitrary point on M^s near x^0 by a solution of (1) tends to zero as the point tends to x^0. This will be the purpose of the next two lemmas, which will relate the time it takes to reach a point on M^s from x^0 by a solution of (2), with control $\tilde{u} \in \tilde{U}$, to the time needed to reach the point by a solution of (1) with control $u \in U$.

LEMMA 1. *Let* $\xi^i(\cdot)$ *denote a solution of* $\dot{x} = b^i(x)$, $x(0) = x^0$, *where* b^i *is obtained from* $b^1, \cdots, b^r$ *by p bracket operations. Then there exists a control* $u \in U$ *such that the corresponding solution* $\varphi(\cdot, u)$ *of* (1) *satisfies*:

(i) $\varphi(4^p\tau, u) - x^0 = \tau^{p+1} b^i(x^0) + o(\tau^{p+1})$ *as* $\tau \to 0$;

(ii) $\varphi(4^p t^{1/(1+p)}, u) - \xi^i(t) = o(t)$ *as* $t \to 0$.

Proof. If b^i is one of the set $b^1, \cdots, b^r$, then $p = 0$ and (i) merely states that there exists a $u \in U$ (in particular in this case $u = e^i$) such that $\varphi(\tau, e^i) - x^0 = \tau b^i(x^0) + o(\tau)$ as $\tau \to 0$, which is evident. Also (ii) merely reduces to $\varphi(\cdot, e^i) = \xi^i(\cdot)$. The proof proceeds by induction. However, the general step is similar to the case $p = 1$. Thus for clarity of presentation and simplicity of notation we will present only this argument.

Suppose b^i is obtained by the use of one bracket operation, i.e., $b^i = [b^j, b^k]$. Let $T^j(t)y$ denote the solution, at time t, of $\dot{x} = B(x)e^j = b^j(x)$, $x(0) = y$, where $1 \leqq j \leqq r$. From the interpretation of the bracket operation (see, for example, [6, § 1.4]) $T^k(-t)T^j(-t)T^k(t)T^j(t)x^0 - x^0 = t^2[b^j(x^0), b^k(x^0)] + o(t^2)$ as $t \to 0$. Let u be defined by

$$u(\tau) = \begin{cases} e^j & \text{if } \tau \in [0, t], \\ e^k & \text{if } \tau \in (t, 2t], \\ -e^j & \text{if } \tau \in (2t, 3t], \\ -e^k & \text{if } \tau \in (3t, 4t]. \end{cases}$$

Then $\varphi(4t,u) = T^k(-t)T^j(-t)T^k(t)T^j(t)x^0$. Let $\xi(\cdot)$ denote the solution of $\dot{x} = b^i(x)$, $x(0) = x^0$; then

$$\xi^i(t) - x^0 = t[b^j(x^0), b^k(x^0)] + o(t).$$

Since

$$\varphi(4t,u) - x^0 = t^2[b^j(x^0), b^k(x^0)] + o(t^2) \quad \text{as } t \to 0,$$

(i) and (ii) follow easily for the case $p = 1$. The results, for arbitrary p, follow in this manner by induction.

LEMMA 2. *Let $\psi(\cdot)$ be a solution of* (2) *corresponding to a control $\tilde{u} \in \tilde{U}$ and initial data $x(0) = x^0$. Then there exists a solution $\varphi(\cdot, u)$ of* (1) *with control $u \in U$ and a $\tau = \tau(t_0)$ such that $\varphi(\tau, u) = \psi(t_0)$ and $\tau(t_0) \to 0$ as $t_0 \to 0$.*

Proof. By Theorem 1, we are assured of a value τ and control $u \in U$ such that $\varphi(\tau, u) = \psi(t_0)$. We concern ourselves, therefore, with showing that $\tau(t_0) \to 0$ as $t_0 \to 0$.

Let y be a regular point for B and $\tilde{B}$ and $T^i(t)y$ denote a solution, at time t, of $\dot{x} = b^i(x)$, $x(0) = y$ for $1 \leqq i \leqq s$. Let p_i denote the minimum number of bracket operations needed to obtain b^i from $b^1, \cdots, b^r$. Then, from Lemma 1, we may find a control $u^i \in U$ such that the corresponding solution $\varphi(\cdot, u^i)$ of (1) through initial data $x(0) = y$ satisfies $\varphi(4^{p_i}t^{1/(1+p_i)}, u^i) - T^i(t)y = o(t)$ as $t \to 0$, or, from Lemma 1 (i), $\varphi(4^{p_i}t^{1/(1+p_i)}, u^i) - y = tb^i(y) + o(t)$ as $t \to 0$. To simplify notation, denote $\varphi(4^{p_i}t^{1/(1+p_i)}, u^i)$ by $S^i(t)y$.

Consider the map $h: R^s \to M^s$ defined by $h/t_1, \cdots, t_s) = S^s(t_s)\cdots S^1(t_1)x^0$. (Note that if, for $1 \leqq k \leqq s-1$, we let $y^k = S^k(t_k)\cdots S^1(t_1)x^0$, then if $t_k, \cdots, t_1$ are sufficiently small, y^k is a regular point for B and $\tilde{B}$.) The Jacobian (differential) $(Dh)(0)$ is just the $s \times s$ matrix with columns $b^1(x^0), \cdots, b^s(x^0)$ and hence is nonsingular. Now $h(0, \cdots, 0) = x^0$ and the implicit function theorem applies to show that h maps a neighborhood of zero onto a neighborhood of x^0. Specifically, for t_0 such that $\psi(t_0)$ is in this neighborhood there exist times $t_1, \cdots, t_s$, each depending on t_0, such that $S^s(t_s)\cdots S^1(t_1)x^0 = \psi(t_0)$, and, for $1 \leqq i \leqq s$, each $t_i \to 0$ as $t_0 \to 0$. Now we may "piece together" a control $u \in U$ in the obvious way such that its corresponding solution φ, through x^0, satisfies

$$\varphi\left(\sum_{i=1}^{s} 4^{p_i}t_i^{1/(1+p_i)}, u\right) = S^s(t_s)\cdots S^1(t_1)x^0 = \psi(t_0).$$

Let

$$\tau = \tau(t_0) = \sum_{i=1}^{s} 4^{p_i}t_i^{1/(1+p_i)}.$$

Then $\varphi(\tau(t_0), u) = \psi(t_0)$ and $\tau(t_0) \to 0$ as $t_0 \to 0$ as required.

THEOREM 2 (Uniform approximation of a solution of (2) by a solution of (1)). *Let ψ be any solution of* (2) *with initial data $x(0) = x^0$ and control $\tilde{u} \in \tilde{U}$, defined on an interval $[0, T]$ such that for $t \in [0, T]$, $\psi(t)$ is regular for both B and $\tilde{B}$. Then given any $\varepsilon > 0$ there exists a solution φ of* (1) *corresponding to initial data $x(0) = x^0$ and some bounded measurable control u, such that* $\max\{|\varphi(t) - \psi(t)| : 0 \leqq t \leqq T\} < \varepsilon$.

Proof. We first note that if $\varphi(\cdot, u)$ denotes a solution of (1) with control u, then for any real α, $\varphi(\alpha t, u) = \varphi(t, \alpha u)$ for all t.

Let $N(\varepsilon, \psi)$ denote a compact $\varepsilon > 0$ neighborhood of $\{\psi(t) : 0 \leqq t \leqq T\}$ and let $\beta = \max\{|b^i(x)| : x \in N(\varepsilon, \psi), 1 \leqq i \leqq s\}$. Note that with $|u| \leqq 1$, if $\varphi(\cdot, u)$ is a

solution of (1), then $|\varphi(t,u) - \psi(t)| \leqq \varepsilon$ on $[0,\mu]$ if $2\mu\beta < \varepsilon$. (The 2 is needed since $\dot{\psi}$ and $\dot{\varphi}$ may have opposite direction.)

For any integer k, consider $\psi(T/k)$. By Lemma 2 there is a control $u \in U$ and a τ_1 such that the corresponding solution φ of (1) satisfies $\varphi(\tau_1, u) = \psi(T/k)$ and we may choose k large enough so that $\tau_1 < \mu$ (i.e., here we need $\tau_1 \to 0$ as $T/k \to 0$). Then there exists an $\alpha > 0$ such that $\alpha T/k = t_1$; hence $\varphi(t_1, u) = \psi(T/k) = \varphi(\alpha T/k, u) = \varphi(T/k, \alpha u)$. Since $t_1 < \mu, |\varphi(t, \alpha u) - \psi(t)| < \varepsilon$ for $0 \leqq t \leqq T/k$.

Now the solutions φ, ψ agree at T/k; we may repeat the procedure with x^0 replaced by $\psi(T/k)$ and obtain the result for $[0, 2T/k]$, etc.

The approximation procedure is probably best illustrated by Example 5, given after the next corollary.

COROLLARY. *If each point $x \in R^n$ is regular for both B and $\tilde{B}$ and* rank $\tilde{B}(x) = n$, *then the system* (1) *is completely controllable and locally-locally controllable at every point. Furthermore, if ψ is any continuously differentiable map, $\psi : [0,1] \to R^n$ and $\varepsilon > 0$, there exists a bounded measurable control u such that the corresponding solution $\varphi(\cdot, u)$ of* (1) *satisfies* $\max\{|\psi(t) - \varphi(t)| : 0 \leqq t \leqq 1\} < \varepsilon$.

Proof. Clearly it suffices to prove the last statement. Let ψ be a continuously differentiable map $\psi : [0,1] \to R^n$. Since rank $\tilde{B}(x) = n$ for all x, define $v(t) = \tilde{B}^{-1}(\psi(t))\dot{\psi}(t)$. Then ψ satisfies $\dot{\psi}(t) = \tilde{B}(\psi(t))v(t)$ and the desired result follows from Theorem 2.

Example 5 (Uniform approximation of a trajectory of the completed system by a trajectory of the original system). The system considered is

(i) $\dot{x} = B(x)u$, B a 3×2 matrix with columns

$$b^1(x) = \left(0, 1, \frac{1}{1 + (1 + x_1)^2}\right), \qquad b^2(x) = (1, 0, 0).$$

Its completed system is

(ii) $\dot{x} = \tilde{B}(x)\tilde{u}$, $\tilde{B}$ a 3×3 matrix with columns $b^1(x), b^2(x)$, as above, and

$$b^3(x) = [b^1, b^2](x) = \left(0, 0, \frac{-2(1 + x_1)}{(2 + 2x_1 + x_1^2)^2}\right).$$

The solution ψ of the completed system which we will approximate will be for $\tilde{u} = (0, 0, -1)$ and initial data $x^0 = (0, 0, 0)$. Thus $\psi(t) = (0, 0, \frac{1}{2}t)$. This solution uses only the contribution of the Jacobi bracket of b^1 and b^2. Thus if $T^i(t)y$ denotes the solution of $\dot{x} = b^i(x), x(0) = y$, $i = 1, 2$, we know from the interpretation of the Jacobi bracket that we should expect to approximate $\psi(t)$ by $T^2(-t)T^1(-t)T^2(t)T^1(t)x^0$. One may note that by varying the magnitude of the control vector u, one may vary the speed of traversing a solution of $\dot{x} = b^i(x)$.

Let $u^1(\alpha) = \begin{pmatrix} \alpha \\ 0 \end{pmatrix}$, $u^2(\alpha) = \begin{pmatrix} 0 \\ \alpha \end{pmatrix}$, and define

$$u(t) = \begin{cases} u^1(\alpha) & \text{if } 0 \leqq t \leqq \gamma, \\ u^2(\alpha) & \text{if } \gamma < t \leqq 2\gamma, \\ -u^1(\alpha) & \text{if } 2\gamma < t \leqq 3\gamma, \\ -u^2(\alpha) & \text{if } 3\gamma < t \leqq 4\gamma \end{cases}$$

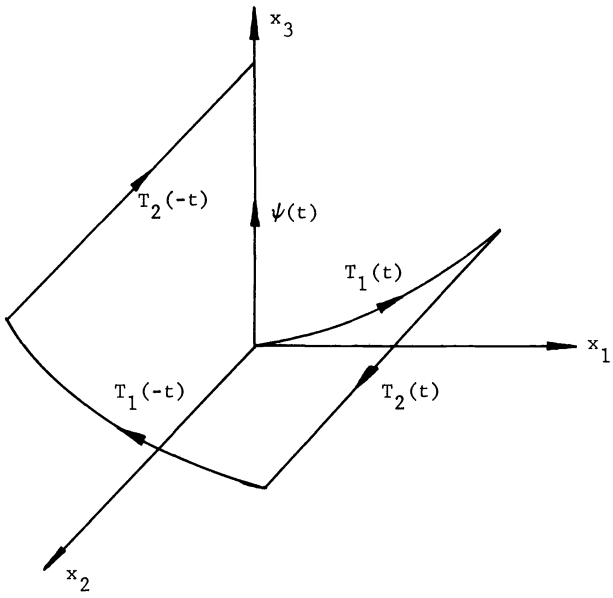


FIG. 1

for $\alpha, \gamma > 0$, and extend $u(\cdot)$ periodically. Let $\varphi(\cdot, u)$ denote the solution of (i) which corresponds to this choice of u and initial data $x^0 = (0, 0, 0)$. We may note that $\varphi(\gamma, u) = T^1(\alpha\gamma)x^0$, $\varphi(2\gamma, u) = T^2(\alpha\gamma)T^1(\alpha\gamma)x^0$, $\varphi(3\gamma, u) = T^1(-\alpha\gamma)T^2(\alpha\gamma)T^1 \cdot (\alpha\gamma)x^0$, $\varphi(4\gamma, u) = T^2(-\alpha\gamma)T^1(-\alpha\gamma)T^2(\alpha\gamma)T^1(\alpha\gamma)x^0$, etc.

Calculating the actual solution of (i) gives

$$\varphi(\gamma, u) = (0, \alpha\gamma, \alpha\gamma/2), \qquad \varphi(2\gamma, u) = (\alpha\gamma, \alpha\gamma, \alpha\gamma/2),$$

$$\varphi(3\gamma, u) = \left(\alpha\gamma, 0, \frac{2(\alpha\gamma)^2 + (\alpha\gamma)^3}{4 + 4\alpha\gamma + 2(\alpha\gamma)^2}\right), \qquad \varphi(4\gamma, u) = \left(0, 0, \frac{2(\alpha\gamma)^2 + (\alpha\gamma)^3}{4 + 4\alpha\gamma + 2(\alpha\gamma)^2}\right),$$

etc. (See Fig. 1.)

Now suppose we wish an $\varepsilon > 0$ uniform approximation to $\psi(\cdot)$ where we take, for $y \in R^3$, $\|y\| = \sum_{i=1}^{3} |y_i|$ and $0 < \varepsilon < 1$. Our goal will be to choose α and γ so that $\varphi(4k\gamma, u) = \psi(4k\gamma)$ for $k = 0, 1, \cdots$, and $\|\varphi(t, u) - \psi(t)\| \leqq \varepsilon$ for all t.

Let $\alpha\gamma = \varepsilon$. Then $|\varphi_1(t, u) - \psi_1(t)| \leqq \varepsilon$, $|\varphi_2(t, u) - \psi_2(t)| \leqq \varepsilon$ for $0 \leqq t \leqq 4\gamma$, and $|\psi_3(4\gamma) - \varphi_3(4\gamma, u)| = |2\gamma - \varepsilon^2[2 + \varepsilon]/(4 + 4\varepsilon + 2\varepsilon^2)|$. Choose $\gamma = \varepsilon^2[2 + \varepsilon]/(8 + 8\varepsilon + 4\varepsilon^2)$; hence $|\psi_3(4\gamma) - \varphi_3(4\gamma, u)| = 0$ and clearly $|\psi_3(t) - \varphi_3(4\gamma, u)| = 0$ and clearly $|\psi_3(t) - \varphi_3(t, u)| \leqq \varepsilon$ for $0 \leqq t \leqq 4\gamma$. This choice of γ gives $\alpha = (8 + 8\varepsilon + 4\varepsilon^2)/(\varepsilon[2 + \varepsilon])$; since α determines the "speed" with which we move along φ, we see that for small ε, γ is small (many switches) and α is large. The above choices of α and γ therefore yield $\psi(4k\gamma) - \varphi(4k\gamma, u) = 0$, $k = 0, 1, 2, \cdots$, and $\|\psi(t) - \varphi(t, u)\| \leqq \varepsilon$ for all t.

It is interesting to compare the results obtained for the systems (1), (2) with those which might be obtained for

$$\dot{x} = a(x) + B(x)u, \qquad \dot{x} = a(x) + \tilde{B}(x)u, \tag{4}$$

where B, $\tilde{B}$ are as in (1), (2) while $a(\cdot)$ is a C^∞, n-vector-valued function.

It is natural to ask whether points attainable by solutions of (5) are attainable by solutions of (4). In general, the answer is no, as will be shown by the following.

Example 6. We shall modify Example 3 so that the rank of the completed system is less than n, the spacial dimension. For $x = (x_1, \cdots, x_4)$ let

$$a(x) = (0, 0, 0, -(x_1^2 + x_2^2)).$$

Let

$$b^1(x) = \left(0, 1, \frac{1}{1 + (1 + x_1)^2}, 0\right), \qquad b^2(x) = (1, 0, 0, 0).$$

Then, as in Example 3,

$$[b^1, b^2](x) = b^3(x) = \left(0, 0, \frac{-2(1 + x_1)}{(2 + 2x_1 + x_1^2)^2}, 0\right),$$

and the completed system (2) has rank 3; i.e., $n = 4$, $s = 3$, $r = 2$.

Starting from the origin, with the completed system $\dot{x} = a(x) + \tilde{B}(x)\tilde{u}$, one may follow the x_3-axis (i.e., $x_1 = x_2 = x_4 = 0$) and hence attain, for instance, the final point $x^f = (0, 0, 1, 0)$. However, with the system $\dot{x} = a(x) + B(x)u$, we cannot keep $x_1 = x_2 = 0$ and increase x_3 to 1. Thus the term $-(x_1^2 + x_2^2)$ will necessarily give a contribution; i.e., we can reach points $(0, 0, 1, \sigma)$ but only with $\sigma < 0$.

Acknowledgment. The authors would like to thank C. Lobry for his helpful remarks and examples.

REFERENCES

[1] L. W. CHOW, *Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann., 117 (1940–41), pp. 98–105.

[2] R. HERMANN, *On the accessibility problem in control theory*, Proc. Symposium Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325–332.

[3] ———, *The differential geometry of foliations. II*, J. Math. Mech., 11 (1962), pp. 303–315.

[4] ———, *Some differential-geometric aspects of the Lagrange variational problem*, Illinois J. Math., 6 (1962), pp. 634–673.

[5] J. KUCHERA, *Solution in large of control problem* $\dot{x} = (Au + Bv)x$, Czechoslovak Math. J., 17 (1967), pp. 91–96.

[6] R. L. BISHOP AND R. J. CRITTENDEN, *Geometry of Manifolds*, Academic Press, New York, 1964.

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

THE OPTIMIZATION OF TRAJECTORIES OF LINEAR FUNCTIONAL DIFFERENTIAL EQUATIONS*

H. T. BANKS AND MARC Q. JACOBS†

Abstract. Our aim in this paper is to examine a number of fundamental questions in the theory of optimal control of processes monitored by certain general systems of linear functional differential equations with finite memories. In our model the controls may appear in a very general nonlinear functional manner which permits us to consider retardations of a rather general character in the control variables. In particular, we prove a maximal principle for such systems. We consider existence questions in the class of admissible Borel measurable (respectively piecewise continuous, almost piecewise continuous) initial functions and controls. We also show that certain solutions of an uncontrolled linear functional differential equation are piecewise analytic or quasi-piecewise analytic.

1. Introduction. The linear functional differential equation describing the controlled systems studied in this paper is given in (2.3) below. Many authors have studied control systems with delays in the state variables and there are several extensive bibliographies available in these areas [7], [18], [48], [20], [53]. Recently, models for systems with delays in the control parameters have been proposed and some results for these systems have been obtained [5], [6], [14], [22], [30], [34], [35], [38], [39]. Such models occur naturally in the study of gas-pressurized bipropellant rocket systems [14], in population models [5], [42], and in some complex economic models currently under study.

In § 2, we set down the notation, definitions and standing hypotheses that will be required throughout. In § 3, we prove (see Theorem 3.1) that the collection of points in R^n, which can be attained at time t from admissible Borel measurable initial functions and controls, is compact and depends continuously (with respect to the Hausdorff metric) on t. The assumptions required for this theorem are in effect no more than is usually required just to prove the existence of solutions to the linear functional differential equation (2.3) (see [3], [6]). Since the Lebesgue–Stieltjes measures, which will appear below in the variation of parameters formula (2.7), can be atomic, we cannot conclude that the abovementioned fixed-time cross sections of the attainable set are convex. However, if we add rather mild assumptions (properties (Δ_1) and (Δ_2) in § 3), then we do obtain the convexity of the fixed-time cross sections of the attainable set (see Theorem 3.2). We then adapt an argument of Lee and Markus [40] for ordinary control problems to obtain the statements of the maximal principle in § 4 (Theorem 4.1 and Remark 4.1). Theorems 3.1 and 3.2 can be regarded as extensions of some well-known results by Neustadt [46] and Olech [47]. Several very special cases of these two theorems have appeared in the literature [13], [38], [39], [48]. The actual statement of the maximal principle is confined to the time optimal control problem, although this is not an essential feature (cf. the remarks preceding Lemma 4.1). This maximal principle complements recent work of Banks [5] and Kharatishvili [34], [35],

* Received by the editors September 25, 1969.

† Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This research was supported in part by the National Aeronautics and Space Administration under Grant NGL-40-002-015 and in part by the Air Force Office of Scientific Research under Grant AF-AFOSR 67-0693A.

and in effect contains some of Lee's work [38], [39] as special cases, although Lee has considered a somewhat different class of cost functionals. Also our work in essence includes the necessary conditions determined by Halanay in [21]. Even in the cases where our work overlaps with that of the above authors, our methods of proof differ in that we have made extensive use of a number of fairly recent developments in the theory of measurable multifunctions [2], [10], [11], [29], [31], [32], [36], [47] to greatly simplify the arguments.[1]

In § 5 we turn to a study of analyticity properties of "fundamental matrix" solutions to certain systems of functional differential equations. Many authors (see the references in [7], [18], [53]) have studied various aspects of the analyticity of the solutions of very special types of functional differential equations, although none of these results appear to include those presented in § 5. Theorem 5.1 is a rather straightforward application of known results in ordinary differential equations. However, Theorem 5.2, which is extremely believable, seems to require a proof involving a substantially more intricate form of analysis than is needed to prove its simple counterpart in the theory of ordinary differential equations. It should be noted that the conclusion of Theorem 5.1 guarantees a type of piecewise analyticity of the "fundamental matrix," whereas Theorem 5.2 gives only what we have termed quasi-piecewise analyticity. One might expect that if the coefficient matrices in system (5.2) are analytic, and if one starts with an analytic initial function, then the solution of the functional differential equation will also be analytic. Indeed, several authors have attempted to prove such results (for example, see [48], [49]), but very simple examples reveal that such general theorems are not true (see Remark 5.1).

Finally, in § 6 we apply the aforementioned piecewise analyticity (respectively quasi-piecewise analyticity) properties to show that under certain circumstances the admissible initial functions and the admissible controls may be delimited to an appropriate class of piecewise continuous (respectively almost piecewise continuous) functions and the attainable set will be the same as if one were using Borel measurable admissible initial functions and controls. These results are simply analogues of those obtained by Halkin [24] for ordinary linear control problems using the work on subintegrals by Halkin and Hendricks [25]. Halkin's paper extends earlier work in [19], [23], [41].

2. Notation, definitions, and general hypotheses. If X and Y are nonempty sets, then a *multifunction* $\Omega: X \to Y$ is simply a subset of $X \times Y$ with domain equal to X; equivalently Ω is a mapping of X into the nonempty subsets of Y. If Y is a topological space and $\Omega(x)$ is compact for each $x \in X$, then we say $\Omega: X \to Y$ is a *compact multifunction*. If $\mathscr{A}$ is a σ-algebra of subsets of X and if Y is a topological space, then we say a multifunction $\Omega: X \to Y$ is *$\mathscr{A}$-measurable* if the set $\Omega^- F$ defined by

$$\Omega^- F \equiv \{x \in X | \Omega(x) \cap F \neq \varnothing\}$$

belongs to $\mathscr{A}$ for each closed $F \subset Y$. If X is a topological space and $\mathscr{A}$ is the collection of Borel sets in X, then we shall write *Borel measurable* instead of $\mathscr{A}$-measurable.

---

[1] The authors are grateful to a referee for pointing out that related results were recently presented by F. M. Kirillova et al., C. Olech, and D. H. Chyung and E. B. Lee during the Fourth IFAC Congress, Warsaw, 1969.

It is assumed that a definite norm $|\cdot|$ is given on any of the finite-dimensional vector spaces which come into our discussion. The closed ball in R^p with center at the origin and radius r will be denoted by $S^p(r)$.

The real vector space of all real $p \times q$ matrices will be denoted by $\mathscr{L}_{pq}$ for any pair of positive integers p and q. Let $[a, b]$ be a compact interval in R, and let $H:[a, b] \to \mathscr{L}_{pq}$ be a function of bounded variation. We shall use μ_H to denote the Lebesgue–Stieltjes measure on $[a, b]$ determined by H (see [16, p. 358 ff.]). In constructing such measures from H, *H will always be taken to be left continuous on* (a, b). We observe that if $t \to T_H(t)$, $t \in [a, b]$, denotes the scalar function defined by

$$T_H(t) = \mathrm{Var}_{s\in[a,t]} H(s), \qquad t \in [a, b],$$

and if $|\mu_H|$ denotes the variation of the Lebesgue–Stieltjes measure μ_H, then one has [16, p. 362]

$$|\mu_H| = \mu_{T_H}. \tag{2.1}$$

For conciseness, we frequently use $|H|(t)$ for $T_H(t)$ (this should not be confused with $|H(t)|$ which is the norm of the matrix $H(t)$). If $g:[a, b] \to R^p$ is μ_H-integrable, then $\int_a^b g(t)\, dH(t)$ denotes the integral of g over $[a, b]$ with respect to the measure μ_H. We use $\mathscr{L}_1([a, b], \mu_H, R^p)$ to denote the collection of all μ_H-integrable functions $g:[a, b] \to R^p$.

If $\Omega:[a, b] \to R^p$ is a multifunction, then $\int_a^b \Omega(t)\, dH(t)$ is used to denote the set (possibly empty)

$$\left\{ \int_a^b g(t)\, dH(t) \,|\, g \in \mathscr{L}_1([a, b], \mu_H, R^p),\ g(t) \in \Omega(t),\ a \leqq t \leqq b \right\}$$

(cf. [2], [10], [11], [15], [26], [47], [33]).

We shall deal frequently with mappings $f:X \times Y \to Z$, where X, Y, Z are sets. It will be convenient to use $f(x, \cdot)$, where x is a fixed element of X, to denote the mapping $y \to f(x, y)$, $y \in Y$. The mapping $f(\cdot, y):X \to Z$ for y a fixed element of Y is similarly defined.

Throughout the paper we make the following standing hypotheses: (H1) F and G are two Lebesgue measurable mappings from $R \times R$ into $\mathscr{L}_{nn}$; (H2) $F(t, s) = 0$ for $s \geqq 0$; (H3) $F(t, s) = F(t, -\tau)$ for $s \leqq -\tau$, where τ is a given positive constant; (H4) $G(t, s) = 0$ for $s \geqq t$; (H5) $G(t, s) = G(t, -\tau)$ for $s \leqq -\tau$; (H6) for each fixed $t \in R$ the functions $G(t, \cdot)$ and $F(t, \cdot)$ are of bounded variation on R; (H7) there is a Lebesgue measurable function $\beta:R \to R$ which is Lebesgue summable on every finite interval and which satisfies

$$\begin{aligned} &\mathrm{Var}_{s\in R} F(t, s) = \mathrm{Var}_{s\in[-\tau,0]} F(t, s) \leqq \beta(t), && t \in R, \\ &\mathrm{Var}_{s\in R} G(t, s) = \mathrm{Var}_{s\in[-\tau,t]} G(t, s) \leqq \beta(t), && t \in R. \end{aligned} \tag{2.2}$$

Let $h:R^m \times R \to R^n$ be a given function such that for each $t \in R$ the function $u \to h(u, t)$, $u \in R^m$, is continuous, and for each $u \in R^m$ the function $t \to h(u, t)$, $t \in R$, is Borel measurable. We shall consider control systems which can be

described by systems of real functional differential equations (FDE's) of the form

$$\dot{x}(t) = \int_{-\tau}^{0} x(t+s)\, d_s F(t,s) + \int_{-\tau}^{t} h(u(s),s)\, d_s G(t,s), \tag{2.3}$$

where both integrals in (2.3) are understood in the Lebesgue–Stieltjes sense with the symbol d_s being used to emphasize that the measures are constructed from the functions $F(t,\cdot)$ and $G(t,\cdot)$.

Let $U:[-\tau,\infty)\to R^m$ and $\Phi:[-\tau,0]\to R^n$ be given Borel measurable, compact multifunctions. It will be assumed that there is a positive constant M such that

$$\begin{aligned} &U(t)\subset S^m(M), \quad h(U(t),t)\subset S^n(M), \quad t\geqq -\tau,\\ &\Phi(t)\subset S^n(M), \qquad -\tau\leqq t\leqq 0. \end{aligned} \tag{2.4}$$

A triple $\{\varphi,u,t_1\}$ is called *admissible* if $\varphi:[-\tau,0]\to R^n$ and $u:[-\tau,t_1]\to R^m$, $t_1\geqq 0$, are Borel measurable functions satisfying

$$\begin{aligned} &\varphi(t)\in\Phi(t), \qquad -\tau\leqq t\leqq 0,\\ &u(t)\in U(t), \qquad -\tau\leqq t\leqq t_1. \end{aligned} \tag{2.5}$$

The selection theorem of Kuratowski and Ryll–Nardzewski [36] assures the existence of admissible triples.

Remark 2.1. It is noted that if $u:[a,b]\to R^m$ is a Borel measurable function, then the function $t\to h(u(t),t)$, $t\in[a,b]$, is also Borel measurable. This follows easily from the fact that there is a sequence of Borel functions, $u_n:[a,b]\to R^m$, whose range is a countable set, and which satisfy $\lim u_n(t)=u(t)$ for each $t\in[a,b]$. It follows now from the assumptions on h that $t\to h(u_n(t),t)$, $t\in[a,b]$, are each Borel measurable functions and $\lim h(u_n(t),t)=h(u(t),t)$, $t\in[a,b]$. Consequently, $t\to h(u(t),t)$, $t\in[a,b]$, is Borel measurable.

For any admissible triple $\{\varphi,u,t_1\}$ there is a unique absolutely continuous function (or response) $t\to x(t,\varphi,u)$, $0\leqq t\leqq t_1$, satisfying (2.3) almost everywhere on $[0,t_1]$ and the initial condition

$$x(t,\varphi,u)=\varphi(t), \qquad -\tau\leqq t\leqq 0. \tag{2.6}$$

According to the variation of parameters formula [3], this response is given by

$$\begin{aligned} x(t,\varphi,u) = \varphi(0)Y(0,t) &+ \int_{-\tau}^{0}\varphi(s)\, d_s\left\{\int_0^{\tau} F(\alpha,s-\alpha)Y(\alpha,t)\, d\alpha\right\}\\ &+\int_0^t\left\{\int_{-\tau}^{\alpha} h(u(s),s)\, d_s G(\alpha,s)\right\}Y(\alpha,t)\, d\alpha, \end{aligned} \tag{2.7}$$

where for fixed $t\geqq 0$ the function $s\to Y(s,t)$, $0\leqq s\leqq t$, is an $n\times n$ matrix solution of

$$Y(s,t)+\int_s^t F(\alpha,s-\alpha)Y(\alpha,t)\, d\alpha = E, \qquad 0\leqq s\leqq t, \tag{2.8}$$

which is of bounded variation and which satisfies $Y(t,t)=E$, the $n\times n$ identity matrix, and $Y(s,t)\equiv 0$ for $s>t$.

A point $x \in R^n$ is *attainable* if there is an admissible triple $\{\varphi, u, t_1\}$ such that $x(t_1, \varphi, u) = x$. The *attainable set* $\mathscr{A}(\Phi, U)$ (or simply $\mathscr{A}$ when Φ and U are understood) is defined by the equation

$$\mathscr{A}(\Phi, U) \equiv \{x \in R^n | x \text{ is attainable}\}.$$

The fixed-time cross sections of $\mathscr{A}(\Phi, U)$ at $t \geqq 0$ are denoted by $\mathscr{A}_t(\Phi, U)$ (or simply $\mathscr{A}_t$ when Φ and U are understood) and are defined by the equation

$$\mathscr{A}_t(\Phi, U) \equiv \{x \in R^n | \text{there exist } \{\varphi, u, t\} \text{ admissible such that } x(t, \varphi, u) = x\}.$$

3. Properties of the attainable set without convexity assumptions. We begin with some simple lemmas and observations.

LEMMA 3.1. *Let the standing hypotheses of* § 2 *be satisfied. Then*

$$|Y(s, t)| \leqq |E| \exp \int_s^t \beta(\xi)\, d\xi, \qquad 0 \leqq s \leqq t.$$

Proof. This is an easy consequence of (2.8) and the boundary conditions.

Remark 3.1. If $\mathscr{I}$ is a compact interval and $H : \mathscr{I} \to \mathscr{L}_{pq}$ is of bounded variation, then H has the well-known decomposition into a sum of a singular function, an absolutely continuous function, and a saltus (jump) function. We note also that if $H = A + N$ where A is the saltus function and N is continuous, then $\operatorname{Var} H = \operatorname{Var} A + \operatorname{Var} N$. It is also observed that if H is continuous, then $t \to T_H(t)$, $t \in \mathscr{I}$, is also continuous. Consequently, from (2.1) it can be shown that $|\mu_H|$ is nonatomic whenever H is continuous.

The next lemma is in essence contained in the papers of Liapunov [43], Blackwell [8], and Olech [47]. There are, however, some technical differences so we include a proof for the sake of completeness.

LEMMA 3.2. *Let $\mathscr{I}$ be a compact interval and let $H : \mathscr{I} \to \mathscr{L}_{pq}$ be of bounded variation on $\mathscr{I}$. Let $\Omega : \mathscr{I} \to R^p$ be a μ_H-measurable compact multifunction. Let $\rho \in \mathscr{L}_1(\mathscr{I}, |\mu_H|, R)$ be such that $\Omega(t) \subset S^p(\rho(t))$, $t \in \mathscr{I}$. Then $\int_{\mathscr{I}} \Omega(t)\, dH(t)$ is compact.*

Proof. First we observe that by the Lebesgue–Nikodym theorem (for example, see [16, p. 263]) there is a $|\mu_H|$-integrable function $B : \mathscr{I} \to \mathscr{L}_{pq}$ such that

$$\int_{\mathscr{I}} g(t)\, dH(t) = \int_{\mathscr{I}} g(t) B(t)\, d|H|(t), \qquad g \in \mathscr{L}_1(\mathscr{I}, \mu_H, R^p).$$

We write $T_H = \alpha + \nu$, where α is the saltus function of T_H and ν is continuous. It is an easy matter to prove that the multifunction $t \to \Omega(t)B(t)$, $t \in \mathscr{I}$, is measurable, where $\Omega(t)B(t) = \{x \in R^q | x = yB(t) \text{ for some } y \in \Omega(t)\}$. Moreover, $\Omega(t)B(t) \subset S^q(\rho(t)|B(t)|)$, $t \in \mathscr{I}$. One can also verify the identities:

$$(3.1) \quad \int_{\mathscr{I}} \Omega(t)\, dH(t) = \int_{\mathscr{I}} \Omega(t) B(t)\, d|H|(t) = \int_{\mathscr{I}} \Omega(t) B(t)\, d\alpha(t) + \int_{\mathscr{I}} \Omega(t) B(t)\, d\nu(t);$$

the proof of the first equality is facilitated by versions of Filippov's selection lemma [11], [31], and the second equality follows from the definition of α and ν. Now μ_α is purely atomic and μ_ν is nonatomic so the conclusion of the lemma follows from (3.1) and a remark of Olech's [47, p. 100] (see [11] also for the nonatomic case).

LEMMA 3.3. *Let H and $\mathscr{I}$ be as in Lemma* 3.2. *Let $\Omega:\mathscr{I} \to R^p$ be a multifunction with a Borel measurable selection; that is, there is a Borel function $g^*:\mathscr{I} \to R^p$ such that $g^*(t) \in \Omega(t)$ for each $t \in \mathscr{I}$. Then $\int_{\mathscr{I}} \Omega(t)\, dH(t)$ coincides with the set $\mathscr{B}(\Omega, H)$*

$$\equiv \left\{ \int_{\mathscr{I}} g(t)\, dH(t) \middle| \text{ the function } g:\mathscr{I} \to R^p \text{ is Borel measurable and } g(t) \in \Omega(t),\ t \in \mathscr{I} \right\}.$$

Proof. Clearly $\mathscr{B}(\Omega, H) \subset \int_{\mathscr{I}} \Omega(t)\, dH(t)$. Conversely, suppose $g \in \mathscr{L}_1(\mathscr{I}, \mu_H, R^p)$, $g(t) \in \Omega(t)$, $t \in \mathscr{I}$. Then there is a Borel set $E_0 \subset \mathscr{I}$ with $|\mu_H|(E_0) = 0$ and there is a Borel function $\bar{g}:\mathscr{I} \to R^p$ such that $\bar{g} = g$ on $\mathscr{I} \setminus E_0$ [50, p. 225]. Using χ_S for the characteristic function of a set S we see that $\tilde{g} \equiv \bar{g} \cdot \chi_{\mathscr{I} \setminus E_0} + g^* \cdot \chi_{E_0}$ is a Borel function satisfying $\tilde{g}(t) \in \Omega(t)$, $t \in \mathscr{I}$, and $\tilde{g} = g$ a.e. $[\mu_H]$. Hence, $\int_{\mathscr{I}} \tilde{g}(t)\, dH(t) = \int_{\mathscr{I}} g(t)\, dH(t)$, and so $\int_{\mathscr{I}} g(t)\, dH(t) \in \mathscr{B}(\Omega, H)$. This completes the proof.

In preparation for the next lemma let us introduce some additional notation. $\tilde{F}, \tilde{G}:[0, \infty) \times R \to \mathscr{L}_{nn}$ are mappings defined by the following two relations:

$$\tilde{F}(t, s) \equiv \int_0^\tau F(\alpha, s - \alpha) Y(\alpha, t)\, d\alpha,$$

$$\tilde{G}(t, s) \equiv \int_0^t G(\alpha, s) Y(\alpha, t)\, d\alpha, \qquad t \geqq 0, \quad s \in R,$$

where F, G, and Y are the functions defined in § 2 which appear in (2.3) and (2.7). We define a function $\mathscr{H}:R^m \times R \to R^m \times R^n$ by the equation

$$\mathscr{H}(u, t) \equiv (u, h(u, t)), \qquad t \in R, \quad u \in R^m,$$

where h is the function introduced in § 2 (see (2.3)). A function $\Gamma:[0, \infty) \times R \to \mathscr{L}_{(m+n)n}$ is defined by the equation

$$\Gamma(t, s) = \begin{bmatrix} O_{mn} \\ \tilde{G}(t, s) \end{bmatrix}, \qquad t \geqq 0, \quad s \in R,$$

where O_{mn} denotes an $m \times n$ matrix all of whose entries are zero. A multifunction $L:R \to R^m \times R^n$ is defined by the condition

$$L(t) = \mathscr{H}(U(t), t), \qquad t \geqq -\tau.$$

Remark 3.2. The sets $L(t)$, $t \geqq -\tau$, are evidently compact. Let $b \geqq 0$ be given. If μ is any Lebesgue–Stieltjes measure on $[-\tau, b]$, then the multifunction $U|[-\tau, b]$ is μ-measurable. This follows from the assumption that U is Borel measurable. Using Lusin's theorem for μ-measurable multifunctions [11], [31] and extensions of Scorza–Dragoni's theorem [11], [32], we can prove that the multifunction $t \to L(t) = \mathscr{H}(U(t), t)$, $t \in [-\tau, b]$, is μ-measurable. Hence $L|[-\tau, b]$ is μ-measurable for every Lebesgue–Stieltjes measure μ on $[-\tau, b]$. We note also that because U is Borel measurable and compact, there is a Borel measurable function $u^*:[-\tau, \infty) \to R^m$ such that $u^*(t) \in U(t)$, $t \geqq -\tau$ (see [36]). Remark 2.1 shows then that L has a Borel measurable selection. Recall now that Φ was also assumed

to be Borel measurable. Hence, in evaluating either of the integrals

$$\int_{-\tau}^{0} \Phi(s)\, d_s\tilde{F}(t,s) \quad \text{or} \quad \int_{-\tau}^{t} L(s)\, d_s\Gamma(t,s),$$

the conclusion of Lemma 3.3 may be applied.

LEMMA 3.4. *Let the standing hypotheses of* § 2 *be satisfied. For* $t \geqq 0$ *define* $\mathscr{R}_t(\Phi, U)$ *to be the set*

$$\Phi(0)Y(0,t) + \int_{-\tau}^{0} \Phi(s)\, d_s\tilde{F}(t,s) + \int_{-\tau}^{t} L(s)\, d_s\Gamma(t,s).$$

Then we have the identity

$$\mathscr{A}_t(\Phi, U) = \mathscr{R}_t(\Phi, U), \qquad t \geqq 0.$$

Proof. Examining the third summand on the right-hand side of (2.7) we use the assumptions on G in § 2 to write

$$\begin{aligned}\int_0^t \left\{\int_{-\tau}^{\alpha} h(u(s),s)\, d_sG(\alpha,s)\right\} Y(\alpha,t)\, d\alpha &= \int_0^t \left\{\int_{-\tau}^{t} h(u(s),s)\, d_sG(\alpha,s)\right\} Y(\alpha,t)\, d\alpha \\ &= \int_{-\tau}^{t} h(u(s),s)\, d_s\left\{\int_0^t G(\alpha,s)Y(\alpha,t)\, d\alpha\right\}.\end{aligned} \tag{3.2}$$

The last equality follows by utilizing an unsymmetric Fubini theorem [9] to interchange the order of integration. We have the identity[2]

$$\int_{-\tau}^{t} h(u(s),s)\, d_s\tilde{G}(t,s) = \int_{-\tau}^{t} \mathscr{H}(u(s),s)\, d_s\Gamma(t,s). \tag{3.3}$$

Consequently from (3.2), (3.3), and (2.7) we have

$$\mathscr{A}_t(\Phi, U) \subset \mathscr{R}_t(\Phi, U).$$

Since $F(t,\cdot)$ is left continuous on $(-\tau, 0)$, we have that if $-\tau \leqq s_n < 0$, and $s_n \to 0$ as $n \to \infty$, then $\lim F(\alpha, s_n - \alpha) = F(\alpha, -\alpha)$, $0 < \alpha \leqq \tau$. Therefore from the Lebesgue dominated convergence theorem and the definition of $\tilde{F}$ we get that $\{0\}$ is not an atom of $\mu_{\tilde{F}(t,\cdot)}$. Hence

$$\int_{-\tau}^{0} \varphi(s)\, d_s\tilde{F}(t,s) = \int_{-\tau}^{0} \bar{\varphi}(s)\, d_s\tilde{F}(t,s)$$

if $\varphi(s) = \bar{\varphi}(s)$ except at $s = 0$. From this remark, the variation of parameters formula (2.7), and Remark 3.2 one can show the reverse inclusion $\mathscr{R}_t(\Phi, U) \subset \mathscr{A}_t(\Phi, U)$. If the detailed proof of this inclusion is carried out, then the meaning of the comment in the preceding footnote becomes clear. This completes the proof of the representation formula of the lemma.

THEOREM 3.1. *Let the standing hypotheses of* § 2 *be satisfied. Then*:

(i) *the sets* $\mathscr{A}_t(\Phi, U)$, $t \geqq 0$, *are compact*;

[2] Our reason for introducing the auxiliary functions $\mathscr{H}$ and Γ is to avoid certain questions concerning the existence of Borel measurable selections. Halkin used a similar device in [24].

(ii) *the mapping* $t \to \mathscr{A}_t(\Phi, U)$, $t \geqq 0$, *taking its values in the compact nonempty subsets of* R^n *is continuous with respect to the Hausdorff metric* [1];

(iii) *for any* $\bar{t} \geqq 0$ *the set* $\bigcup_{t\in[0,\bar{t}]} \mathscr{A}_t(\Phi, U)$ *is compact.*

Proof of (i). This is an immediate consequence of the representation formula in Lemma 3.4 and Lemma 3.2.

Proof of (iii). This is readily deduced from (ii).

Proof of (ii). Let S denote the closed unit ball in R^n with center at the origin. We must prove that given $t_1 \geqq 0$ and $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\mathscr{A}_{t_1} + \varepsilon S \supset \mathscr{A}_t \quad \text{and} \quad \mathscr{A}_t + \varepsilon S \supset \mathscr{A}_{t_1}, \qquad |t - t_1| \leqq \delta, \quad t \geqq 0. \tag{3.4}$$

The relations in (3.4) can be verified by considering two cases: (a) $t \geqq t_1$; (b) $0 \leqq t < t_1$. Consider case (a) first. Suppose $x_t \in \mathscr{A}_t$; then there is an admissible triple $\{\varphi_t, u_t, t\}$ such that $x_t = x(t, \varphi_t, u_t)$. Define the function $\bar{u}:[-\tau, t_1] \to R^m$ to be the restriction of u_t to $[-\tau, t_1]$. Using the variation of parameters formula (2.7), Lemma 3.1, inequalities (2.2) and (2.4), hypothesis (H4) of § 2 and some standard manipulations with Lebesgue–Stieltjes integrals, we obtain the estimate

$$\begin{aligned} |x(t, \varphi_t, u_t) - x(t_1, \varphi_t, \bar{u})| &\leqq M|Y(0, t) - Y(0, t_1)| \\ &\quad + M \int_0^\tau \beta(\alpha)|Y(\alpha, t) - Y(\alpha, t_1)|\, d\alpha \\ &\quad + M|E| \int_{t_1}^t \beta(\alpha)\left[\exp \int_\alpha^t \beta(\xi)\, d\xi\right] d\alpha \\ &\quad + M \int_0^{t_1} \beta(\alpha)|Y(\alpha, t) - Y(\alpha, t_1)|\, d\alpha. \end{aligned} \tag{3.5}$$

We now give a similar estimate for (b). By the Kuratowski–Ryll–Nardzewski selection theorem [36] there is a Borel function $u^*:[-\tau, \infty) \to R^m$ such that $u^*(t) \in \Omega(t), t \geqq -\tau$. We note that $u_1 \equiv u_t \cdot \chi_{[-\tau,t]} + u^* \cdot \chi_{(t,t_1]}$ is a Borel function and $\{\varphi_t, u_1, t_1\}$ is admissible. For reasons similar to those adduced to support (3.5) we get the inequality

$$\begin{aligned} |x(t, \varphi_t, u_t) - x(t_1, \varphi_t, u_1)| &\leqq M|Y(0, t) - Y(0, t_1)| \\ &\quad + M \int_0^\tau \beta(\alpha)|Y(\alpha, t) - Y(\alpha, t_1)|\, d\alpha \\ &\quad + M|E| \int_t^{t_1} \beta(\alpha)\left[\exp \int_\alpha^{t_1} \beta(\xi)\, d\xi\right] d\alpha \\ &\quad + M \int_0^t \beta(\alpha)|Y(\alpha, t) - Y(\alpha, t_1)|\, d\alpha. \end{aligned} \tag{3.5'}$$

From the continuity of $Y(\alpha, \cdot)$, the Lebesgue dominated convergence theorem and inequalities (3.5) and (3.5'), there results the following statement.

Statement 3.6. Given $t_1 \geqq 0$ and $\varepsilon > 0$ there is a $\delta > 0$ depending only on t_1 and ε such that $|t - t_1| \leqq \delta$, $t \geqq 0$, implies $|x(t, \varphi_t, u_t) - x(t_1, \varphi_t, \bar{u})| \leqq \varepsilon$ and $|x(t, \varphi_t, u_t) - x(t_1, \varphi_t, u_1)| \leqq \varepsilon$.

Statement 3.6 implies

$$\mathscr{A}_{t_1} + \varepsilon S \supset \mathscr{A}_t, \quad |t - t_1| \leqq \delta, \quad t \geqq 0. \tag{3.6}$$

The other inclusion relationship in (3.4) is proved by a symmetric argument which is omitted.

Remark 3.3. Let $t \to \mathscr{T}(t)$, $t \geqq 0$, be a compact multifunction which is continuous with respect to the Hausdorff metric. If we impose a terminal condition of the form

$$x(t_1, \varphi, u) \in \mathscr{T}(t_1), \tag{3.7}$$

then by the usual device [46], Theorem 3.1 yields an existence theorem for the time optimal control problem. If we consider only admissible controls whose domains $[-\tau, t_1]$ lie in some fixed interval $[-\tau, \bar{t}]$, and if there is a terminal constraint (3.7) or, indeed, if the right end is free, then Theorem 3.1 can be used to give an existence theorem for the problem of minimizing $P(x(t_1, \varphi, u))$ on the class of admissible triples $\{\varphi, u, t_1\}$ such that (3.7) is satisfied,[3] or for the problem of minimizing $P(x(t_1, \varphi, u))$ on the class of admissible triples $\{\varphi, u, t_1\}$, where P is a real-valued continuous function on R^n.

In order to deduce necessary conditions for the optimization problems mentioned in Remark 3.3 it is desirable to have that the sets $\mathscr{A}_t(\Phi, U)$ are convex. Simple examples show that this cannot be deduced under the general circumstances of Theorem 3.1 because the Lebesgue–Stieltjes measures involved in the representation formula, $\mathscr{A}_t(\Phi, U) = \mathscr{R}_t(\Phi, U)$, of Lemma 3.4 can be atomic. It is noted that any function on an interval $[a, b]$ into $\mathscr{L}_{pq}$ which is of bounded variation has only a denumerable number of discontinuities. We say that F has *property* (Δ_1) if for each $t \in R$ it is possible to index the points $-\theta_i(t)$, $i = 1, 2, \cdots$, in the interior of $[-\tau, 0]$ at which $F(t, \cdot)$ is discontinuous, in such a way that continuous functions $t \to \theta_i(t)$, $t \in R$, are defined and $t \to t - \theta_i(t)$, $t \in R$, is strictly increasing, $i = 1, 2, \cdots$. We say that G has *property* (Δ_2) if for each $t \in R$, $G(t, \cdot)$ is continuous at $-\tau$ and it is possible to index the points $\zeta_i(t)$, $i = 1, 2, \cdots$, in the interior of $[-\tau, t]$, at which $G(t, \cdot)$ is discontinuous in such a way that continuous strictly increasing functions $t \to \zeta_i(t)$, $t \in R$, $i = 1, 2, \cdots$, are defined.

THEOREM 3.2. *If in addition to the standing hypotheses of* § 2 *we assume that* $\Phi(0)$ *is convex, F and G are Borel measurable, F has property* (Δ_1) *and G has property* (Δ_2), *then conclusions* (i), (ii), *and* (iii) *of Theorem* 3.1 *are still valid and* $\mathscr{A}_t(\Phi, \Omega)$, $t \geqq 0$, *are convex.*

Before proceeding with the proof we give another lemma that will be useful in the proof.

LEMMA 3.5. *Let* $\rho : R \to R$ *be a continuous strictly increasing function. Let* $f : [a, b] \to \mathscr{L}_{pq}$ *be a Lebesgue summable function. We define three functions* W_1,

[3] Actually for these existence statements it is not necessary to assume that the multifunction $\mathscr{T}$ is continuous or even compact. It suffices to have the multifunction $\mathscr{T}$ closed (i.e., $\mathscr{T}(t)$ is closed for $t \geqq 0$) and upper semicontinuous in the Kuratowski sense (see for example [37], [12], [31]). We keep the stronger hypothesis of continuity because it is needed in proving necessary conditions for a minimum.

$W_2, W_3 : R \to \mathscr{L}_{pq}$ by the equations

$$W_1(s) = \int_a^b f(\xi) I(s - \rho(\xi))\, d\xi, \qquad s \in R,$$

$$W_2(s) = \int_a^b f(\xi) J(s - \rho(\xi))\, d\xi, \qquad s \in R,$$

$$W_3(s) = \int_a^b f(\xi) J(-s + \rho(\xi))\, d\xi, \qquad s \in R,$$

where $I, J : R \to R$ are the step functions defined by the relations

$$I(x) = \begin{cases} 0, & x \leqq 0, \\ 1, & x > 0, \end{cases} \qquad J(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geqq 0. \end{cases}$$

Then W_i is continuous, $i = 1, 2, 3$.

Proof. First we remark that $\rho : [a, b] \to [\rho(a), \rho(b)]$ has a continuous inverse $\rho^{-1} : [\rho(a), \rho(b)] \to [a, b]$ which is also strictly increasing. Some elementary calculations yield the following formulas:

$$W_1(s) = \begin{cases} 0 & \text{for } s \leqq \rho(a), \\ \displaystyle\int_a^b f(\xi)\, d\xi & \text{for } s > \rho(b), \\ \displaystyle\int_a^{\rho^{-1}(s)} f(\xi)\, d\xi & \text{for } \rho(a) < s \leqq \rho(b), \end{cases}$$

$$W_2(s) = \begin{cases} 0 & \text{for } s < \rho(a), \\ \displaystyle\int_a^b f(\xi)\, d\xi & \text{for } s \geqq \rho(b), \\ \displaystyle\int_a^{\rho^{-1}(s)} f(\xi)\, d\xi & \text{for } \rho(a) \leqq s < \rho(b) \end{cases}$$

and

$$W_3(s) = \begin{cases} 0 & \text{for } s > \rho(b), \\ \displaystyle\int_a^b f(\xi)\, d\xi & \text{for } s \leqq \rho(a), \\ \displaystyle\int_{\rho^{-1}(s)}^b f(\xi)\, d\xi & \text{for } \rho(a) < s \leqq \rho(b). \end{cases}$$

The continuity of the functions W_i, $i = 1, 2, 3$, is an immediate consequence of these formulas and the continuity of ρ^{-1} on $[\rho(a), \rho(b)]$.

Proof of Theorem 3.2. We write $F(t, s) = A_F(t, s) + N_F(t, s)$ and $G(t, s) = A_G(t, s) + N_G(t, s)$, where $A_F(t, \cdot)$ is the saltus function for $F(t, \cdot)$, $A_G(t, \cdot)$ is the saltus function for $G(t, \cdot)$, and both $N_F(t, \cdot)$ and $N_G(t, \cdot)$ are continuous. Denote the jump of $F(t, \cdot)$ at $-\theta_i(t)$ by $B_i(t)$ and the jump of $G(t, \cdot)$ at $\zeta_i(t)$ by $C_i(t)$, $i = 1, 2, \cdots$. The jump of $F(t, \cdot)$ at $-\tau$ is denoted by $B_{-1}(t)$ and the jump of $F(t, \cdot)$ at 0 is denoted

by $B_0(t)$. The jump of $G(t,\cdot)$ at t is denoted by $C_0(t)$. From Remark 3.1 and inequalities (2.2) it follows that

$$\sum_{i=1}^{\infty} |B_i(t)|,\ \sum_{i=1}^{\infty} |C_i(t)| \leqq \beta(t), \qquad t \in R. \tag{3.8}$$

Since F and G are Borel measurable, the functions $B_i, C_j, i = -1, 0, 1, 2, \cdots$, $j = 0, 1, 2, \cdots$, are all Lebesgue measurable. For example, let us show B_i is Lebesgue measurable, $i \geqq 1$. Define $s_n(t) = 1/n - \theta_i(t)$, $n = 1, 2, 3, \cdots$; then $s_n(t) > -\theta_i(t)$ and $\lim s_n(t) = -\theta_i(t)$. Since $F(t,\cdot)$ is left continuous on $(-\tau, 0)$, we have

$$B_i(t) = \lim F(t, s_n(t)) - F(t, -\theta_i(t)).$$

Since F is Borel measurable, the functions $t \to F(t, s_n(t))$ and $t \to F(t, -\theta_i(t))$ are Borel measurable (a fortiori Lebesgue measurable). Hence B_i is a Borel function and thus Lebesgue measurable. The proof of the measurability of the other functions is similar. Define $B(t) \equiv \sum_{i=1}^{\infty} B_i(t)$ and $C(t) \equiv \sum_{i=1}^{\infty} C_i(t)$ (both series converge by (3.8)). The saltus functions A_F and A_G can be written in the form

$$A_F(t, s) = -B_{-1}(t)J(-s - \tau) + B_0(t)J(s) + \sum_{i=1}^{\infty} B_i(t)I(s + \theta_i(t)) \tag{3.9}$$

and

$$A_G(t, s) = C_0(t)J(s - t) + \sum_{i=1}^{\infty} C_i(t)I(s - \zeta_i(t)).$$

We have

$$\begin{aligned} \tilde{F}(t, s) &= \int_0^{\tau} A_F(\alpha, s - \alpha)Y(\alpha, t)\, d\alpha + \int_0^{\tau} N_F(\alpha, s - \alpha)Y(\alpha, t)\, d\alpha, \\ \tilde{G}(t, s) &= \int_0^{t} A_G(\alpha, s)Y(\alpha, t)\, d\alpha + \int_0^{t} N_G(\alpha, s)Y(\alpha, t)\, d\alpha, \end{aligned} \tag{3.10}$$

and the second terms on the right-hand side of both equations depend continuously on s by the Lebesgue dominated convergence theorem. Using (3.8), (3.9) and the dominated convergence theorem we get

$$\begin{aligned} \int_0^{\tau} A_F(\alpha, s - \alpha)Y(\alpha, t)\, d\alpha = &- \int_0^{\tau} B_{-1}(\alpha)Y(\alpha, t)J(-s + \alpha - \tau)\, d\alpha \\ &+ \int_0^{\tau} B_0(\alpha)Y(\alpha, t)J(s - \alpha)\, d\alpha \\ &+ \sum_{i=1}^{\infty} \int_0^{\tau} B_i(\alpha)Y(\alpha, t)I(s - \alpha + \theta_i(\alpha))\, d\alpha. \end{aligned} \tag{3.11}$$

According to property (Δ_1) and Lemma 3.5, each term in the series (3.11) is continuous in s. We also have

$$\left| \int_0^{\tau} B_i(\alpha)Y(\alpha, t)I(s - \alpha + \theta_i(\alpha))\, d\alpha \right| \leqq K|E| \int_0^{\tau} |B_i(\alpha)|\, d\alpha, \qquad i = 1, 2, \cdots, \tag{3.12}$$

where $K = \exp \int_0^t \beta(\xi)\, d\xi$. Moreover, the series $\sum_{i=1}^{\infty} \int_0^{\tau} |B_i(\alpha)|\, d\alpha$ converges by (3.8). Hence, by the Weierstrass M-test and (3.12), the series in (3.11) converges uniformly for $s \in [-\tau, 0]$ ($t \geqq 0$ is fixed). Therefore the function $s \to \int_0^{\tau} A_F(\alpha, s-\alpha) \cdot Y(\alpha, t)\, d\alpha$, $-\tau \leqq s \leqq 0$, is continuous, and we conclude that $\tilde{F}(t, \cdot)$ is continuous on $[-\tau, 0]$ for each fixed $t \geqq 0$. By an entirely parallel argument it can be shown that $\tilde{G}(t, \cdot)$ (also $\Gamma(t, \cdot)$) is continuous for each fixed $t \geqq 0$. Using the Lebesgue–Nikodym theorem [16, p. 263] it is determined that there exist integrable Borel functions $V_{\tilde{F}}: [-\tau, 0] \to \mathscr{L}_{nn}$ and $V_{\Gamma}: [-\tau, t] \to \mathscr{L}_{(m+n)n}$ such that

$$\int_{-\tau}^{0} \varphi(s)\, d_s \tilde{F}(t, s) = \int_{-\tau}^{0} \varphi(s) V_{\tilde{F}}(s)\, d|\mu_{\tilde{F}(t,\cdot)}|$$

for $\varphi \in \mathscr{L}_1([-\tau, 0], \mu_{\tilde{F}(t,\cdot)}, R^n)$ and

$$\int_{-\tau}^{t} g(s)\, d_s \Gamma(t, s) = \int_{-\tau}^{t} g(s) V_{\Gamma}(s)\, d|\mu_{\Gamma(t,\cdot)}|$$

for $g \in \mathscr{L}_1([-\tau, t], \mu_{\Gamma(t,\cdot)}, R^n)$. From the representation formula in Lemma 3.4 and an extension of Filippov's selection principle [11], [31] we obtain

$$\mathscr{A}_t(\Phi, U) = \Phi(0) Y(0, t) + \int_{-\tau}^{0} \Phi(s) V_{\tilde{F}}(s)\, d|\mu_{\tilde{F}(t,\cdot)}| + \int_{-\tau}^{t} L(s) V_{\Gamma}(s)\, d|\mu_{\Gamma(t,\cdot)}|. \tag{3.13}$$

By Remark 3.1, $|\mu_{\tilde{F}(t,\cdot)}|$ and $|\mu_{\Gamma(t,\cdot)}|$ are nonatomic, and we conclude that $\mathscr{A}_t(\Phi, U)$ is convex (see [47], [11]).

We shall use co (B) to denote the convex hull of a set $B \subset R^p$.

COROLLARY 3.1. *Let Ψ denote the multifunction $t \to \text{co}\,(\Phi(t))$, $-\tau \leqq t < 0$. Let $U^*: [-\tau, \infty) \to R^m$ be a Borel measurable compact multifunction such that $U^*(t) \subset S^m(M)$ and* co $(h(U(t), t)) =$ co $(h(U^*(t), t))$ *for $t \geqq -\tau$. Let the hypotheses of Theorem* 3.2 *be satisfied. Then*

$$\mathscr{A}_t(\Phi, U) = \mathscr{A}_t(\Psi, U^*), \qquad t \geqq 0.$$

Proof. The identity is easily verified by using (3.13), the linearity of $V_{\tilde{F}}(s)$ and $V_{\Gamma}(s)$ and Theorem 7.1 in [11].

As a particular case of Corollary 3.1 we obtain the following corollary.

COROLLARY 3.2. *Let the hypotheses of Corollary* 3.1 *be satisfied. In addition suppose $\ddot{\Psi}(t)$ is the set of extreme points of $\Psi(t)$, and $U^{\#}: [-\tau, \infty) \to R^m$ is a multifunction such that $U^{\#}(t) \subset S^m(M)$ and the set of extreme points of* co $(h(U(t), t))$ *is equal to $h(U^{\#}(t), t)$, $t \geqq -\tau$. If the multifunctions $\ddot{\Psi}$ and $U^{\#}$ are compact and Borel measurable, then*

$$\mathscr{A}_t(\Phi, U) = \mathscr{A}_t(\ddot{\Psi}, U^{\#}), \qquad t \geqq 0.$$

Remark 3.4. In Theorem 3.2 it was assumed that $\Phi(0)$ is convex. If this should happen not to be the case, then one can always select Φ_0, a compact convex subset of $\Phi(0)$ (for example, Φ_0 could be a singleton point set), and define

$$\Phi^*(t) = \begin{cases} \Phi(t), & t \neq 0, \\ \Phi_0, & t = 0. \end{cases}$$

Since Φ^* is also Borel measurable, compact, and satisfies $\Phi(t) \subset S^n(M)$, $t \in [-\tau, 0]$, we could replace Φ by Φ^* and Theorem 3.2 could be applied.

4. Necessary conditions for an optimal control. The properties of the attainable sets deduced in § 3 suggest that the main geometric ideas involved in proving the maximal principle for ordinary linear control problems (see [40]) are going to retain their validity for certain of the optimization problems formulated in Remark 3.3. We shall only consider the time optimal control problem mentioned in Remark 3.3. It will be clear from the discussion that the results can be used to prove a maximal principle for the other problem discussed in the aforementioned remark if we add additional assumptions which assure that on compact convex subsets $\mathscr{A}_t$ of R^n the mapping P assumes its minimum on $\partial\mathscr{A}_t$, the boundary of $\mathscr{A}_t$, e.g., when P is linear (cf. [21]).

The following lemma is true and the arguments are entirely similar to those given in [40, pp. 72, 129].

LEMMA 4.1. *Let* $\mathfrak{F}$, $\mathfrak{G}:[a,b] \to R^n$ *be compact multifunctions which are continuous with respect to the Hausdorff metric. Let* $\mathfrak{F}(t)$ *be convex for* $a \leqq t \leqq b$. *Let* $t^* \in (a,b]$ *be such that* $\mathfrak{F}(t^*) \cap \mathfrak{G}(t^*) \neq \varnothing$ *and* $\mathfrak{F}(t) \cap \mathfrak{G}(t) = \varnothing$ *if* $a \leqq t < t^*$. *Then* $x^* \in \mathfrak{F}(t^*) \cap \mathfrak{G}(t^*)$ *implies* $x^* \in \partial\mathfrak{F}(t^*)$.

We shall use $\langle x, y\rangle$ to denote the scalar product, $x, y \in R^n$, and A' to denote the transpose of a matrix A.

THEOREM 4.1. *Let the hypotheses of Theorem* 3.2 *be satisfied. If* $\{\varphi^*, u^*, t^*\}$ *is an optimal solution to the time optimal control problem in Remark* 3.3, *then there is a function* $\psi:[0,t^*] \to R^n$ *which is of bounded variation and satisfies the adjoint equation*

$$\psi(s) + \int_s^{t^*} \psi(\alpha)F'(\alpha, s-\alpha)\,d\alpha = e, \qquad 0 \leqq s \leqq t^*,$$

where e *is an outward normal to a support hyperplane to the set* $\mathscr{A}_{t^*}(\Phi, U)$ *through the point* $x(t^*, \varphi^*, u^*)$ *on the boundary of* $\mathscr{A}_{t^*}(\Phi, U)$ *such that*

(i) $\langle \varphi^*(0), \psi(0)\rangle \geqq \langle \varphi_0, \psi(0)\rangle, \qquad \varphi_0 \in \Phi(0);$

(ii) $$\int_0^\tau \left\langle \int_{-\tau}^0 \varphi^*(s)\,d_sF(\alpha, s-\alpha), \psi(\alpha) \right\rangle d\alpha \geqq \int_0^\tau \left\langle \int_{-\tau}^0 \varphi(s)\,d_sF(\alpha, s-\alpha), \psi(\alpha) \right\rangle d\alpha$$

for every admissible φ;

(iii) $$\int_0^{t^*} \left\langle \int_{-\tau}^{t^*} h(u^*(s), s)\,d_sG(\alpha, s), \psi(\alpha) \right\rangle d\alpha \geqq \int_0^{t^*} \left\langle \int_{-\tau}^{t^*} h(u(s), s)\,d_sG(\alpha, s), \psi(\alpha) \right\rangle d\alpha$$

for every admissible u. *Moreover, if* $\mathscr{T}(t)$ *is equal to a fixed compact convex set* $\mathfrak{T} \subset R^n$ *for* $t \geqq 0$, *then* e *can be picked to satisfy the transversality condition*: e *is normal to a common support hyperplane separating* $\mathscr{A}_{t^*}(\Phi, U)$ *and* $\mathfrak{T}$.

Proof. Let $x^* = x(t^*, \varphi^*, u^*)$. By Theorem 3.2 and Lemma 4.1 we infer that x^* belongs to the boundary of $\mathscr{A}_{t^*}$. There is a vector $e \in R^n$ with $|e| = 1$ such that

$$\max\{\langle e, x\rangle | x \in \mathscr{A}_{t^*}\} = \langle e, x^*\rangle.$$

Using the fact that the value of $\varphi(0)$ does not affect the value of the second term on the right-hand side of (2.7) (cf. proof of Lemma 3.4) and some elementary reasoning involving formula (2.7), we can show that

$$\langle \varphi^*(0), eY'(0,t^*)\rangle \geqq \langle \varphi_0, eY'(0,t^*)\rangle, \qquad \varphi_0 \in \Phi(0); \tag{4.1a}$$

$$\left\langle \int_{-\tau}^{0} \varphi^*(s)\, d_s\left[\int_0^{\tau} F(\alpha, s-\alpha)Y(\alpha,t^*)\, d\alpha\right], e\right\rangle \geqq \left\langle \int_{-\tau}^{0} \varphi(s)\, d_s\left[\int_0^{\tau} F(\alpha, s-\alpha)Y(\alpha,t^*)\, d\alpha\right], e\right\rangle \tag{4.1b}$$

for every admissible φ;

$$\left\langle \int_0^{t^*}\left[\int_{-\tau}^{t^*} h(\mu^*(s), s)\, d_s G(\alpha, s)\right] Y(\alpha, t^*)\, d\alpha, e\right\rangle \geqq \left\langle \int_0^{t^*}\left[\int_{-\tau}^{t^*} h(u(s), s)\, d_s G(\alpha, s)\right] Y(\alpha, t^*)\, d\alpha, e\right\rangle \tag{4.1c}$$

for every admissible u. Define $\psi(\alpha) = eY'(\alpha, t^*)$, $0 \leqq \alpha \leqq t^*$. Then by appropriately using the unsymmetric Fubini theorem [9] and some standard manipulation with the scalar product in (4.1b), (4.1c), relations (ii) and (iii) are proved. The fact that $\alpha \to \psi(\alpha)$, $0 \leqq \alpha \leqq t^*$, is of bounded variation and satisfies the adjoint equation is an immediate consequence of (2.8).

The transversality condition is just a geometric property. In proving this condition we use the norm in R^n defined by $|x|^2 = \langle x, x\rangle$. We have $\mathscr{A}_t \cap \mathfrak{T} = \varnothing$ for $0 \leqq t < t^*$. Let $t_n \in [0, t^*)$ be such that $t_n \to t^*$ as $n \to \infty$. Let $a_n \in \mathscr{A}_{t_n}$, $b_n \in \mathfrak{T}$ be such that $|a_n - b_n|$ is the minimum value that the function $(x, y) \to |x - y|$, $(x, y) \in \mathscr{A}_{t_n} \times \mathfrak{T}$, assumes. Then $a_n - b_n \neq 0$ and $e_n = (b_n - a_n)/|a_n - b_n|$ is a unit outer normal to $\mathscr{A}_{t_n}$ at a_n and a unit inner normal to $\mathfrak{T}$ at b_n. Hence

$$\begin{aligned} &\{x \mid \langle e_n, x - a_n\rangle \leqq 0\} \supset \mathscr{A}_{t_n}, \\ &\{x \mid \langle e_n, x - b_n\rangle \geqq 0\} \supset \mathfrak{T}, \qquad n = 1, 2, 3, \cdots. \end{aligned} \tag{4.2}$$

We might as well assume $e_n \to e$ and $b_n \to b$ as $n \to \infty$. Then a_n also converges to b. Using (4.2) and the fact that $\mathscr{A}_{t_n} \to \mathscr{A}_{t^*}$ as $n \to \infty$ (the limit is taken with respect to the Hausdorff metric) we find that

$$\{x \mid \langle e, x - b\rangle \leqq 0\} \supset \mathscr{A}_{t^*}, \qquad \{x \mid \langle e, x - b\rangle \geqq 0\} \supset \mathfrak{T}$$

so that $\pi = \{x \mid \langle e, x - b\rangle = 0\}$ is a hyperplane satisfying the transversality condition.

Remark 4.1. We can put conditions (ii) and (iii) of Theorem 4.1 in a form which will in many cases be more manageable if we assume that $F(t, \cdot)$ and $G(t, \cdot)$ have no singular part and if the functions θ_i, ζ_i, $i = 1, 2, 3, \cdots$, introduced in properties (Δ_1) and (Δ_2) are of class C^1. Let us indicate the form which (ii) and (iii) take in this case. We use the decompositions $F = A_F + N_F$ and $G = A_G + N_G$ which were introduced in the proof of Theorem 3.2. According to our assumptions, $N_F(t, \cdot)$ and $N_G(t, \cdot)$ are absolutely continuous. By some rather involved analysis, which includes several applications of the unsymmetric Fubini theorem [9], it

can be shown that condition (ii) of Theorem 4.1 implies

(ii′) $\langle \varphi^*(s), P(s)\rangle \geqq \langle \varphi(s), P(s)\rangle$ a.e. on $[-\tau, 0]$

for every admissible φ, where P is defined by

$$P(s) \equiv \psi(s+\tau)B'_{-1}(s+\tau) + \sum_{i=1}^{\infty} \psi(\rho_i^{-1}(s))B'_i(\rho_i^{-1}(s))K_i(s)v_i(s) + \int_0^{\tau} \psi(\alpha)\frac{\partial N'_F}{\partial s}(\alpha, s-\alpha)\,d\alpha, \qquad s \in [-\tau, 0],$$

and $\rho_i(\alpha) \equiv \alpha - \theta_i(\alpha)$ for $\alpha \in [0, \tau]$, K_i denotes the characteristic function of $[\rho_i(0), \rho_i(\tau)] \cap [-\tau, 0]$, and $v_i(s) \equiv 1/\dot{\rho}_i(\rho_i^{-1}(s))$ for $s \in [-\tau, 0]$, $i = 1, 2, 3, \cdots$. By a similar type of analysis which is again omitted, (iii) can be shown to imply

(iii′) $\langle h(u^*(\xi), \xi), Q(\xi)\rangle \geqq \langle h(u(\xi), \xi), Q(\xi)\rangle$ a.e. on $[-\tau, t^*]$

for every admissible u, where Q is defined by

$$Q(\xi) \equiv \psi(\xi)C'_0(\xi)\mathscr{K}_0(\xi) + \sum_{i=1}^{\infty} \psi(\zeta_i^{-1}(\xi))C'_i(\zeta_i^{-1}(\xi))\mathscr{K}_i(\xi)\beta_i(\xi) + \int_0^{t^*} \psi(\alpha)\frac{\partial N'_G}{\partial \xi}(\alpha, \xi)\,d\alpha$$

and $\mathscr{K}_0$ is the characteristic function of $[0, t^*]$, $\mathscr{K}_i$ is the characteristic function of $[\zeta_i(0), \zeta_i(t^*)]$, and $\beta_i(\xi) \equiv 1/\dot{\zeta}_i(\zeta_i^{-1}(\xi))$, $-\tau \leqq \xi \leqq t^*$.

5. Analyticity results for solutions of FDE's. As the reader is by now well aware, the representation (2.7) of solutions to (2.3) in terms of "fundamental" or "adjoint" matrix solutions [3] is of immense importance in the study of control of such systems. In this section we investigate analyticity properties of these fundamental matrix solutions for certain types of linear (in the state variable) systems which are special cases of (2.3); namely,

$$\dot{x}(t) = \sum_{i=0}^{K} x(t-\theta_i)A_i(t) + \int_{-\tau}^{0} x(t+s)A(t,s)\,ds + h(u(t), t) \tag{5.1}$$

with $0 = \theta_0 < \theta_1 < \cdots < \theta_K \leqq \tau$, which corresponds to an $F(t, \cdot)$ consisting of an absolutely continuous function plus a saltus function with a finite number of constant (in t) jump points. A somewhat more general system allowing retardations in the control variable u is discussed further in Remark 6.1 at the end of § 6. The associated fundamental matrices $X(t, \sigma)$ satisfy (as a function of t)

$$\begin{aligned} &\dot{X}(t) = \sum_{i=0}^{K} X(t-\theta_i)A_i(t) + \int_{-\tau}^{0} X(t+s)A(t,s)\,ds, \qquad t > \sigma, \\ &X(\sigma) = E, \qquad X(t) = 0 \quad \text{for } t < \sigma. \end{aligned} \tag{5.2}$$

Since the corresponding adjoint matrices $Y(\sigma, t)$ satisfy (in σ) systems [20] which can be put in a form similar to that of (5.2) and since we have $X(t, \sigma) = Y(\sigma, t)$, to investigate analyticity properties of X and Y in σ or t it suffices to examine the analyticity in t of solutions to (5.2). Considering the following two examples one

sees that systems of the type (5.2) with analytic coefficients and analytic initial function need not possess an analytic solution.

Example 5.1. The scalar system

$$\dot{x}(t) = x(t-1), \qquad t > 0,$$
$$x(t) = 1, \qquad t \in [-1, 0],$$

has a unique solution on $[-1, 2]$ given by

$$x(t) = \begin{cases} 1, & t \in [-1, 0], \\ 1 + t, & t \in [0, 1], \\ 3/2 + t^2/2, & t \in [1, 2], \end{cases}$$

which is not analytic at $t = 1$.

Example 5.2. The scalar system

$$\dot{x}(t) = \int_{-1}^{0} x(t+s)\,ds, \qquad t > 0,$$
$$x(t) = 1, \qquad t \in [-1, 0],$$

has a unique solution on $[-1, 2]$ given by

$$x(t) = \begin{cases} 1, & t \in [-1, 0], \\ 1 + \sinh t, & t \in [0, 1], \\ 1 + \sinh t - \sum_{n=1}^{\infty} (n-1)\dfrac{(t-1)^{2n-1}}{(2n-1)!}, & t \in [1, 2], \end{cases}$$

which is not analytic at $t = 1$.

Remark 5.1. Example 5.1 can be used to contradict a theorem of Pinney [49, p. 237], while Example 5.2 contradicts a result due to Oguztöreli [48, p. 52]. (It is not difficult to show that the right side of the system in Example 5.2 is analytic in x in the sense of Volterra [51], [48] as required in Oguztöreli's theorem.)

In light of the previous examples and remarks one might expect under reasonable assumptions on the coefficients to obtain not analyticity but some type of piecewise analyticity for solutions to (5.2). We are thus motivated to introduce the following concepts (see also Halkin [23] and Levinson [41]). A function $f: R \to R$ is *analytic on* $[a, b]$ if there exist $\varepsilon > 0$ and a function g analytic on $(a - \varepsilon, b + \varepsilon)$ such that $f = g$ on $[a, b]$. We say that f is *piecewise analytic* (pwa) on $[a, b]$ if there exists a partition $a = s_0 < s_1 < \cdots < s_\nu = b$ such that f is analytic on $[s_{i-1}, s_i]$, $i = 1, 2, \cdots, \nu$. Finally, f is said to be *quasi-piecewise analytic* (qpwa) on $[a, b]$ if there exists a partition $a = s_0 < s_1 < \cdots < s_\nu = b$ such that f is analytic on (s_{i-1}, s_i), $i = 1, 2, \cdots, \nu$.

Combining a modification of the step method [18] with known results for ordinary linear differential equations we can prove the following theorem.

THEOREM 5.1. *Let* $A(t, s) \equiv 0$ *in* (5.2) *and* $t \to A_i(t)$, $i = 0, 1, \cdots, K$, *be* (*real*) *analytic on* $[\sigma, \infty]$ *into* $\mathscr{L}_{nn}$. *If the lags* θ_i, $i = 1, 2, \cdots, K$, *are commensurate, then the solution to* (5.2) *is pwa on* $[\sigma, \sigma + T]$ *for any* $T > 0$.

Proof. We shall only give the proof for $K = 1$, $\theta_1 = 1$. In the case of a finite number of commensurate lags (i.e., there exist $\delta > 0$ and positive integers q_i such that $\theta_i = q_i\delta$, $i = 1, 2, \cdots, K$) one uses the following arguments on intervals of length δ instead of intervals of length 1. Thus we consider the system

$$\begin{aligned} &\dot{X}(t) = X(t)A_0(t) + X(t-1)A_1(t), \qquad t > \sigma, \\ &X(\sigma) = E, \\ &X(t) = 0, \qquad t < \sigma, \end{aligned} \tag{5.3}$$

and denote by $\mathfrak{W}$ the solution of

$$\begin{aligned} \dot{\mathfrak{W}}(t) &= \mathfrak{W}(t)A_0(t), \\ \mathfrak{W}(\sigma) &= E. \end{aligned}$$

From the theory of ordinary differential equations it is known that $\mathfrak{W}$ and $\mathfrak{W}^{-1}$ exist and are analytic on $(\sigma - \varepsilon, \sigma + T + \varepsilon)$ for some $\varepsilon > 0$. Since the solution X of (5.3) agrees with $\mathfrak{W}$ on $[\sigma, \sigma + 1]$, we have that X is analytic on $[\sigma, \sigma + 1]$. Furthermore, we see that

$$X(t) = \left\{X(\sigma + n)\mathfrak{W}^{-1}(\sigma + n) + \int_{\sigma+n}^{t} X(s-1)A_1(s)\mathfrak{W}^{-1}(s)\,ds\right\}\mathfrak{W}(t) \tag{5.4}$$

for $t \in [\sigma + n, \sigma + n + 1]$, $n \geqq 1$. Hence the analyticity of $\mathfrak{W}$, $\mathfrak{W}^{-1}$, A_1 on $[\sigma + 1, \sigma + 2]$, and that of X on $[\sigma, \sigma + 1]$ imply that X is analytic on $[\sigma + 1, \sigma + 2]$, which by the same reasoning leads to the analyticity of X on $[\sigma + 2, \sigma + 3]$. A finite number of repetitions of this reasoning using (5.4) completes the proof.

Remark 5.2. From the definition of the determinant it follows immediately that if $[\alpha, \beta]$ is any interval of analyticity of X (the solution to (5.2) with $A(t, s) \equiv 0$), then either $X(t)$ is singular for every $t \in [\alpha, \beta]$ or else there are at most a finite number of points in $[\alpha, \beta]$ where $X^{-1}(t)$ fails to exist.

Just as the step method fails in existence proofs for (5.2) whenever $A(t, s) \not\equiv 0$, this form of the step method will not be of use in proving analyticity results for solutions to the general system (5.2). We can, however, obtain the following result by utilization of successive approximations with step-like procedures.

THEOREM 5.2. *Suppose that* $(t, s) \to A(t, s)$ *and* $t \to A_i(t)$, $i = 0, 1, \cdots, K$, *are (real) analytic on* $[\sigma, \infty) \times [-\tau, 0]$ *and* $[\sigma, \infty)$ *respectively into* $\mathscr{L}_{nn}$. *If the lags* $\theta_1, \theta_2, \cdots, \theta_K, \tau$ *are commensurate, then the solution to* (5.2) *is qpwa on* $[\sigma, \sigma + T]$ *for any* $T > 0$.

Again, we shall here give a proof of this theorem only for the special case

$$\begin{aligned} &\dot{X}(t) = X(t)A_0(t) + X(t-1)A_1(t) + \int_{-1}^{0} X(t+s)A(t,s)\,ds, \qquad t \in [0, T], \\ &X(0) = E, \qquad X(t) = 0 \quad \text{for } t < 0, \end{aligned} \tag{5.5}$$

as it will then be clearly seen how one modifies the ideas to obtain the result for commensurate lags on $[\sigma, \sigma + T]$. Since the uniform limit of a sequence of real analytic functions need not be analytic, if we wish to use successive approximation techniques to obtain analyticity results, then we must work with complex systems. That is, we must somehow replace (5.5) by a system defined on a domain in the

complex plane $\mathbb{C}$ which contains $[-1, T]$ so that the system is equivalent to (5.5) on $[-1, T]$. Before beginning the proof we give some preliminary results which will be needed.

LEMMA 5.1. *If f is analytic in a region $\mathscr{S}(a,b) = \{z = x + iy | a < x < b, -d < y < d\}$ and continuous at $z = a$ from within $\mathscr{S}(a,b)$, then F defined by $F(z) \equiv \int_a^z f(\zeta)\, d\zeta$ is an analytic function on $\mathscr{S}(a,b)$.*

Proof. From the extended form of Cauchy's theorem [52] it follows that F is independent of path in $\mathscr{S}(a,b)$ and is thus well-defined. For $z \in \mathscr{S}(a,b)$ in a neighborhood of $z_0 \in \mathscr{S}(a,b)$ we have $F(z) = F(z_0) + \int_{z_0}^z f(\zeta)\, d\zeta$ which is analytic at z_0 by well-known results.

LEMMA 5.2. *Suppose $(t,s) \to \alpha(t,s)$ is real analytic on $(-\varepsilon, T+\varepsilon) \times (-1-\varepsilon, \varepsilon) \subset R^2$. Then there are sets $\mathscr{T}_\delta$ and $\mathscr{S}_\delta$ in $\mathbb{C}$ of the form*

$$\mathscr{T}_\delta = \{z = x + iy | x \in (-\delta, T+\delta), y \in (-\delta, \delta)\},$$

$$\mathscr{S}_\delta = \{z = x + iy | x \in (-1-\delta, \delta), y \in (-\delta, \delta)\}$$

and a function $(z,w) \to \alpha^*(z,w)$ analytic on $\mathscr{T}_\delta \times \mathscr{S}_\delta \subset \mathbb{C}^2$ such that $\alpha^*|K = \alpha$, where $K \equiv \{(x,0) | x \in [0,T]\} \times \{(x,0) | x \in [-1,0]\}$ (or $K = [0,T] \times [-1,0]$ as a subset of R^2).

Proof. Define $\mathscr{D} = (-\varepsilon, T+\varepsilon) \times (1-\varepsilon, \varepsilon)$, which is an open region in R^2 on which α is analytic. It then follows [45, p. 5], [27, pp. 41–42] that there is an open set $\mathscr{D}^*$ in $\mathbb{C}^2$ with $\mathscr{D}^* \cap R^2 = \mathscr{D}$ and an analytic function α^* on $\mathscr{D}^*$ such that $\alpha^*|\mathscr{D} = \alpha$. The set K defined in the lemma is compact in $\mathbb{C}^2$ and $K \cap (\mathbb{C}^2 \setminus \mathscr{D}^*)$ is empty since $K \cap R^2 \subset \mathscr{D}$. It is then not difficult to show that there is a $\delta > 0$ such that $\mathscr{T}_\delta$ and $\mathscr{S}_\delta$ as defined in the lemma satisfy $K \subset \mathscr{T}_\delta \times \mathscr{S}_\delta \subset \mathscr{D}^*$.

Proof of Theorem 5.2. Our first task is to somehow extend system (5.5) (or, as in the usual case of successive approximations, its equivalent in integral form) to a system on a complex domain where of course we want all coefficients involved to be analytic. From Lemma 5.2 and standard arguments it follows that there exist domains $\mathscr{S}_\delta$, $\mathscr{T}_\delta$ (see Lemma 5.2) and analytic continuations (which we again denote by A and A_i) of the mappings $(t,s) \to A(t,s)$ and $t \to A_i(t)$ to $\mathscr{S}_\delta \times \mathscr{S}_\delta$ and $\mathscr{T}_\delta$ respectively. Let $\mathscr{D}_\mathscr{S}$ and $\mathscr{D}_\mathscr{T}$ be $\delta/2$ neighborhoods in $\mathbb{C}$ (using the usual norm in $\mathbb{C}$) of the sets $[-1,0]$ and $[0,T]$ respectively. These are the regions on which we shall work throughout the remainder of the proof.

For k any integer, we define $S_k = \{z \in \mathbb{C} | k < \operatorname{Re}(z) < k+1\} \cap (\mathscr{D}_\mathscr{S} \cup \mathscr{D}_\mathscr{T})$. We shall consider the system defined for $z \in S_k \cup \{k+1\}$, $k \geqq 0$, by

$$X(z) = E + \int_{[0,z]} \Big\{ X(\zeta) A_0(\zeta) + X(\zeta - 1) A_1(\zeta) + \int_{[-1,0]} X(\zeta + w) A(\zeta, w)\, dw \Big\} d\zeta, \tag{5.6}$$

where we must indicate the paths of integration to be used. The path $[0,z]$ for $z \in S_k \cup \{k+1\}$ consists of straight-line segments joining z and k, k and $z-1$, $z-1$ and $k-1, \cdots, z-(k-1)$ and 1, 1 and $z-k$, $z-k$ and 0. Note that the

path will always lie in $\mathscr{D}_{\mathscr{F}}$. The integral $\int_{[-1,0]} X(\zeta + w)A(\zeta, w)\,dw$, for ζ on the polygonal path joining 0 and z described above and $\zeta \in S_m \cup \{m+1\}$, is to be integrated along the w-path of straight-line segments joining 0 and $-\zeta + m$, $-\zeta + m$ and -1. Hence for any $\zeta \in S_m \cup \{m+1\}$ this latter integral depends on the values of X along the segments joining ζ and m, m and $\zeta - 1$. Note that for z real the system (5.6) with the proper initial conditions reduces to the integrated form (in the usual sense) of (5.5).

We next obtain a quasi-slabwise analytic (i.e., analytic on S_k, $k = 0, 1, \cdots$) solution to (5.6). To do this we define successive approximations X_n, show that each is analytic on S_k, $k = 0, 1, \cdots$, and show that $\{X_n\}$ converges uniformly on each S_k. The limit function will be the desired solution. Define for $n = 0, 1, 2, \cdots$, $X_n(z) = 0$ for $\operatorname{Re}(z) < 0$ and $X_n(0) = E$. For $z \in S_k \cup \{k+1\}$, $k = 0, 1, 2, \cdots$, define $X_0(z) = E$ and

$$X_n(z) = E + \int_{[0,z]} \left\{ X_{n-1}(\zeta)A_0(\zeta) + X_{n-1}(\zeta - 1)A_1(\zeta) + \int_{[-1,0]} X_{n-1}(\zeta + w)A(\zeta, w)\,dw \right\} d\zeta \tag{5.7}$$

for $n = 1, 2, \cdots$, where the paths of integration are the polygonal paths described above. Note that each X_n, $n \geqq 1$, is defined on $\mathscr{D}_{\mathscr{G}} \cup \mathscr{D}_{\mathscr{F}}$ less the rays $\cup_{k \geqq 0} \{z = k + iy | y \neq 0\}$.

We shall say that a function g is left continuous at $z = k$, $k \geqq 0$, if $g(\zeta) \to g(k)$ as $\zeta \to k$, $\zeta \in S_{k-1}$. A similar meaning is attached to "right continuous at $z = k$." Finally, we shall say that g is continuous at $z = k$ if it is both left and right continuous at $z = k$ in the above sense. We now state and prove an induction lemma which will yield analyticity of X_n on the S_k.

INDUCTION LEMMA. *Let $n \geqq 1$. Let $k \geqq 0$. Then X_{n-1} analytic on $S_{-1}, S_0, S_1, \cdots, S_k$ and continuous at $z = 0, 1, 2, \cdots, k$ imply X_n analytic on S_k and continuous at $z = k$.*

Note. Since clearly none of the approximations are left continuous at $z = 0$, we understand "continuous at $z = 0$" to mean "right continuous at $z = 0$" in the above lemma.

Proof. Suppose the assumptions of the induction lemma are true. We can then establish the following lemma.

LEMMA 5.3. *Let m be a fixed integer, $0 \leqq m \leqq k$. For $\zeta \in S_m \cup \{m\} \cup \{m+1\}$ define*

$$F(\zeta) = \int_{[-1,0]} X_{n-1}(\zeta + w)A(\zeta, w)\,dw.$$

Then F is analytic on S_m, right continuous at $z = m$, and, if $m < k$, left continuous at $z = m + 1$.

Use of the hypotheses of the induction lemma and Lemma 5.3 yield that the integrand

$$\mathscr{I}(\zeta) = X_{n-1}(\zeta)A_0(\zeta) + X_{n-1}(\zeta - 1)A_1(\zeta) + \int_{[-1,0]} X_{n-1}(\zeta + w)A(\zeta, w)\,dw$$

in (5.7) is analytic on $S_0, S_1, \cdots, S_{k-1}$ and continuous at $z = 0, 1, \cdots, k-1$, left continuous at $z = k$. Hence by the extension of Cauchy's theorem (see the proof of Lemma 5.1) the part of the integral in (5.7) from 0 to k along the polygonal paths is actually independent of path (as long as the paths cross the lines $\mathrm{Re}\,(z) = m$ through the point $z = m$). Thus, (5.7) may be written

$$X_n(z) = E + \int_0^k \mathscr{I}(\zeta)\,d\zeta + \int_k^z \mathscr{I}(\zeta)\,d\zeta, \tag{5.8}$$

where as usual $\int_{z_1}^{z_2}$ denotes integration along the straight-line segment joining z_1 and z_2. The first integral in (5.8) is now independent of $z \in S_k$. Thus we need only show that the second integral is analytic for $z \in S_k$. But this follows immediately from the hypotheses of the induction lemma, Lemma 5.3 and Lemma 5.1. We therefore have X_n analytic on S_k.

We shall next argue that X_n is right continuous at $z = k$; the arguments for left continuity are not dissimilar and will be omitted. From (5.8) we have $X_n(z) - X_n(k) = \int_k^z \mathscr{I}(\zeta)\,d\zeta$ for $z \in S_k$, the integrand $\mathscr{I}$ being analytic on S_k and right continuous at $z = k$. Thus $\mathscr{I}$ is bounded in some "right neighborhood" of $z = k$, from which the desired result follows immediately. To complete the proof of the induction lemma it remains only to establish the validity of Lemma 5.3.

Proof of Lemma 5.3. Making the assumptions given in the statement of the induction lemma, we let m be a fixed integer, $0 \leqq m \leqq k$. Then $F(\zeta), \zeta \in S_m$, can be written

$$\begin{aligned} F(\zeta) &= \int_{-1}^{-\zeta+m} X_{n-1}(\zeta + w)A(\zeta, w)\,dw + \int_{-\zeta+m}^{0} X_{n-1}(\zeta + w)A(\zeta, w)\,dw \\ &= \int_{\zeta-1}^{m} X_{n-1}(w)A(\zeta, w - \zeta)\,dw + \int_{m}^{\zeta} X_{n-1}(w)A(\zeta, w - \zeta)\,dw. \end{aligned}$$

The right continuity of F at $z = m$ follows from the continuity of A, the boundedness of X_{n-1} in right and left neighborhoods of $z = m$ and a right neighborhood of $z = m - 1$, and the theorem on dominated convergence. For $m < k$ the proof that F is left continuous at $z = m + 1$ is similar. (If $m = k$, these arguments are no longer valid in obtaining left continuity of F at $m + 1$ since at this stage in the induction we do not have that X_{n-1} is left continuous at $k + 1$, which is needed for the boundedness conclusions about X_{n-1}.)

We turn next to the analyticity arguments for F on S_m. We shall argue that the function f defined by $f(\zeta) = \int_m^\zeta X_{n-1}(w)A(\zeta, w - \zeta)\,dw$ is analytic on S_m, similar arguments being valid for the term $\int_{\zeta-1}^{m} X_{n-1}(w)A(\zeta, w - \zeta)\,dw$ in F above.

Fix $\zeta_0 \in S_m$. For ζ in a sufficiently small neighborhood $\mathscr{N}_0$ of ζ_0 we can write

$$\begin{aligned} \int_m^\zeta X_{n-1}(w)A(\zeta, w - \zeta)\,dw &= \int_m^{\zeta_0} X_{n-1}(w)A(\zeta, w - \zeta)\,dw \\ &\quad + \int_{\zeta_0}^{\zeta} X_{n-1}(w)A(\zeta, w - \zeta)\,dw \equiv h_1(\zeta) + h_2(\zeta), \end{aligned}$$

where the integrands are analytic in ζ on $\mathcal{N}_0$ for each fixed $w \in \mathcal{N}_0$ and analytic in w on $\mathcal{N}_0$ for each fixed $\zeta \in \mathcal{N}_0$. A straightforward application of Morera's theorem establishes the analyticity of h_1 on $\mathcal{N}_0$. Use of a theorem of Hartogs–Osgood [28, p. 28] yields that h_2 is of the form $\int_{\zeta_0}^{\zeta} g(w, \zeta)\, dw$, where g is analytic in $\mathcal{N}_0 \times \mathcal{N}_0$, from which the analyticity of h_2 follows easily.

Having confirmed the validity of the induction lemma, we point out that it follows directly from the analyticity properties of X_0 (recall $X_0(z) = E$ for $z \in S_k \cup \{k+1\}$, $k = 0, 1, \cdots$ and $X_0(z) = 0$ for $\operatorname{Re}(z) < 0$) and the induction lemma that each X_n is analytic on each S_k.

We next prove that the sequence $\{X_n\}$ converges uniformly on the region of interest. Let l be the positive integer ($l > T$) such that $S_k = \varnothing$ for $k \geqq l$ and $S_{l-1} \neq \varnothing$. (Recall the definition of S_k, $\mathscr{D}_{\mathscr{G}}$ and $\mathscr{D}_{\mathscr{F}}$.) We shall show that the sequence $\{X_n\}$ converges uniformly on $\mathfrak{R} \equiv \bigcup_{k=0}^{l-1} S_k \cup \{k\}$. We note that we trivially have uniform convergence of $\{X_n\}$ on $S_{-1} \cup \{0\}$ to the function X defined by $X(z) = 0$ for $z \in S_{-1}$, $X(0) = E$. Recall now the definition of X_n given in (5.7) and the integration paths employed. For any $z \in \mathfrak{R}$ let $s(z)$ denote the *arclength* of the polygonal path described above (see (5.6)) which joins 0 to z. Let M be a bound for $|A_0(\zeta)|$, $|A_1(\zeta)|$, $\zeta \in \mathscr{D}_{\mathscr{F}}$, and $|A(\zeta, w)|$, $(\zeta, w) \in \mathscr{D}_{\mathscr{F}} \times \mathscr{D}_{\mathscr{G}}$. Then for $z \in \mathfrak{R}$ we have

$$
\begin{aligned}
|X_1(z) - X_0(z)| &= \left| \int_{[0,z]} \left\{ X_0(\zeta)A_0(\zeta) + X_0(\zeta - 1)A_1(\zeta) \right.\right. \\
&\qquad \left.\left. + \int_{[-1,0]} X_0(\zeta + w)A(\zeta, w)\, dw \right\} d\zeta \right| \\
&\leqq \int_{[0,z]} \left\{ |A_0(\zeta)| + |A_1(\zeta)| + \int_{[-1,0]} |A(\zeta, w)|\, |dw| \right\} |d\zeta| \\
&\leqq \int_{[0,z]} \left\{ M + M + M\left(1 + 2\left(\frac{\delta}{2}\right)\right) \right\} |d\zeta| \\
&\leqq 3M(1 + \delta)s(z) \equiv \rho s(z).
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
|X_2(z) - X_1(z)| \leqq \int_{[0,z]} \Big\{ & M|X_1(\zeta) - X_0(\zeta)| + M|X_1(\zeta - 1) - X_0(\zeta - 1)| \\
& + \int_{[-1,0]} M|X_1(\zeta + w) - X_0(\zeta + w)|\, |dw| \Big\} |d\zeta|.
\end{aligned}
$$

For $\zeta \in S_0$ we have

$$
\begin{aligned}
\int_{[-1,0]} M|X_1(\zeta + w) - X_0(\zeta + w)|\, |dw| &= \int_{-\zeta}^{0} M|X_1(\zeta + w) - X_0(\zeta + w)|\, |dw| \\
&\leqq \int_{-\zeta}^{0} M\rho s(\zeta + w)|dw| \leqq \int_{-\zeta}^{0} M\rho s(\zeta)|dw| \\
&\leqq M\rho s(\zeta)(1 + \delta).
\end{aligned}
$$

For $\zeta \in S_k$, $k \geqq 1$, we find that

$$s(\zeta + w) \leqq s(\zeta) \tag{5.9}$$

for any w lying on the path consisting of straight-line segments joining -1 to $-\zeta + k$ and $-\zeta + k$ to 0. Hence the above estimate is also valid for these values of ζ. It follows that

$$\begin{aligned} |X_2(z) - X_1(z)| &\leqq \int_{[0,z]} \{M\rho s(\zeta) + M\rho s(\zeta) + M\rho s(\zeta)(1 + \delta)\}|d\zeta| \\ &\leqq 3M(1 + \delta)\rho \int_{[0,z]} s(\zeta)|d\zeta| \\ &= \rho^2 \frac{[s(z)]^2}{2}. \end{aligned}$$

Using this estimate and the above ideas it is easily shown that

$$|X_3(z) - X_2(z)| \leqq \rho^3 \frac{[s(z)]^3}{3!},$$

and, in general,

$$|X_n(z) - X_{n-1}(z)| \leqq \rho^n \frac{[s(z)]^n}{n!}$$

for $z \in \mathfrak{R}$. Hence for $n > m$, we have

$$|X_n(z) - X_m(z)| \leqq \sum_{j=m+1}^{n} |X_j(z) - X_{j-1}(z)| \leqq \sum_{j=m+1}^{n} \rho^j \frac{[s(z)]^j}{j!}. \tag{5.10}$$

But for $z \in \mathfrak{R}$ we have that $s(z) \leqq l(1 + 2(\delta/2)) = l(1 + \delta)$. Using this with (5.10) yields the uniform convergence of $\{X_n\}$ on $\mathfrak{R}$. Let us denote by X this limit function on $\mathfrak{R} \cup S_{-1}$. Since each X_n is analytic on S_k and continuous at $z = k$, we have that X also possesses these properties. Furthermore, for each n, $X_n(z)$ is real-valued whenever z is real, from which it follows that X is real analytic on $(0, 1)$, $(1, 2)$, etc. Finally, since X_n converges to X uniformly on $[-1, T]$ it is not difficult to argue that X is the unique solution to (5.5), which completes the proof of Theorem 5.2.

One might reasonably expect a stronger type of analyticity (say pwa) than that obtained in Theorem 5.2 to be true for systems of the type (5.2) even with $A \not\equiv 0$. The authors have tried unsuccessfully so far to obtain these stronger results. Several ideas using different integration paths in defining the successive approximations (see the proof of Theorem 5.2) and stronger assumptions on the coefficients have been tried. These lead to either a lack of analyticity of the estimates in the desired regions, or else an inability to obtain uniform convergence of the estimates. The authors were able to prove that the solution to (5.5) is analytic on $[0, 1]$, but could not adapt these methods to prove analyticity on $[k, k + 1]$ for $k > 0$. The fact that one is using a zero initial matrix on $[-1, 0)$ appears to be essential in obtaining analyticity on $[0, 1]$. (Note that in this case the system loses some of its lag behavior on $[0, 1]$ and is much more like an integral equation.)

The analyticity results obtained in this section can be used to study the zeros of the multipliers in the maximal principle for control problems involving functional differential equations (see Remark 5.2 and [4], [5], [6], [21]). The information thus obtained can be especially useful when the maximal principle is also a sufficient condition for optimality (see [21]). Another application of these analyticity results is discussed in the next section.

6. Application of the analyticity results. We shall use pwc as an abbreviation for *piecewise continuous*, and when we say a function $f:[a, b] \to R^p$ is piecewise continuous we are taking the standard definition. We shall say that f is *almost piecewise continuous* (apwc) if there is a finite number of points $s_i \in [a, b]$, $i = 0, 1, \cdots, N$, with the property that $f|[\alpha, \beta]$ is pwc for $[\alpha, \beta] \subset [a, b]$ for which $s_i \notin [\alpha, \beta]$, $i = 0, 1, \cdots, N$.

In this section we shall demonstrate how some of the work with subintegrals of multifunctions by Halkin and Hendricks [25] and the related existence theory for piecewise continuous optimal controls [24] can be applied in special cases of (2.3) to give analogues of Theorem 3.2 when the admissible triples $\{\varphi, u, t\}$ are required to be pwc (or apwc) (i.e., φ and u are pwc (or apwc)).

Lebesgue measure will be understood in all of the integrals appearing in this section. Suppose a multifunction $H:[a, b] \to R^p$ is given. Then we have defined $\int_a^b H(t)\,dt$ and with $[a, b]$ understood we denote this by $\int H$. We define

$$\int^* H \equiv \left\{ \int_a^b g(t)\,dt \,|\, g:[a, b] \to R^p \text{ is apwc and } g(t) \in H(t), t \in [a, b] \right\}.$$

LEMMA 6.1 (Halkin-Hendricks). $\int^* H$ *is convex.*

We omit the proof. Let it suffice to say that the proof of Theorem 1 [25, p. 365] may in effect be repeated. One need only take f_1 and f_2 to be apwc in that proof and observe that $f_1 \cdot \chi_{[a,b]\setminus E} + f_2 \cdot \chi_E$ is apwc if $E \subset [a, b]$ is the union of a finite number of intervals.

A set $E \subset R^q$ is said to be *semianalytic* (see Lojasiewicz [44] or Halkin and Hendricks [25]) if for every point in R^q there exists a neighborhood V of that point such that

$$E \cap V = \bigcup_{i=1}^{k} \{x \in R^q \,|\, f_i(x) = 0 \text{ and } g_{ij}(x) > 0 \text{ for } j = 1, 2, \cdots, l\},$$

where g_{ij} and f_i are real-valued functions which are analytic on V.

LEMMA 6.2 (Halkin–Hendricks). *Let $H:[a, b] \to R^p$ be a compact multifunction and suppose the graph of H is bounded. Let there exist a finite set of points s_i, $i = 0, 1, \cdots, N$, such that $a \leqq s_0 < s_1 < \cdots < s_N \leqq b$ and such that for each compact interval $[\alpha, \beta] \subset [a, b]$ which contains none of the points s_i the graph of H restricted to $[\alpha, \beta]$ is semianalytic. Then $\int H = \int^* H$.*

Again this is only a slight extension of the main result (Theorem 2) in [25]. Indeed, the proof is clear upon examining the proof of that theorem. In effect one observes that $\int H \supset \int^* H \supset \mathscr{E}\left(\int H\right)$, where $\mathscr{E}\left(\int H\right)$ denotes the set of

extreme points of the convex set $\int H$, and that $\int^* H$ is convex, and then the proof is immediate. To show that $\int^* H \supset \mathscr{E}\left(\int H\right)$ one need only show that the function $g:[a, b] \to R^p$ is apwc, where g is defined by the condition that $g(t)$ is the lexicographic maximum (with respect to an arbitrary orthonormal basis for R^p as in Olech [47]) of $H(t)$, $t \in [a, b]$. If the s_i, $i = 0, 1, \cdots, N$, and an interval $[\alpha, \beta]$ are chosen as in the hypotheses of Lemma 6.2, then Halkin and Hendricks [25] have shown $g|[\alpha, \beta]$ is pwc. Hence $g:[a, b] \to R^p$ is apwc.

LEMMA 6.3 (Halkin). *Let $B:[a, b] \times R^q \to R^p$ be a continuous function with the property that there is a finite set of points s_i, $i = 0, 1, \cdots, N$, such that $a \leqq s_0 < s_1 < \cdots < s_N \leqq b$ and such that $B|(a, s_0) \times R^q$, $B|(s_N, b) \times R^q$ and $B|(s_{i-1}, s_i) \times R^q$, $i = 1, 2, \cdots, N$, are analytic. Define $\mathscr{B}:[a, b] \times R^q \to R^{p+q}$ by the relation $\mathscr{B}(t, u) = (u, B(t, u))$. Let $\Omega:[a, b] \to R^q$ be a compact multifunction satisfying the generic hypotheses of Lemma 6.2. Then the compact multifunction $\mathscr{W}:[a, b] \to R^{p+q}$ defined by $\mathscr{W}(t) = \mathscr{B}(t, \Omega(t))$, $t \in [a, b]$, also satisfies the generic hypotheses of Lemma 6.2.*

This result is a modification of a statement of Halkin's [24]. Since Halkin omitted a proof and since the above lemma differs somewhat from his result, we shall suggest a proof which is straightforward. There will be no loss in generality if we assume that the same points satisfy the hypothesis of Lemma 6.2 with $H:[a, b] \to R^p$ replaced by $\Omega:[a, b] \to R^q$. It will suffice for us to show that if $[\alpha, \beta] \subset [a, b]$ and $s_i \notin [\alpha, \beta]$, $i = 0, 1, \cdots, N$, then $\mathscr{W}|[\alpha, \beta]$ has a semianalytic graph. Let $\mathscr{G}$ denote the graph of $\mathscr{W}|[\alpha, \beta]$ and let $P_0 = (t_0, u_0, x_0)$ be an arbitrary point in $R \times R^q \times R^p = R^{p+q+1}$. Then there is a neighborhood V_0 of (t_0, u_0) in $R \times R^q = R^{q+1}$ and analytic functions f_i, g_{ij}, $i = 1, \cdots, k$, $j = 1, \cdots, l$, on V_0 such that

$$E \cap V_0 = \bigcup_{i=1}^{k} \{(t, u) \in R^{q+1} | f_i(t, u) = 0 \text{ and } g_{ij}(t, u) > 0 \text{ for } j = 1, \cdots, l\},$$

where E is the graph of $\Omega|[\alpha, \beta]$. Let $\mathfrak{B}$ denote the set $\{(t, u, x) \in R^{p+q+1} | (t, u) \in V_0\}$; then $\mathfrak{B}$ is a neighborhood of P_0 in R^{p+q+1}. A generic point (t, u, x) in R^{p+q+1} is also denoted by $(t, u^1, \cdots, u^q, x^1, \cdots, x^p)$, and we write $(B^1, \cdots, B^p)$ for the function B. Let $\pi:R^{p+q+1} \to R^{q+1}$ be defined by $\pi(t, u, x) = (t, u)$. Define functions $\bar{f}_i$ and $\bar{g}_{ij}$ on $\mathfrak{B}$ by the equations

$$\bar{f}_i(t, u, x) = [f_i(\pi(t, u, x))]^2 + \sum_{n=1}^{p} [B^n(t, u) - x^n]^2,$$

$$\bar{g}_{ij}(t, u, x) = g_{ij}(\pi(t, u, x))$$

for $i = 1, \cdots, k$, $j = 1, \cdots, l$. Then we have that there are real numbers $\bar{\alpha}, \bar{\beta}$ such that $(\bar{\alpha}, \bar{\beta}) \supset [\alpha, \beta]$ and such that B is analytic on $(\bar{\alpha}, \bar{\beta}) \times R^q$. One can now verify that

$$\mathscr{G} \cap \mathfrak{B} \cap ((\bar{\alpha}, \bar{\beta}) \times R^{p+q}) = \bigcup_{i=1}^{k} \{(t, u, x) \in R^{p+q+1} | \bar{f}_i(t, u, x) = 0 \text{ and } \bar{g}_{ij}(t, u, x) > 0 \text{ for } j = 1, 2, \cdots, l\},$$

where the functions $\bar{f}_i$ and $\bar{g}_{ij}$ are analytic on $\mathfrak{B} \cap ((\bar{\alpha}, \bar{\beta}) \times R^{p+q})$. We can assume $P_0 \in (\bar{\alpha}, \bar{\beta}) \times R^{p+q}$ since the contrary case can be dealt with trivially. Thus $\mathscr{G}$ is semianalytic and this proves the lemma.

We now turn our attention to the control system (5.1). Let $\mathscr{A}_t^*(\Phi, U)$ denote the collection of all points in $\mathscr{A}_t(\Phi, U)$ which are attainable from admissible triples $\{\varphi, u, t\}$, where $\varphi : [-\tau, 0] \to R^n$ and $u : [0, t] \to R^m$ are apwc. Let $\underline{\mathscr{A}}_t(\Phi, U)$ denote the collection of all points in $\mathscr{A}_t(\Phi, U)$ which are attainable from admissible triples $\{\varphi, u, t\}$ where both φ and u are pwc. The variation of parameters formula[4] (2.7) when applied to the FDE (5.1) gives

$$x(t, \varphi, u) = \varphi(0)Y(0, t) + \int_0^t h(u(\alpha), \alpha)Y(\alpha, t)\, d\alpha + \int_{-\tau}^0 \varphi(\alpha)\mathfrak{R}(\alpha, t)\, d\alpha, \tag{6.1}$$

where $\mathfrak{R}(\alpha, t)$ is defined by the equation

$$\mathfrak{R}(\alpha, t) \equiv \sum_{i=1}^{K} A_i(\alpha + \theta_i)Y(\alpha + \theta_i, t)\chi_{[-\theta_i, 0]}(\alpha) + \int_0^{\alpha+\tau} A(s, \alpha - s)Y(s, t)\, ds, \qquad -\tau \leqq \alpha \leqq 0.$$

Let a function $\mathfrak{M} : [-\tau, 0] \times R^n \to R^{2n}$ be defined by

$$\mathfrak{M}(\alpha, \varphi) \equiv (\varphi, \varphi\mathfrak{R}(\alpha, t)), \qquad -\tau \leqq \alpha \leqq 0, \quad \varphi \in R^n,$$

and let $\mathfrak{H} : R^m \times [0, t] \to R^{m+n}$ be the function defined by

$$\mathfrak{H}(u, \alpha) \equiv (u, h(u, \alpha)Y(\alpha, t)), \qquad 0 \leqq \alpha \leqq t, \quad u \in R^m.$$

Define projections $\pi_1 : R^{m+n} \to R^n$ and $\pi_2 : R^{2n} \to R^n$ by the equations

$$\pi_1(u, x) = x, \qquad (u, x) \in R^m \times R^n = R^{m+n},$$
$$\pi_2(\varphi, x) = x, \qquad (\varphi, x) \in R^n \times R^n = R^{2n}.$$

In each of the following three formulas the first integral on the right-hand side of the equation is over the interval $[0, t]$ and the second integral is over the interval $[-\tau, 0]$. Using (6.1) it can be shown that[5]

$$\begin{aligned}
\mathscr{A}_t(\Phi, U) &= \Phi(0)Y(0, t) + \pi_1\left[\int \mathfrak{H}(U(\alpha), \alpha)\, d\alpha\right] + \pi_2\left[\int \mathfrak{M}(\alpha, \Phi(\alpha))\, d\alpha\right], \\
\mathscr{A}_t^*(\Phi, U) &= \Phi(0)Y(0, t) + \pi_1\left[\int^* \mathfrak{H}(U(\alpha), \alpha)\, d\alpha\right] + \pi_2\left[\int^* \mathfrak{M}(\alpha, \Phi(\alpha))\, d\alpha\right], \\
\underline{\mathscr{A}}_t(\Phi, U) &= \Phi(0)Y(0, t) + \pi_1\left[\underline{\int} \mathfrak{H}(U(\alpha), \alpha)\, d\alpha\right] + \pi_2\left[\underline{\int} \mathfrak{M}(\alpha, \Phi(\alpha))\, d\alpha\right],
\end{aligned} \tag{6.2}$$

whenever the left-hand sides are nonempty.

[4] We remark that the representation theorems can easily be shown to be valid under the analyticity hypotheses placed on (5.1) in § 5 (Theorems 5.1, 5.2).

[5] Here $\underline{\int}$ denotes the subintegral in Halkin and Hendricks [25]; i.e., if $H : [a, b] \to R^p$ is some multifunction, then $\underline{\int} H \equiv \left\{ \int_a^b g(t)\, dt \middle| g : [a, b] \to R^p \text{ is pwc and } g(t) \in H(t) \text{ for } t \in [a, b] \right\}$.

THEOREM 6.1. *Let the homogeneous part of* (5.1) *satisfy the hypotheses of Theorem* 5.2, *and let the functions* $h|R^m \times [0, t]$, $t \geqq 0$, *satisfy the same conditions as the function B in Lemma* 6.3. *Let* $\Phi:[-\tau, 0] \to R^n$ *and* $U:[0, \infty) \to R^m$ *be compact multifunctions satisfying the generic hypotheses of Lemma* 6.2. *Then*

$$\mathscr{A}_t(\Phi, U) = \mathscr{A}_t^*(\Phi, U), \qquad t \geqq 0.$$

Proof. With the aid of Theorem 5.2 and a few rudimentary deductions one can show that the function $\mathfrak{H}:R^m \times [0, t] \to R^{m+n}$ and the multifunction $U:[0, t] \to R^m$ satisfy the generic hypotheses of Lemma 6.3, and similarly for the function $\mathfrak{M}:[-\tau, 0] \times R^n \to R^{2n}$ and the multifunction $\Phi:[-\tau, 0] \to R^n$. Thus Lemmas 6.3 and 6.2 and the second formula in relation (6.2) apply to give the desired conclusion.

THEOREM 6.2. *Let the homogeneous part of* (5.1) *satisfy the hypotheses of Theorem* 5.1 *and let the function h be analytic on* $R^m \times [0, \infty)$. *Let* $\Phi:[-\tau, 0] \to R^n$ *and* $U:[0, \infty) \to R^m$ *be compact multifunctions such that the graph of* Φ *and the graph of* $U|[0, t]$ *for* $t \geqq 0$ *are bounded and semianalytic. Then*

$$\mathscr{A}_t(\Phi, U) = \underline{\mathscr{A}}_t(\Phi, U), \qquad t \geqq 0.$$

With the aid of Theorem 5.1 and the above remarks the proof of this theorem will be so similar to Halkin's proof [24] of the corresponding result for nondelay systems that it can safely be omitted.

Recalling Remark 3.3 one sees that Theorems 6.1 and 6.2 give new existence theorems for certain optimal control problems in the class of apwc admissible triples $\{\varphi, u, t\}$ and the class of pwc admissible triples $\{\varphi, u, t\}$ respectively.

Remark 6.1. Finally, we point out that the conclusions of this section can be extended to include systems with certain types of delays in the controls. In particular, results similar to those given in Theorems 6.1 and 6.2 can be obtained for systems of the form

$$\begin{aligned} \dot{x}(t) = &\sum_{i=0}^{K} x(t - \theta_i)A_i(t) + \int_{-\tau}^{0} x(t + s)A(t, s)\, ds \\ &+ \sum_{i=0}^{v} h(u(t - h_i), t - h_i)B_i(t) + \int_{-\tau}^{t} h(u(s), s)B(t, s)\, ds, \end{aligned} \tag{6.3}$$

$0 = h_0 < h_1 < \cdots < h_v \leqq \tau$, which corresponds to a $G(t, \cdot)$ in (2.3) consisting of an absolutely continuous function plus a saltus function with a finite number of jump points. The statements and proofs of these results are so similar to those above that we shall omit them here.

We note that these results can be obtained under hypotheses on (6.3) so as to include as special cases systems of the type

$$\dot{x}(t) = \sum_{i=0}^{K} x(t - \theta_i)A_i(t) + \int_{-\tau}^{0} x(t + s)A(t, s)\, ds + \sum_{i=0}^{v} u(t - h_i)B_i(t).$$

REFERENCES

[1] P. ALEXANDROFF AND H. HOPF, *Topologie*, vol. 1, Springer-Verlag, Berlin, 1935.

[2] R. J. AUMANN, *Integrals of set-valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1–12.

[3] H. T. BANKS, *Representations for solutions of linear functional differential equations*, J. Differential Equations, 5 (1969), pp. 399–409.

[4] H. T. BANKS, *Variational problems involving functional differential equations*, this Journal, 7 (1969), pp. 1–17.
[5] ———, *A maximum principle for optimal control problems with functional differential systems*, Bull. Amer. Math. Soc., 75 (1969), pp. 158–161.
[6] ———, *Necessary conditions for control problems with variable time lags*, this Journal, 6 (1968), pp. 9–47.
[7] R. BELLMAN AND K. L. COOKE, *Differential Difference Equations*, Academic Press, New York, 1963.
[8] D. BLACKWELL, *The range of certain vector integrals*, Proc. Amer. Math. Soc., 2 (1951), pp. 390–395.
[9] R. H. CAMERON AND W. T. MARTIN, *An unsymmetric Fubini theorem*, Bull. Amer. Math. Soc., 47 (1941), pp. 121–125.
[10] C. CASTAING, *Sur les multi-applications mesurables*, Doctoral thesis, Université de Caen, France, 1967.
[11] ———, *Sur les multi applications mesurables*, Rev. Française Inf. Rech. Oper., 1 (1967), pp. 91–126.
[12] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints. I*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412.
[13] D. H. CHYUNG AND E. B. LEE, *Optimal systems with time delays*, Proc. 3rd IFAC Conf., Institute of Mechanical Engineers, London, 1966.
[14] K. S. DAY AND T. C. HSAI, *Optimal control of linear time-lag systems*, Proc. JACC (Ann Arbor, Mich., 1968), American Society of Mechanical Engineers, New York, 1968, pp. 1046–1055.
[15] G. DEBREU, *Integration of correspondences*, Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1966, pp. 351–372.
[16] N. DINCULEANU, *Vector Measures*, Pergamon Press, New York, 1967.
[17] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part I*, Interscience, New York, 1958.
[18] L. E. EL'SGOL'TS, *Introduction to the Theory of Differential Equations with Deviating Arguments*, Holden-Day, San Francisco, 1966.
[19] W. C. GRIMMELL, *The existence of piecewise-continuous fuel optimal controls*, this Journal, 5 (1967), pp. 515–519.
[20] A. HALANAY, *Differential Equations*, Academic Press, New York, 1966.
[21] ———, *Le "principe du maximum" pour les systèmes optimaux linéares à retardement*, C. R. Acad. Sci. Paris, 254 (1962), pp. 2277–2279.
[22] ———, *Optimal controls for systems with time lag*, this Journal, 6 (1968), pp. 215–234.
[23] H. HALKIN, *A generalization of LaSalle's "bang-bang" principle*, this Journal, 2 (1964), pp. 199–202.
[24] ———, *A new existence theorem in the class of piecewise continuous control functions*, Calculus of Variations and Control Theory, A. V. Balakrishnan, ed., Academic Press, New York, 1969.
[25] H. HALKIN AND E. C. HENDRICKS, *Subintegrals of set-valued functions with semianalytic graphs*, Proc. Nat. Acad. Sci. U.S.A., 59 (1968), pp. 365–367.
[26] H. HERMES, *Calculus of set-valued functions and control*, J. Math. Mech., 18 (1968), pp. 47–60.
[27] M. HERVÉ, *Several Complex Variables. Local Theory*, Oxford University Press, London, 1963.
[28] L. HÖRMANDER, *An Introduction to Complex Analysis in Several Variables*, Van Nostrand, New York, 1966.
[29] C. J. HIMMELBERG, M. Q. JACOBS AND F. S. VANVLECK, *Measurable multifunctions, selectors, and Filippov's implicit functions lemma*, J. Math. Anal. Appl., 25 (1969), pp. 276–284.
[30] D. K. HUGHES, *Variational and optimal control problems with delayed argument*, J. Optimization Theory and Applications, 2 (1968), pp. 1–14.
[31] M. Q. JACOBS, *Measurable multivalued mappings and Lusin's theorem*, Trans. Amer. Math. Soc., 134 (1968), pp. 471–481.
[32] ———, *Remarks on some recent extensions of Filippov's implicit functions lemma*, this Journal, 5 (1967), pp. 622–627.
[33] ———, *On the approximation of integrals of multivalued functions*, this Journal, 7 (1969), pp. 158–177.
[34] G. L. KHARATISHVILI, *Extremal problems in linear topological spaces*, Doctoral thesis, Tbilisi State University, U.S.S.R., 1968.

[35] ———, *A maximal principle in extremal problems with delays*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967.
[36] K. KURATOWSKI AND C. RYLL-NARDZEWSKI, *A general theorem on selectors*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 13 (1965), pp. 397–403.
[37] K. KURATOWSKI, *Les fonctions semi-continues dans l'espace des ensembles fermés*, Fund. Math., 18 (1932), pp. 148–160.
[38] E. B. LEE, *Variational problems for systems having delay in the control action*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 697–699.
[39] ———, *Geometric theory of linear controlled systems*, to appear.
[40] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[41] N. LEVINSON, *Minimax, Liapunov and "bang-bang,"* J. Differential Equations, 2 (1966), pp. 218–241.
[42] W. LI, *Mathematical models in the biological sciences*, Master's thesis, Brown University, Providence, R.I., 1970.
[43] A. A. LYAPUNOV, *Sur les fonctions-vecteurs complètement additives*, Izv. Akad. Nauk SSSR Ser. Mat., 8 (1940), pp. 465–478.
[44] S. LOJASIEWICZ, *Ensembles semi-analytiques*, Lecture Notes, Institut des Hautes Études Scientifiques, Bures-sur-Yvette, 1965.
[45] R. NARASIMHAN, *Introduction to the Theory of Analytic Spaces*, Springer-Verlag Lecture Notes in Mathematics, Berlin–Heidelberg, 1966.
[46] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.
[47] C. OLECH, *Extremal solutions to a control system*, J. Differential Equations, 2 (1966), pp. 74–101.
[48] M. N. OGUZTÖRELI, *Time-Lag Control Systems*, Academic Press, New York, 1966.
[49] E. PINNEY, *Ordinary Difference–Differential Equations*, University of California Press, Berkeley, 1958.
[50] H. L. ROYDEN, *Real Analysis*, 2nd ed., Macmillan, New York, 1968.
[51] V. VOLTERRA, *Theory of Functionals*, Blackie, London, 1930.
[52] J. L. WALSH, *The Cauchy–Goursat theorem for rectifiable Jordan curves*, Proc. Nat. Acad. Sci. U.S.A., 19 (1933), pp. 540–541.
[53] A. M. ZVERKIN, G. A. KEMENSKII, S. B. NORKIN AND L. E. EL'SGOL'TS, *Differential equations with a perturbed argument. I and II*, Russian Math. Surveys, 17 (1962), pp. 61–146 and Trudy Sem. Teor. Differential. Uravnenii s Otklon. Argumentom Univ. Druzhby Narodov Patrisa Lumumby, 2 (1963), pp. 3–49.

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

SOME SUFFICIENT CONDITIONS FOR THE GLOBAL AND LOCAL CONTROLLABILITY OF NONLINEAR TIME-VARYING SYSTEMS*

E. J. DAVISON† AND E. G. KUNZE‡

Summary. Sufficient conditions are derived for global and local controllability of nonlinear time-varying systems with control appearing linearly. It is shown that the controllability of $\dot{x} = A(t, x)x + B(t, x)u$ can be related to the controllability of the linear system $\dot{x} = A(t, z)x + B(t, z)u$, where z belongs to a certain set of continuous vector functions. This result is then used to specify a class of nonlinear systems which are globally controllable.

1. Introduction. This paper deals with the controllability of nonlinear time-varying systems with control appearing linearly. This problem has been studied by Hermes [1], who showed that a nonlinear system is locally controllable, if the associated Pfaffian system is not integrable. Markus and Lee [2] and Kalman [3] obtained conditions for local controllability of a more general class of systems than will be considered here, by showing that the nonlinear system is locally controllable if the linearized system is completely controllable.

The results obtained in this paper provide sufficient conditions for *complete* and *total controllability* as defined by Kreindler and Sarachik [4] for linear systems. A distinction is made between *global controllability* (i.e., the system is controllable in the whole of the state space R^n) and *local controllability* (i.e., the system is controllable only in some domain of R^n). It is shown (Theorem 1) that the system $\dot{x} = A(t, x)x + B(t, x)u$ is globally completely (totally) controllable, if the linear system $\dot{x} = A(t, z)x + B(t, z)u$ is completely (totally) controllable for all functions $z \in C_n[t_0, t_f]$. If the linear system is controllable only for z in some bounded family $\mathscr{C}$, then a criterion for local controllability results (Theorem 2). Theorem 3 shows how the controllability matrix $Q(t, z)$ introduced by Silverman and Meadows [5] can be used to test whether the linear system is controllable for all functions z belonging to $C_n[t_0, t_f]$ or $\mathscr{C}$, thereby giving a simple computable criterion for the global or local controllability of single input nonlinear systems. This criterion is then used to specify a class of nonlinear systems which are globally controllable (Theorem 4).

In deriving these results, the problem is transformed into one of showing the existence of a fixed point for a mapping $x = P(z)$, which is solved by using Schauder's fixed-point theorem. The existence of a fixed point requires that the determinant of Kalman's controllability matrix [1] of the linear system, here denoted by $G(t_0, t_f; z)$ with initial time t_0 and final time t_f, have a positive lower bound relative to z in $C_n[t_0, t_f]$ or $\mathscr{C}$.

2. Preliminaries. Consider the nonlinear time-varying system with linear control represented by the equation

$$dx/dt = A(t, x)x + B(t, x)u, \qquad t_0 \leqq t < \infty, \tag{1}$$

* Received by the editors November 6, 1969, and in revised form March 3, 1970. This work was supported by the National Research Council of Canada under Grant A 4396.

† Department of Electrical Engineering, University of Toronto, Toronto, Ontario.

‡ Department of Electrical Engineering, University of Toronto, Toronto, Ontario. Now at Institut für Schwingungsforschung, 75 Karlsruhe-Waldstadt, Germany.

where the state x is an n-vector, the control input u an m-vector, A is an $n \times n$ and B an $n \times m$ matrix. Assume that the elements $a_{ik}(t, x)$ of A $(i, k = 1, 2, 3, \cdots, n)$ and the elements $b_{il}(t, x)$ of B $(i = 1, 2, 3, \cdots, n, l = 1, 2, 3, \cdots, m)$ are continuous functions of x for fixed t and piecewise continuous functions of t for fixed x and fulfill the following conditions:

$$|a_{ik}(t, x)| \leqq M, \quad |b_{il}(t, x)| \leqq N \quad \text{for all } x \in R^n, \quad t \in [t_0, t_f], \tag{2}$$

where M and N are positive real constants.

The following definitions are due to Kreindler and Sarachik [4].

DEFINITION 1. The system (1) is said to be *completely state controllable* at t_0 in the domain of controllability $D \subset R^n$, if each initial state $x(t_0)$ in D can be transferred to any final state x_f in D in some finite time $t_f(x_f) \geqq t_0$. (If D is the whole state space R^n, the controllability is said to be *global*. If D is not the whole of R^n, the controllability is said to be *local*.)

DEFINITION 2. The system given by (1) is said to be *totally state controllable* in the domain of controllability D, if it is completely state controllable in D on every interval $[t_0, t_f]$, $t_f > t_0$. (If D is the whole of R^n, the controllability is said to be *global*, otherwise it is said to be *local*.)

To derive sufficient conditions for the controllability of system (1) consider first the simpler system

$$dx/dt = A(t, z)x + B(t, z)u, \tag{3}$$

where the argument x of A and B has been replaced by a specified function $z \in C_n[t_0, t_f]$, the Banach space of continuous R^n-valued functions on $[t_0, t_f]$. For each fixed $z \in C_n[t_0, t_f]$, system (3) is linear; and with $x(t_0) = x_0$, the solution is given by

$$x(t) = \phi(t, t_0; z)x_0 + \int_{t_0}^{t} \phi(t, \tau; z)B(\tau, z)u(\tau)\, d\tau. \tag{4}$$

In (4), $\phi(t, t_0; z)$ is the state transition matrix of the system

$$dx/dt = A(t, z)x \tag{5}$$

and is determined by

$$\frac{d}{dt}\phi(t, t_0; z) = A(t, z)\phi(t, t_0; z), \qquad \phi(t_0, t_0; z) = I, \tag{6}$$

where I is the identity matrix. Define

$$H(t_0, \tau; z) = \phi(t_0, \tau; z)B(\tau, z), \tag{7}$$

$$G(t_0, t; z) = \int_{t_0}^{t} H(t_0, \tau; z)H'(t_0, \tau; z)\, d\tau. \tag{8}$$

The prime indicates the matrix transpose.

Necessary and sufficient conditions for system (3) to be controllable are summarized by the following lemmas (Kreindler and Sarachik [4]).

LEMMA 1. *System* (3) *is completely state controllable at* t_0 *if and only if there exists a finite time* $t_f > t_0$ *such that the rows of the matrix* $H(t_0, \tau; z)$ *are linearly independent functions of* τ *on* $[t_0, t_f]$.

LEMMA 2. *System* (3) *is totally state controllable if and only if for all* t_0 *and for all* $t_f > t_0$ *the rows of matrix* $H(t_0, \tau; z)$ *are linearly independent functions of* τ *on* $[t_0, t_f]$.

3. Derivation of results—global controllability. Assume that system (3) is either completely or totally controllable for all $z \in C_n[t_0, t_f]$. For complete controllability, by Lemma 1, the rows of $H(t_0, \tau; z)$ are linearly independent functions of τ on some $[t_0, t_f]$. This implies that the matrix $G(t_0, t_f; z)$ defined by (8) (the Gramian matrix for the set of m-dimensional vector functions H on the interval $[t_0, t_f]$) is positive definite for $t = t_f$. Total controllability, by Lemma 2, then implies that the Gramian matrix $G(t_0, t_f; z)$ is positive definite for all t_0 and all $t_f > t_0$. In either case a control u always exists such that the system (3) can be transferred from any $x_0 \in R^n$ to any $x_f \in R^n$ in a finite time. Consider the control

$$u(t_0, t, t_f; z) = H'(t_0, t; z)G(t_0, t_f; z)^{-1}[\phi(t_f, t_0; z)^{-1}x_f - x_0]. \tag{9}$$

Using (7) and (8) and inserting (9) into (4), we obtain from (3):

$$x(t) = \phi(t, t_0; z)\{x_0 + G(t_0, t; z)G(t_0, t_f; z)^{-1}[\phi(t_f, t_0; z)^{-1}x_f - x_0]\}; \tag{10}$$

and it is easily verified that

$$x(t_0) = x_0, \qquad x(t_f) = x_f.$$

Clearly $u(t_0, t, t_f; z)$ as defined by (9) will transfer the system from x_0 to x_f for all $z \in C_n[t_0, t_f]$. In the following discussion it will be convenient to view the right side of (10) as an operator $P(z)(t)$, i.e.,

$$P(z)(t) = \phi(t, t_0; z)\{x_0 + G(t_0, t; z)G(t_0, t_f; z)^{-1}[\phi(t_f, t_0; z)^{-1}x_f - x_0]\} \tag{11}$$

so that (10) can be written in the form

$$x = P(z). \tag{12}$$

The following theorem now gives conditions under which the nonlinear system (1) is globally controllable.

THEOREM 1 (Global controllability). *The system*

$$dx/dt = A(t, x)x + B(t, x)u$$

is globally (a) *completely state controllable at* t_0 *or* (b) *totally state controllable, if the following three conditions all hold*:

(A) *The elements* $a_{ik}(t, x)$ *of* A $(i, k = 1, 2, \cdots, n)$ *and* $b_{il}(t, x)$ *of* B $(l = 1, 2, \cdots, m,\ i = 1, 2, \cdots, n)$ *are piecewise continuous functions of* t *and continuous functions of* x.

(B) $|a_{ik}(t, x)| \leqq M$, $|b_{il}(t, x)| \leqq N$, *for all* $x \in R^n$, $t \in [t_0, t_f]$, *where* M *and* N *are positive real constants.*

(C) *There exists a constant* $c > 0$ *such that*

$$\inf_{z \in C_n[t_0, t_f]} \det G(t_0, t_f; z) \geqq c$$

(a) *for some* $t_f > t_0$, *in the case of complete state controllability at* t_0,

(b) *for all* t_0 *and for all* $t_f > t_0$, *in the case of total state controllability.*

The proof of the theorem will be based on the following lemma.

LEMMA 3. *If conditions* (A), (B), (C) *of Theorem* 1 *are satisfied, then for every pair* $x_0, x_f \in R^n$, *the operator* P *defined by* (11) *has a fixed point in* $C_n[t_0, t_f]$.

Proof of Lemma 3. Define $|z(t)| = \sum_{i=1}^{n} |z_i(t)|$ and let the norm in $C_n[t_0, t_f]$ be

$$\|z\| = \max \{|z(t)| : t_0 \leqq t \leqq t_f\}. \tag{13}$$

Consider the closed and convex subset of $C_n[t_0, t_f]$:

$$\psi \equiv \{z | z \in C_n[t_0, t_f], \|z\| \leqq K\}, \tag{14}$$

where the constant K is defined by

$$K = \{(1 + C)|x_0| + C|x_f|e^{nM(t_f - t_0)}\}e^{nM(t_f - t_0)} \tag{15}$$

with

$$C = \sup_{z \in C_n[t_0, t_f]} \|G(t_0, t_f; z)^{-1}\| \, nmN^2(t_f - t_0)e^{2nM(t_f - t_0)} \tag{16}$$

where

$$\|G(t_0, t_f; z)^{-1}\| = \max_j \sum_{i=1}^{n} |g_{ij}(t_0, t_f; z)|,$$

where $G(t_0, t_f; z)^{-1} = \{g_{ij}(t_0, t_f; z)\}$. Let Ω be the image of ψ:

$$\Omega \equiv \{x | x = P(z), z \in \psi\}. \tag{17}$$

It is clear that the operator P as defined by (11) is continuous and it is easily established from the Arzela–Ascoli theorem [6] that the image set Ω defined by (17) is compact and is a subset of ψ defined by (14). Hence by Schauder's theorem [6], the operator has a fixed point.

Proof of Theorem 1. The significance of Lemma 3 is that there always exists at least one function $z^* \in C_n[t_0, t_f]$, which, introduced into (10), provides an x^* such that $x^* = z^*$. This x^*, however, is a solution to system (1) for the control input $u(t_0, t, t_f; z^*)$, which is easily verified by differentiating x^* with respect to t. Since $u(t_0, t, t_f; z^*)$ takes system (1) from x_0 to x_f on the interval $[t_0, t_f]$, and since by Lemma 3 there is a $u(t_0, t, t_f; z^*)$ for all $x_0, x_f \in R^n$, system (1) is *globally controllable.* In particular, if condition (C(a)) of Theorem 1 holds, the above conclusion is true for *some* finite time interval $[t_0, t_f]$, and the system is *completely controllable.* If condition (C(b)) of Theorem 1 holds, the above conclusion is true for *every* finite time interval $[t_0, t_f]$, and the system is therefore *totally controllable.*

4. Local controllability. The method used to establish Theorem 1 for global controllability can be used to derive a theorem for local controllability under less restrictive conditions; i.e., it will no longer be necessary that the elements of A and B be bounded for all $x \in R^n$, and the Gramian determinant need only have a lower bound on a bounded set of functions z. This bounded set is defined by

$$\mathscr{C} = \{z | z = C_n[t_0, t_f]; z(t_0) = x_0, z(t_f) = x_f; x_0, x_f \in R^n; \|z\| \leqq K_1\}, \tag{18}$$

where K_1 is some real positive nonzero constant.

THEOREM 2 (Local controllability). *The system*

$$dx/dt = A(t, x)x + B(t, x)u$$

is locally, (a) *completely state controllable at* t_0, *or* (b) *totally state controllable, about the origin if the following three conditions all hold*:

(A) *The elements* $a_{ik}(t, x)$ *of* A $(i, k = 1, 2, \cdots, n)$ *and* $b_{il}(t, x)$ *of* B $(l = 1, 2, \cdots, m,\ i = 1, 2, \cdots, n)$ *are piecewise continuous functions of* t *and continuous functions of* x.

(B) $|a_{ik}(t, z)| \leqq M, |b_{il}(t, z)| \leqq N$ *for all* $z \in \mathscr{C}$, $t \in [t_0, t_f]$, *where* M *and* N *are positive real constants.*

(C) $$\inf_{z \in \mathscr{C}} \det G(t_0, t_f; z) \geqq c \quad \text{for some } c > 0$$

(a) *for some* $t_f > t_0$, *in the case of complete state controllability at* t_0,

(b) *for all* t_0 *and for all* $t_f > t_0$, *in the case of total state controllability.*

The proof of the theorem is based on the following lemma, which is a counterpart to Lemma 3.

LEMMA 4. *The operator* P *defined by* (11) *has a fixed point in* $\mathscr{C}$ *defined by* (18) *such that* $x^* = P(x^*)$, *if* $|x_0| < K_2$, $|x_f| < K_3$, *where* K_2 *and* K_3 *are real positive constants which are sufficiently small and not both zero, and if conditions* (A) *to* (C) *of Theorem* 2 *are fulfilled.*

Proof of Lemma 4. The proof follows exactly the same reasoning as Lemma 3.

Proof of Theorem 2. The proof is the same as for Theorem 1, if one uses $\mathscr{C}$ instead of $C_n[t_0, t_f]$ and Lemma 4 instead of Lemma 3.

5. Relation of Gramian matrix to controllability matrix. A serious difficulty in the application of Theorems 1 or 2 is to show that condition (C) holds. Therefore a relation which shows that condition (C) holds (at least for certain cases) will now be given.

If the additional assumption is introduced that $A(t, z)$ and $B(t, z)$ are piecewise differentiable on $[t_0, t_f]$ at least $n - 2$ and $n - 1$ times, respectively, then the controllability matrix Q of Silverman and Meadows [5] can be introduced.

Define the matrix

$$Q(t; z, z^{(1)}), \cdots, z^{(n-1)} = [P_0(t; z), P_1(t; z, z^{(1)}), \cdots, P_{n-1}(t; z, z^{(1)}, \cdots, z^{(n-1)})], \tag{19}$$

where $P_k(t; z, z^{(1)}, \cdots, z^{(k)})$ is recursively defined by

$$P_k(t; z, z^{(1)}, \cdots, z^{(k)}) = -A(t, z)P_{k-1}(t; z, z^{(1)}, \cdots, z^{(k-1)}) + \frac{d}{dt}P_{k-1}(t; z, z^{(1)}, \cdots, z^{(k-1)}), \tag{20}$$

$$P_0(t, z) = B(t, z). \tag{21}$$

For simplicity denote $Q(t; z, z^{(1)}, \cdots, z^{(n-1)})$ by $Q(t, z)$. The results obtained in [7] allow the formulation of the following theorem.

THEOREM 3. *Assume that* $A(t, z)$ *and* $B(t, z)$ *of* (3) *are piecewise differentiable on* $[t_0, t_f]$ *at least* $n - 2$ *and* $n - 1$ *times, respectively, and that* $B(t, z)$ *is an* $n \times 1$ *vector. If* $\inf_{z \in C_n[t_0, t_f]\ (\text{or } \mathscr{C})} [\det Q(t, z)]^2 \geqq \gamma$ *for some* $\gamma > 0$ *and for some* t *in* $[t_\alpha, t_\beta]$, *where* $[t_\alpha, t_\beta]$ *is a subinterval of* $[t_0, t_f]$, *then* $\det G(t_\alpha, t_\beta; z)$ *of Theorems* 1 *and* 2

has a lower bound such that

$$\inf_{z \in C_n[t_0, t_f] \,(\text{or } \mathscr{C})} \det G(t_\alpha, t_\beta; z) \geqq \varepsilon \quad \textit{for some } \varepsilon > 0.$$

6. Some numerical examples.

Example 1. Consider the system

$$\begin{aligned} \dot{x}_1 &= x_2 + \sin[g(x_1, x_2, t)]u, \\ \dot{x}_2 &= -x_1 + \sin[g(x_1, x_2, t)]u, \end{aligned} \tag{22}$$

where $g(x_1, x_2, t)$ is a continuous function of x_1, x_2 and a piecewise continuous function of t and satisfies the following inequality:

$$0 < \varepsilon \leqq g(x_1, x_2, t) \leqq \pi - \varepsilon \qquad \text{for all } x_1, x_2 \in C_n[t_0, t_f], \quad t \in [t_0, t_f].$$

By using Theorems 1 and 3 global total controllability is easily established. The coefficients of x_1, x_2 and u fulfill the conditions (A) and (B) of Theorem 1. Condition (C(b)) is established by using Theorem 3. The determinant of the controllability matrix is

$$\det Q(t, z) = \begin{vmatrix} \sin[g(z_1, z_2, t)] & \sin[g(z_1, z_2, t)] - \dfrac{d}{dt}\sin[g(z_1, z_2, t)] \\ \sin[g(z_1, z_2, t)] & -\sin[g(z_1, z_2, t)] - \dfrac{d}{dt}\sin[g(z_1, z_2, t)] \end{vmatrix} \tag{23}$$

which yields

$$\det Q(t, z) = -2\sin^2[g(z_1, z_2, t)]. \tag{24}$$

From (24) the smallest lower bound is readily determined:

$$\inf_{z \in C_n[t_0, t_f]} [\det Q(t, z)]^2 \geqq 4\sin^4 \varepsilon > 0, \tag{25}$$

which holds for all t. Hence by Theorem 3, condition (C(b)) of Theorem 1 is satisfied, and system (22) is therefore globally totally controllable.

Example 2. Consider the system

$$\dot{x} = \begin{pmatrix} 0 & \dfrac{1}{1 - x_1^2 x_2^2} \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u. \tag{26}$$

For $|\max(x_1, x_2)| \leqq K < 1$, conditions (A) and (B) of Theorem 2 are satisfied. Theorem 3 will be used to establish the third condition. The determinant of the controllability matrix is

$$\det Q(t, z) = \frac{-1}{1 - z_1^2 z_2^2}, \tag{27}$$

and

$$\inf_{z \in \mathscr{C}} [\det Q(t, z)]^2 = \inf_{z \in \mathscr{C}} \left(\frac{-1}{1 - z_1^2 z_2^2} \right)^2 = \frac{1}{(1 - K^4)^2}. \tag{28}$$

Hence condition (C(b)) of Theorem 2 holds and system (26) is locally totally controllable about the origin.

7. A class of nonlinear systems which is globally controllable. Theorems 1 and 3 can be used to establish classes of nonlinear systems which are globally controllable. The following is an example.

THEOREM 4. *Consider the system*

$$\dot{x} = A(x, t)x + B(x, t)u, \tag{29a}$$

where $A(x, t)$ *and* $B(x, t)$ *have the following form*:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & 0 & \cdots & 0 & \cdots & 0 \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots & 0 & & 0 \\ \cdot & \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot & & \cdot \\ & & & & a_{n-2,n-1} & & 0 \\ a_{n-1,1} & a_{n-1,2} & a_{n-1,3} & \cdots & a_{n-1,n-1} & & a_{n-1,n} \\ a_{n,1} & a_{n,2} & a_{n,3} & \cdots & a_{n,n-1} & & a_{n,n} \end{bmatrix}, \tag{29b}$$

$$B = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ b_n \end{bmatrix}; \tag{29c}$$

then a sufficient condition that system (29) *be*

(a) *globally*

(i) *completely state controllable at* t_0,

(ii) *totally state controllable*,

(b) *locally about the origin*

(i) *completely state controllable at* t_0,

(ii) *totally state controllable*

is that the following three conditions all hold:

(A) *The elements* $a_{ik}(t, z)$ *of* $A(i, k = 1, 2, \cdots, n)$ *are piecewise differentiable on* $[t_0, t_f]$ *at least* $n - 2$ *times and* $b_n(t, z)$ *is piecewise differentiable on* $[t_0, t_f]$ *at least* $n - 1$ *times.*

(B) (a) *In the case of global controllability*

$$|a_{il}(t, x)| \leqq M, \qquad |b_n(t, x)| \leqq N \qquad \text{for all } x \in R^n, \quad t \in [t_0, t_f].$$

(b) *In the case of local controllability*

$$|a_{ii}(t, z)| \leqq M, \qquad |b_n(t, z)| \leqq N \qquad \text{for all } z \in \mathscr{C}, \quad t \in [t_0, t_f],$$

where $\mathscr{C}$ *is defined by* (18).

(C) (a) *In the case of global controllability, there exists a constant* $c > 0$ *such that*

$$b_n^2(t_j, z) \geqq c, \qquad a_{i,i+1}^2(t_j, z) \geqq c \quad \text{for all } z \in C_n[t_0, t_f],$$

$i = 1, 2, \cdots, n-1$, *for some* $t_j \in [t_0, t_f]$.

(b) *In the case of local controllability, there exists a constant* $c > 0$ *such that*

$b_n^2(t_j, z) \geqq c$, $a_{i,i+1}^2(t_j, z) \geqq c$ *for all* $z \in \mathscr{C}$, $i = 1, 2, \cdots, n-1$, *for some* $t_j \in [t_0, t_f]$,

(i) *for some* $t_f > t_0$ *in the case of complete state controllability at* t_0,

(ii) *for all* t_0 *and for all* $t_f > t_0$, *in the case of total state controllability.*

Proof. It is easily established for (29), that

$$\det Q(t, z) = b_n^n \prod_{i=1}^{n-1} a_{i,i+1}^i. \tag{30}$$

Theorem 4 immediately follows on using this result together with Theorems 1, 2 and 3.

Remark. It is seen then that for the class of systems given by (29), the system is globally totally controllable if all the elements are bounded and if the product of the superdiagonal elements of A with b_n is equal to zero at most only a countable number of times. This is a generalization of the nonlinear time-varying system considered in [8] which is as follows:

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ & & & & 0 & 1 \\ a_{n,1} & a_{n,2} & a_{n,3} & \cdots & a_{n,n-1} & a_{n,n} \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u. \tag{31}$$

Acknowledgment. Initial discussion of the problem with Professors L. M. Silverman and P. P. Varaiya is greatly appreciated.

REFERENCES

[1] H. Hermes, *Controllability and the singular problem,* this Journal, 2 (1965), pp. 241–260.

[2] L. Markus and E. B. Lee, *On the existence of optimal controls*, Trans. ASME Ser. D. J. Basic Engrg., 84 (1962), pp. 13–20.

[3] R. E. Kalman, *Discussion to* [2], Ibid., 84 (1962), pp. 21–22.

[4] E. KREINDLER AND P. E. SARACHIK, *On the concepts of controllability and observability of linear systems,* IEEE Trans. Automatic Control, AC-9 (1964), pp. 129–136.

[5] L. M. SILVERMAN AND H. E. MEADOWS, *Controllability and observability in time-variable linear systems,* this Journal, 5 (1967), pp. 64–73.

[6] L. W. KANTOROVICH AND G. P. AKILOV, *Functional Analysis in Normed Spaces*, Pergamon Press, Oxford, 1964.

[7] L. SILVERMAN AND B. D. O. ANDERSON, *Controllability, observability and stability of linear systems,* this Journal, 6 (1968), pp. 121–130.

[8] E. J. DAVISON, L. M. SILVERMAN AND P. P. VARAIYA, *Controllability of a class of nonlinear time-variable systems,* IEEE Trans. Automatic Control, AC-12 (1967), pp. 791–792.

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

SOME SUFFICIENT CONDITIONS FOR OPTIMALITY IN CONTROL PROBLEMS WITH STATE SPACE CONSTRAINTS*

J. E. FUNK† AND E. G. GILBERT‡

1. Introduction. The sufficient conditions obtained in this paper are an outgrowth of the work of Mangasarian [5]. As in his paper, a chain of inequalities and some ad hoc assumptions lead to a simple and direct proof of the main results. However, the hypotheses here are weaker, a somewhat different problem is treated, and jumps in the "multipliers" corresponding to the differential constraints are allowed. The inclusion of the jumps is important, for, without them it is impossible to prove optimality in almost all optimal control problems where there is a state constraint of the form $g(x, t) \leqq 0$. If certain convexity and normality assumptions are imposed in the optimal control problem considered, the sufficient conditions become necessary. The corresponding result does not follow for the conditions given in [5]. The sufficient conditions given here also apply to a number of interesting control problems without state space constraints, including those where sufficient conditions of a similar type have been obtained previously [3], [4], [5].

Very recently Mangasarian and Schumaker [6] extended the conditions in [5] to allow for jumps in the multipliers. These conditions were used to solve certain problems (the determination of spline functions) in which the sufficiency conditions of [5] were not adequate. For some of the problems which were considered the sufficient conditions were also shown to be necessary, but the close relations between necessary conditions and sufficient conditions developed in this paper were not made clear.

The general method of attack is to start with a set of conditions which are necessary and then suitably strengthen them to obtain the sufficient conditions. Depending on the necessary conditions and optimization problem formulation, one can obtain different sets of sufficient conditions. The treatment in this paper is restricted to the rather general problem formulation and necessary conditions given by Neustadt [7], [8]. Similar developments can also be carried out for the problem formulation and necessary conditions given by Gamkrelidze [9], Hestenes and Guinn [2], [1], Warga [10], [11], and others.

2. Problem formulation and necessary conditions. Let the following problem data and conditions be given: $I = [t_1, t_2] \subset R^1$, a compact interval; $G \subset R^n$, an open set; $U \subset R^r$, a compact set; $U_t \subset U$, an arbitrary set for each $t \in I$; $f(x, u, t)$, a continuous function from $G \times U \times I$ into R^n whose derivative with respect to x, $f_x(x, u, t)$, exists and is continuous on $G \times U \times I$; $g(x, t)$, a function from $G \times I$ into R^1 which is C^2 on $G \times I$; $\chi_i(x_1, x_2)$ for $i = -\mu, \cdots, 0, \cdots, m$ (μ and m are nonnegative integers), functions from $G \times G$ into R^1 whose derivatives, $\chi_{ix_1}(x_1, x_2)$ and $\chi_{ix_2}(x_1, x_2)$, exist and are continuous on $G \times G$.

* Received by the editors September 4, 1969, and in revised form February 4, 1970. This work was supported by the United States Air Force under Grant AFOSR-69-1767.

† Air Force Institute of Technology, Wright-Patterson Air Force Base, Dayton, Ohio 45433.

‡ Computer, Information and Control Engineering Program, University of Michigan, Ann Arbor, Michigan 48104.

Let $\mathcal{U}$ denote the set of those functions $u(t)$ from I into R^r which are measurable and satisfy $u(t) \subset U_t$ for almost all $t \in I$. Given $u \in \mathcal{U}$, $x(t)$ is said to be a solution of the equations of motion,

$$\dot{x}(t) = f(x(t), u(t), t), \tag{2.1}$$

corresponding to u if $x(t)$ is an absolutely continuous function from I into G which satisfies (2.1) for almost all $t \in I$. The optimal control problem can now be stated.

Optimal control problem. Find $u \in \mathcal{U}$ such that the corresponding solution x satisfies

$$\chi_i(x(t_1), x(t_2)) \leqq 0, \quad i = -\mu, \cdots, -1 \quad \text{(omit if } \mu = 0\text{)}, \tag{2.2}$$

$$\chi_i(x(t_1), x(t_2)) = 0, \quad i = 1, \cdots, m \quad \text{(omit if } m = 0\text{)}, \tag{2.3}$$

$$g(x(t), t) \leqq 0 \quad \text{for all } t \in I, \tag{2.4}$$

and $\chi_0(x(t_1), x(t_2))$ is minimum.

THEOREM 2.1 (Necessary conditions [7], [8]). *Let $w \in \mathcal{U}$ and z denote respectively an optimal control and the corresponding solution of the equations of motion. Then there exist a (row) function ψ from I into R^n, a (scalar) function λ from I into R^1 and elements $\alpha_i \in R^1$ for $i = -\mu, \cdots, m$ such that the following conditions hold:*

(N1) *$\psi(t)$ is absolutely continuous and satisfies*

$$\dot{\psi}(t) = -\psi f_x(z(t), w(t), t) + \lambda(t) p_x(z(t), t) \tag{2.5}$$

for almost all $t \in I$, where

$$p(x, t) = g_x(x, t) f(x, w(t), t) + g_t(x, t). \tag{2.6}$$

Also $\psi(t) \not\equiv g_x(z(t), t)$ on a subset of I of positive measure.

(N2) *$\lambda(t)$ is continuous from the right in (t_1, t_2), is nonincreasing on I, is constant on all subintervals of I such that $g(z(t), t) < 0$ and satisfies*

$$\lambda(t_2) = 0. \tag{2.7}$$

(N3)

$$\alpha_0 \leqq 0, \tag{2.8}$$

$$\left.\begin{array}{ll} \alpha_i \leqq 0, & \chi_i(z(t_1), z(t_2)) = 0 \\ \alpha_i = 0, & \chi_i(z(t_1), z(t_2)) < 0 \end{array}\right\} \quad i = -\mu, \cdots, -1, \tag{2.9}$$

$$\sum_{i=-\mu}^{m} |\alpha_i| + \lambda(t_1) > 0, \tag{2.10}$$

$\psi(t_1)$ and $\psi(t_2)$ satisfy

$$\psi(t_1) = -\sum_{i=-\mu}^{m} \alpha_i \chi_{ix_1}(z(t_1), z(t_2)) + \lambda(t_1) g_x(z(t_1), t_1), \tag{2.11}$$

$$\psi(t_2) = \sum_{i=-\mu}^{m} \alpha_i \chi_{ix_2}(z(t_1), z(t_2)). \tag{2.12}$$

(N4) *$w(t)$ satisfies the integral maximum condition*

$$\int_{t_1}^{t_2} [\psi(t) - \lambda(t)g_x(z(t), t)][f(z(t), w(t), t) - f(z(t), u(t), t)]\, dt \geqq 0 \tag{2.13}$$

for all $u \in \mathcal{U}$,

which, if $U_t = U$ *for all* $t \in I$, *can be replaced by the pointwise maximum condition*

$$[\psi(t) - \lambda(t)g_x(z(t), t)][f(z(t), w(t), t) - f(z(t), u, t)] \geqq 0 \tag{2.14}$$

for all $u \in U$ *and almost all* $t \in I$.

It should be emphasized that the continuity of $\psi(t)$ in the above necessary conditions is not in conflict with the jumps in adjoint variables which appear in some sets of necessary conditions (see, e.g., [9]). In fact, the jumps occur in the multiplier $\psi(t) - \lambda(t)g_x(z(t), t)$ because of discontinuities in $\lambda(t)$. If λ is piecewise differentiable and proper allowances for notation and problem differences are made, the substitutions $\bar{w} = -\dot{\lambda}$ and $\bar{v} = -\psi + \lambda g_x$ will yield a set of conditions which, while not equivalent to the sufficient conditions of Mangasarian and Schumaker [6], may be identified with them.

3. The sufficient conditions. First the sufficiency theorem will be stated and proved. Then some general remarks will be made. In the next section several corollaries which are of interest in applications are given.

THEOREM 3.1 (Sufficient conditions). *Let* $w \in \mathcal{U}$ *and* z *denote respectively a control and corresponding solution of the equations of motion such that* $z(t)$ *satisfies the constraints* (2.2), (2.3) *and* (2.4) *when* z *replaces* x. *Further assume that there exist* $\psi, \lambda, \alpha_{-\mu}, \cdots, \alpha_m$ (*as in Theorem* 2.1) *such that conditions* (S1)–(S9) *given below hold. Then* $w(t)$ *is optimal.*

(S1) *Same as* (N1) *except the requirement that* $\psi(t) \neq \lambda(t)g_x(z(t), t)$ *is omitted.*

(S2) *Same as* (N2) *except the requirement of continuity of* $\lambda(t)$ *from the right is omitted.*

(S3) *Same as* (N3) *except requirements* (2.8), (2.9) *and* (2.10) *are omitted.*

(S4) *Same as* (N4).

(S5) $\alpha_0 < 0$.

(S6) *For all* $x_1 \in G$, $x_2 \in G$ *such that* $\chi_i(x_1, x_2) \leqq 0$ *for* $i = -\mu, \cdots, -1$ *and* $\chi_i(x_1, x_2) = 0$ *for* $i = 1, \cdots, m$,

$$\begin{aligned}\chi_0(x_1, x_2) - \chi_0(z(t_1), z(t_2)) \geqq\ & \chi_{0x_1}(z(t_1), z(t_2))(x_1 - z(t_1)) \\ & + \chi_{0x_2}(z(t_1), z(t_2))(x_2 - z(t_2)).\end{aligned} \tag{3.1}$$

(S7) *For all* $x_1 \in G$, $x_2 \in G$ *such that* $\chi_i(x_1, x_2) \leqq 0$ *for* $i = -\mu, \cdots, -1$ *and* $\chi_i(x_1, x_2) = 0$ *for* $i = 1, \cdots, m$,

$$\sum_{\substack{i=-\mu \\ i \neq 0}}^{m} \alpha_i[\chi_{ix_1}(z(t_1), z(t_2))(x_1 - z(t_1)) + \chi_{ix_2}(z(t_1), z(t_2))(x_2 - z(t_2))] \geqq 0. \tag{3.2}$$

(S8) *For all* $t \in I$ *and* $x \in G$ *such that* $g(z(t), t) = 0$ *and* $g(x, t) \leqq 0$,

$$g_x(z(t), t)(x + z(t)) \leqq 0. \tag{3.3}$$

(S9) *For all* $t \in I$, $u \in U_t$ *and* $x \in G$ *such that* $g(x, t) \leqq 0$,

$$[\psi(t) - \lambda(t)g_x(z(t), t)][f_x(z(t), w(t), t)(x - z(t)) - f(x, u, t) + f(z(t), u, t)] \geqq 0. \tag{3.4}$$

Proof. Let $u(t) \in \mathcal{U}$ and $x(t)$ be any control-solution pair such that $x(t)$ satisfies (2.2), (2.3) and (2.4). It must be shown that $\Delta\chi_0 = \chi_0(x(t_1), x(t_2)) - \chi_0(z(t_1), z(t_2)) \geqq 0$. Begin by using (S6) to obtain

$$\Delta\chi_0 \geqq \chi_{0x_1}(z(t_1), z(t_2))(x(t_1) - z(t_1)) + \chi_{0x_2}(z(t_1), z(t_2))(x(t_2) - z(t_2)). \tag{3.5}$$

From (S3) and (S5) this is equivalent to

$$\begin{aligned}\Delta\chi_0 \geqq -\alpha_0^{-1}\Big\{&\psi(t_1)(x(t_1) - z(t_1)) - \psi(t_2)(x(t_2) - z(t_2)) \\ &+ \sum_{\substack{i=-\mu \\ i\neq 0}}^{m} \alpha_i[\chi_{ix_1}(z(t_1), z(t_2))(x(t_1) - z(t_1)) + \chi_{ix_2}(z(t_1), z(t_2))(x(t_2) - z(t_2))] \\ &- \lambda(t_1)g_x(z(t_1), t_1)(x(t_1) - z(t_1))\Big\}.\end{aligned} \tag{3.6}$$

Now application of (S7) gives a stronger inequality which through (2.7) of (S2) can be written

$$\Delta\chi_0 \geqq \alpha_0^{-1}\int_{t_1}^{t_2} d[(\psi(t) - \lambda(t)g_x(z(t), t))(x(t) - z(t))]. \tag{3.7}$$

Because ψ, x and z are absolutely continuous and λ is nonincreasing the integral exists and may be expanded to give

$$\begin{aligned}\Delta\chi_0 \geqq \alpha_0^{-1}\int_{t_1}^{t_2}\Big[&\dot{\psi}(t)(x(t) - z(t)) + \psi(t)(\dot{x}(t) - \dot{z}(t)) \\ &- \lambda(t)\frac{d}{dt}(g_x(z(t), t)(x(t) - z(t)))\Big]\,dt - \alpha_0^{-1}\int_{t_1}^{t_2} g_x(z(t), t)(x(t) - z(t))\,d\lambda(t).\end{aligned} \tag{3.8}$$

Expressing $\dot{\psi}$ by (S1), $\dot{x}$ and $\dot{z}$ by the equations of motion, and noting that $(d/dt)g_x(z(t), t) = p_x(z(t), t) - g_x(z(t), t)f_x(z(t), w(t), t)$ yields

$$\begin{aligned}\Delta\chi_0 \geqq -\alpha_0^{-1}\int_{t_1}^{t_2}&(\psi(t) - \lambda(t)g_x(z(t), t))[f_x(z(t), w(t), t)(x(t) - z(t)) \\ &+ f(z(t), w(t), t) - f(x(t), u(t), t)]\,dt - \alpha_0^{-1}\int_{t_1}^{t_2} g_x(z(t), t)(x(t) - z(t))\,d\lambda(t).\end{aligned} \tag{3.9}$$

From (S5), (S2) and (S8) it is clear that the second integral in (3.9) is nonnegative. By (S5), (S4) and (S9) the same can be said for the first integral. Thus the desired result $\Delta\chi_0 \geqq 0$ is obtained.

It is perhaps worthwhile to emphasize the following facts. Conditions (S1)–(S4) are necessary conditions. Thus if (S5)–(S9) are satisfied automatically through further impositions on the problem data, (S1)–(S4) are necessary and sufficient for optimality. As will be seen in the next section, (S6)–(S9) are satisfied automatically if certain convexity conditions are satisfied. To have (S5) hold automatically, it is necessary to impose a "normality" condition such as: (S1)–(S4) imply (S5). The question of normality in the present context is a difficult one and

will not be pursued further here. In those problems where (2.4) is missing or inactive, $\lambda(t) \equiv 0$ on I, and the resulting simplification of Theorem 3.1 leads easily to generalizations of results obtained by Lee [3].

A number of variations of Theorem 3.1 are apparent. First of all it is clear from (3.9) that (S4) and (S9) may be replaced by the single condition

$$(3.10) \qquad (\psi(t) - \lambda(t)g_x(z(t), t)[f_x(z(t), w(t), t)(x - z(t)) + f(z(t), w(t), t) - f(x, u, t)] \geqq 0$$

for all $t \in I$, $u \in U_t$, and $x \in G$ such that $g(x, t) \leqq 0$.

Perhaps of greater interest is the fact the proof remains unchanged if U_t is replaced by $U_t(x)$, where $U_t(x) \subset U$ for all $t \in I$, $x \in G$. It is only necessary to substitute $U_t(x)$ for U_t in conditions (S4) and (S9) and generalize appropriately the set of admissible control functions $\mathcal{U}$. For state-dependent control constraints of this form it is not clear that the corresponding generalization of Theorem 2.1 is valid. For control constraints of the form $h(x, u, t) \leqq 0$, a new derivation based on a different set of necessary conditions appears to be needed (see [7, p. 136] and [1], [2]).

It is also possible to obtain sufficient conditions for a strong relative minimum in the following sense. Let $G_t \subset G$ for all $t \in I$ be a neighborhood of $z(t)$, where $z(t)$ is a solution corresponding to $w(t) \in \mathcal{U}$. Then (definition) the pair w, z is a strong relative minimum if for any control-solution pair $u(t) \in \mathcal{U}$, $x(t)$ satisfying $x(t) \in G_t$ for all $t \in I$ and (2.2), (2.3) and (2.4), it follows that $\chi_0(x(t_1), x(t_2)) - \chi_0(z(t_1), z(t_2)) \geqq 0$. Sufficient conditions for a strong relative minimum are obtained by replacing $x_1 \in G$, $x_2 \in G$ in (S6) and (S7) by $x_1 \in G_{t_1}$, $x_2 \in G_{t_2}$ and replacing $x \in G$ in (S8) and (S9) by $x \in G_t$.

The conditions (S1)–(S9) are quite simple to apply because they do not involve comparing $w(t)$, $z(t)$ with other admissible control-solution pairs $u(t)$, $x(t)$. If such comparisons can be carried out, it is possible to weaken (S4), (S8) and (S9) to corresponding conditions on (2.13) and the integrals of the left-hand members of (3.3) and (3.4) (make the following substitutions: $x = x(t), x_1 = x(t_1)$ and $x_2 = x(t_2)$).

4. Some corollaries of Theorem 3.1. In this section additional hypotheses are imposed on the problem data so that conditions (S6)–(S9) need not be verified explicitly.

COROLLARY 4.1. *Let G, χ_i, g and f satisfy the following conditions:*

(C1) *G is convex.*

(C2) *χ_i is a convex function on $G \times G$ for $i = -\mu, \cdots, 0$.*

(C3) *χ_i is an affine function on $G \times G$ for $i = 1, \cdots, m$.*

(C4) *$g(\cdot, t)$ is convex on G for each $t \in I$.*

(C5) *f is affine in x, i.e.,*

$$(4.1) \qquad f(x, u, t) = A(t)x + F(u, t).$$

Then Theorem 3.1 *is true if* (S1)–(S9) *are replaced by* (S1), (S2), (S3′), (S4) *and* (S5), where (S3′) *is* (N3) *with requirements* (2.8) *and* (2.10) *omitted.*

Proof. Conditions (S6)–(S8) follow directly from the convexity conditions (C1)–(C4) (for (S7) condition (2.9) must be noted). Condition (C5) implies (S9) holds with (3.4) as an equality.

It is also possible to prove a sufficiency theorem similar to Corollary 4.1, where the (convex) terminal set $X = \{(x_1, x_2) \in G \times G : \chi_i(x_1, x_2) \leqq 0, i = -\mu, \cdots, -1; \chi_i(x_1, x_2) = 0, i = 1, \cdots, m\}$ and (convex) state set $G_t = \{x \in G : g(x, t) \leqq 0\}$ are replaced by quite arbitrary convex sets. To do this many of the arguments used in the proof of Theorem 3.1 are repeated, with (χ_{x_1}, χ_{x_2}) and g_x playing the role of normals to support planes of X and G_t. Lee and Markus [4, p. 346] have obtained sufficiency conditions for a minimum-time problem under similar circumstances.

The corollary is also valid if (C5) is replaced by two conditions: $f(x, u, t) = F_1(x, t) + F_2(u, t)$, where $F_1(\cdot, t)$ is convex on G, and $\psi(t) - \lambda(t)g_x(z(t), t) \leqq 0$ for all $t \in I$. These conditions are more in the nature of those proposed by Mangasarian [5], [6]. However, the condition $\psi(t) - \lambda(t)g_x(z(t), t) \leqq 0$ limits greatly the applicability of the corollary and the sufficient conditions are no longer "nearly" necessary.

For problems with integral cost terms it is possible to let the cost function be convex rather than affine in x. For this purpose the following notation is introduced: $x = (\tilde{x}, \bar{x})$, where $\tilde{x} \in R^1$, $\bar{x} \in R^{n-1}$ and $f = (\hat{f}, \bar{f})$ where $\hat{f}$ is into R^1 and $\bar{f}$ is into R^{n-1}.

COROLLARY 4.2. *Let G, χ_i, g and f satisfy the following additional conditions:*

($\overline{\text{C1}}$) *$G = R^1 \times \bar{G}$, where $\bar{G} \in R^{n-1}$ is convex.*

($\overline{\text{C2}}$) *$\chi_i(x_1, x_2) = \bar{\chi}_i(\bar{x}_1, \bar{x}_2)$, where $\bar{\chi}_i$ is a convex function on $\bar{G} \times \bar{G}$ for $i = -\mu, \cdots, -1$, and $\chi_0(x_1, x_2) = \tilde{x}_2 - \tilde{x}_1 + \bar{\chi}_0(\bar{x}_1, \bar{x}_2)$, where $\bar{\chi}_0$ is a convex function on $\bar{G} \times \bar{G}$.*

($\overline{\text{C3}}$) *$\chi_i(x) = \bar{\chi}_i(\bar{x})$, where $\bar{\chi}_i$ is an affine function on $\bar{G} \times \bar{G}$ for $i = 1, \cdots, m$.*

($\overline{\text{C4}}$) *$g(x, t) = \bar{g}(\bar{x}, t)$, where $\bar{g}(\cdot, t)$ is convex on $\bar{G}$ for each $t \in I$.*

($\overline{\text{C5}}$) *$\bar{f}$ is affine in $\bar{x}$, i.e.,*

$$\bar{f} = \bar{A}(t)\bar{x} + \bar{F}(u, t). \tag{4.2}$$

($\overline{\text{C6}}$) *$\hat{f} = \hat{f}(\bar{x}, t) + \hat{F}(u, t)$, where $\hat{f}(\cdot, t)$ is convex on $\bar{G}$ for each $t \in I$.*

Then Theorem 3.1 *is true if* (S1)–(S9) *are replaced by* (S1), (S2), (S3′), (S4) *and* (S5), *where* (S3′) *is* (N3) *with requirements* (2.8) *and* (2.10) *omitted.*

Proof. The proof is the same as Corollary 4.1 except ($\overline{\text{C6}}$) and (S5) must be added to ($\overline{\text{C5}}$) to obtain (S9).

Suitably modified, the remarks following Corollary 4.1 concerning X, G_t and f apply in the present context.

L. W. Neustadt has shown the authors results similar to those of Corollaries 4.1 and 4.2. These results were obtained by considering sufficient conditions for optimality in the abstract optimization problem treated in [8].

REFERENCES

[1] T. GUINN, *Weakened hypotheses for the variational problem considered by Hestenes,* this Journal, 3 (1965), pp. 418–423.

[2] M. R. HESTENES, *On variational theory and optimal control theory*, this Journal, 3 (1965), pp. 23–48.

[3] E. B. LEE, *A sufficient condition in the theory of optimal control*, this Journal, 1 (1963), pp. 241–245.

[4] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[5] O. L. MANGASARIAN, *Sufficient conditions for the optimal control of nonlinear systems,* this Journal, 4 (1966), pp. 139–152.

[6] O. L. MANGASARIAN AND L. L. SCHUMAKER, *Splines via optimal control,* Approximations, with Special Emphasis on Spline Functions (Conference, University of Wisconsin, Madison, 1969), Academic Press, New York, 1969.

[7] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems. II: Applications,* this Journal, 5 (1967), pp. 90–137.
[8] ———, *A general theory of extremals,* J. Comput. System Sci., 3 (1969), pp. 57–92.
[9] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes,* Interscience, New York, 1962.
[10] J. WARGA, *Minimizing variational curves restricted to a preassigned set,* Trans. Amer. Math. Soc., 112 (1964), pp. 432–455.
[11] ———, *Unilateral variational problems with several inequalities,* Michigan Math. J., 12 (1965), pp. 449–480.

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

ASYMPTOTIC KUHN–TUCKER CONDITIONS FOR MATHEMATICAL PROGRAMMING PROBLEMS IN A BANACH SPACE*

SANJO ZLOBEC†

1. Introduction. The Kuhn–Tucker optimality conditions were recently formulated by Guignard [3] for mathematical programming problems in a Banach space. The purpose of this paper is to give the asymptotic extensions of Theorems 2 and 3 in [3]. These asymptotic extensions can be applied in some situations where Guignard's theorems are not applicable, e.g., Example 1. Guignard's necessity conditions are already asymptotic in nature and will not be treated in this paper, which is restricted to sufficiency conditions.

2. Notations and preliminaries. The following notations will be used:

R^n is the *n-dimensional real Euclidean space*,
$R^{m\times n}$ the *space of* $m \times n$ *real matrices.*

For any $Z \in R^{m\times n}$,

Z^T is the *transpose of* Z,
$R\{Z\}$ the *range space of* Z,
$N\{Z\}$ the *null space of* Z,
$Z^\dagger$ the *Moore-Penrose generalized inverse of* Z, e.g., [8],
X, Y *real Banach spaces*,
$\alpha: X \to Y$ a *mapping from* X *into* Y,
$\nabla\alpha(\bar{x})$ the *Fréchet derivative of* α *at* $\bar{x}$,
$l(x) = \langle l, x\rangle$ the *value of* l *at* x, if $l: X \to Y$ is a continuous linear mapping from X into Y,
$\langle \nabla^2\alpha(\bar{x}), (x, w)\rangle$ the *value of the second Fréchet derivative* $\nabla^2\alpha(\bar{x})$ at $(x, w) \in X \times X$,
$\overline{M}$ the *closure of the set* M,
$\beta(\bar{x}, r)$ a *closed ball with the center at* $\bar{x}$ *and radius* $r > 0$,
$M \setminus N$ $= \{x \in X : x \in M, x \notin N\}$,
X^* the *dual space of* X, e.g., [6, p. 106].

A set C is a *cone* if $x \in C$ and $\lambda \geqq 0$ imply $\lambda x \in C$. If $C \subset X$ is a cone, then

$$C^- = \{u \in X^* : \langle u, x\rangle \leqq 0 \text{ for all } x \in C\},$$
$$C^+ = \{u \in X^* : \langle u, x\rangle \geqq 0 \text{ for all } x \in C\}.$$

The following concepts from [3] will also be used:

Let $\bar{x} \in M \subset X$. Then the vector w is *tangent to* M *at* $\bar{x}$ if there exist a sequence $\{x^k\}$ in M converging to $\bar{x}$ and a sequence $\{\lambda_k\}$ on nonnegative numbers such that the sequence $\{\lambda_k(x^k - \bar{x})\}$ converges to w. The set $T(M, \bar{x})$ of all the vectors tangent to M at $\bar{x}$ is the *cone tangent to* M *at* $\bar{x}$. The set $P(M, \bar{x})$, the closure of the convex hull of $T(M, \bar{x})$, is the *cone pseudotangent to* M *at* $\bar{x}$. M is *pseudoconvex at* $\bar{x}$ if $x - \bar{x} \in P(M, \bar{x})$ for all $x \in M$.

* Received by the editors December 30, 1969, and in revised form April 15, 1970.

† Department of Engineering Sciences, Northwestern University, Evanston, Illinois 60201. Now at Department of Mathematics, McGill University, Montreal 110, Canada. This work was supported in part by the National Science Foundation under Project GP 13546 and by the National Research Council of Canada under Grant A7692.

Let $\psi(x)$ be a real function of $x \in X$. Then ψ is *pseudoconcave over M at $\bar{x}$* if ψ is Fréchet-differentiable at $\bar{x}$ and if $x \in M$, $\langle \nabla\psi(\bar{x}), x - \bar{x} \rangle \leqq 0$ implies $\psi(x) - \psi(\bar{x}) \leqq 0$. ψ is *quasiconcave* if the set $\{x \in X : \psi(x) \geqq \lambda\}$ is convex for all $\lambda \in R^1$.

We are concerned with the mathematical programming problem

$$\text{maximize } \{\psi(x) : \alpha(x) \in B, x \in C\},$$

where ψ is a real function of $x \in X$ and B and C are nonempty subsets in Y and X, respectively. We denote by A and Δ,

$$A = \{x \in C : \alpha(x) \in B\}, \qquad \Delta = \{x \in X : \alpha(x) \in B\}.$$

The subset A is assumed to be nonempty. All our results will be of a local character and stated at an arbitrary but fixed point $\bar{x} \in A$.

3. Results.

THEOREM 1. *Assumptions.*

(i) *ψ is continuous.*

(ii) *G is a closed convex cone in X such that $x - \bar{x} \in G$ for all $x \in A$.*

(iii) *A or Δ is pseudoconvex at $\bar{x}$.*

(iv) *ψ is pseudoconcave over A at $\bar{x}$ or quasiconcave with $\nabla\psi(\bar{x}) \neq 0$.*

(v) *There exist a sequence $\{u^i\}$, $u^i \in P^+[B, \alpha(\bar{x})]$, and some $g^- \in G^-$ such that*

$$\lim_{i \to \infty} \langle \nabla\psi(\bar{x}) + u^i \cdot \nabla\alpha(\bar{x}), p \rangle = \langle g^-, p \rangle \qquad \text{for all } p \in P(\Delta, \bar{x}).$$

Conclusion. $\bar{x}$ maximizes ψ over A.

Proof. The parts of the proof which are the same as in [3] will be omitted. Take an $x \in A$. Then $x - \bar{x} \in P(\Delta, \bar{x})$ by (iii). Since $\langle \nabla\alpha(\bar{x}), y \rangle \in P[B, \alpha(\bar{x})]$ for all $y \in P(\Delta, \bar{x})$, e.g., [3, p. 234, proof of Lemma], and $u^i \in P^+[B, \alpha(\bar{x})]$, we have

$$\langle u^i \cdot \nabla\alpha(\bar{x}), x - \bar{x} \rangle \geqq 0.$$

Therefore,

$$\langle \nabla\psi(\bar{x}) + u^i \cdot \nabla\alpha(\bar{x}), x - \bar{x} \rangle - \langle \nabla\psi(\bar{x}), x - \bar{x} \rangle = \langle u^i \cdot \nabla\alpha(\bar{x}), x - \bar{x} \rangle \geqq 0,$$

i.e.,

$$\langle \nabla\psi(\bar{x}) + u^i \cdot \nabla\alpha(\bar{x}), x - \bar{x} \rangle \geqq \langle \nabla\psi(\bar{x}), x - \bar{x} \rangle.$$

Since

$$\lim_{i \to \infty} \langle \nabla\psi(\bar{x}) + u^i \cdot \nabla\alpha(\bar{x}), x - \bar{x} \rangle = \langle g^-, x - \bar{x} \rangle \leqq 0,$$

by (v) and (ii), we conclude that $\langle \nabla\psi(\bar{x}), x - \bar{x} \rangle \leqq 0$. This implies $\psi(x) \leqq \psi(\bar{x})$, by the sufficiency proof of Theorem 2 in [3].

COROLLARY 1 (Sufficiency part of Theorem 2 in [3]).

Assumptions.

Assumptions (i), (ii), (iii) *and* (iv) *are as in Theorem* 1.

(v′) *There exists $u \in P^+[B, \alpha(\bar{x})]$ such that*

$$\nabla\psi(\bar{x}) + u \cdot \nabla\alpha(\bar{x}) \in G^-.$$

Conclusion. $\bar{x}$ *maximizes* ψ *over* A.

In order to get an idea of why Theorem 1 is more general than Corollary 1, we are going to characterize the assumptions (v) and (v′) in the finite-dimensional case with $G \equiv X$. The Fréchet derivative $\nabla\alpha(\bar{x})$ is now the Jacobian matrix.

LEMMA 1. *Let* $\nabla\alpha(\bar{x}) \in R^{m\times n}$, $\nabla\psi(\bar{x}) \in R^n$ *and* $P^+[B, \alpha(\bar{x})] \subset R^m$ *be as above. Then the following are equivalent:*

(a) *There exists a sequence* $\{u^i\}, u^i \in P^+[B\alpha(\bar{x})]$, *such that*

$$\lim_{i\to\infty} [\nabla\psi(\bar{x}) + \nabla^T\alpha(\bar{x}) \cdot u^i] = 0.$$

(b) $\nabla\alpha(\bar{x}) \cdot y \in P[B, \alpha(\bar{x})]$ *implies* $\nabla^T\psi(\bar{x}) \cdot y \leqq 0$.

(c) $\nabla\psi(\bar{x}) \in R\{\nabla^T\alpha(\bar{x})\}$ *and* $-[\nabla^T\alpha(\bar{x})]^\dagger\nabla\psi(\bar{x}) \in \overline{N\{\nabla^T\alpha(\bar{x})\} + P^+[B, \alpha(\bar{x})]}$.

Proof. We use the fact that $P^+[B, \alpha(\bar{x})]$ is a closed convex cone and $P[B, \alpha(\bar{x})] = P^{+\,+}[B, \alpha(\bar{x})]$. Now apply Theorem 2.2 from [1].

LEMMA 2. *Let* $\nabla\alpha(\bar{x}) \in R^{m\times n}$, $\nabla\psi(\bar{x}) \in R^n$ *and* $P^+[B, \alpha(\bar{x})] \subset R^m$ *be as above. Then the following are equivalent:*

(a′) *The system*

$$\nabla\psi(\bar{x}) + \nabla^T\alpha(\bar{x}) \cdot u = 0, \qquad u \in P^+[B, \alpha(\bar{x})],$$

is consistent.

(c′) $\nabla\psi(\bar{x}) \in R\{\nabla^T\alpha(\bar{x})\}$ *and* $-[\nabla^T\alpha(\bar{x})]^\dagger\nabla\psi(\bar{x}) \in N\{\nabla^T\alpha(\bar{x})\} + P^+[B, \alpha(\bar{x})]$.

Proof. This is Lemma 2.3 in [1].

By comparison of (c) and (c′) we conclude that Theorem 1, even with the uniform limit in (v), is more general than Corollary 1, if $N\{\nabla^T\alpha(\bar{x})\} + P^+[B, \alpha(\bar{x})]$ is not a closed cone.

In order to formulate the asymptotic second order optimality conditions we assume that X is finite-dimensional. This assures the compactness of the unit sphere, which is needed in the proof.

THEOREM 2. *Assumptions.*

(i) ψ *and* α *are twice continuously differentiable at* $\bar{x}$.

(ii) G *is a closed convex cone in* X.

(iii) *If* $x \in A \cap \beta(\bar{x}, r)$, *then* $x - \bar{x} \in G$.

(iv) *If* $y \in B \cap \beta[\alpha(\bar{x}), r]$, *then* $y - \alpha(\bar{x}) \in P[B, \alpha(\bar{x})]$.

(v) *There exists a sequence* $\{u^i\}, u^i \in P^+[B, \alpha(\bar{x})]$, *and some* $g^- \in G^-$, *such that*

$$\lim_{i\to\infty} \langle \nabla\psi(\bar{x}) + u^i \cdot \nabla\alpha(\bar{x}), g\rangle = \langle g^-, g\rangle \qquad \textit{for all } g \in G.$$

(vi) *For every* $\beta(\bar{x}, r)$ *the set* $\{A \cap \beta(\bar{x}, r)\} \setminus \bar{x}$ *is nonempty. For every* $w \in \overline{W}$, *where*

$$W = \left\{ w : w = \frac{z - \bar{x}}{\|z - \bar{x}\|}, z \in \{A \cap \beta(\bar{x}, r)\} \setminus \bar{x} \right\}$$

for some $\beta(\bar{x}, r)$, *it follows that*

$$\lim_{i\to\infty} \langle \nabla^2\psi(\bar{x}) + u^i \cdot \nabla^2\alpha(\bar{x}), (w, w)\rangle = L(w)$$

exists. Further, if $\lim_{k\to\infty} w^k = h$, *where* $w^k \in W$, *then* $\lim_{k\to\infty} L(w^k) = L(h)$. *The sequence* $\{u^i\}$ *here is the one used in* (v).

(vii) *For every nontrivial* $h \in X$ *such that*

$$\lim_{i\to\infty} \langle u^i \cdot \nabla\alpha(\bar{x}), h\rangle = 0 \quad \textit{and} \quad \langle \nabla\alpha(\bar{x}), h\rangle \in P[B, \alpha(\bar{x})]^{\setminus} - P[B, \alpha(\bar{x})],$$

and for every nontrivial $h \in X$ *such that*

$$\langle \nabla\psi(\bar{x}), h\rangle = 0 \quad \textit{and} \quad \langle \nabla\alpha(\bar{x}), h\rangle \in -P[B, \alpha(\bar{x})] \cap P[B, \alpha(\bar{x})],$$

it follows that

$$\lim_{i\to\infty} \langle \nabla^2\psi(\bar{x}) + u^i \cdot \nabla^2\alpha(\bar{x}), (h, h)\rangle < 0.$$

The sequence $\{u^i\}$ *here is the one used in* (v).

Conclusion. $\bar{x}$ *is an isolated local maximum for* ψ *over* A.

Proof. Let us suppose that $\bar{x}$ is not an isolated local maximum for ψ over A. Then there exists a sequence $\{x^k\}$, $x^k \in A$, $x^k \neq \bar{x}$ for all k, such that $\lim_{k\to\infty} x^k = \bar{x}$, $\psi(x^k) \geqq \psi(\bar{x})$ for all k. Since X is finite-dimensional, the unit sphere in X is compact and therefore we may assume that

$$\lim_{k\to\infty} \frac{x^k - \bar{x}}{\|x^k - \bar{x}\|} = h \neq 0.$$

For x^k close to $\bar{x}$ we have $x^k - \bar{x} \in G$, by (iii). Therefore,

$$\lim_{i\to\infty} \langle \nabla\psi(\bar{x}) + u^i \cdot \nabla\alpha(\bar{x}), x^k - \bar{x}\rangle = \langle g^-, x^k - \bar{x}\rangle \leqq 0,$$

by (v), and

$$\lim_{i\to\infty} \langle \nabla\psi(\bar{x}) + u^i \cdot \nabla\alpha(\bar{x}), h\rangle = \langle g^-, h\rangle \leqq 0,$$

by continuity of g^-. Also $\langle \nabla\psi(\bar{x}), h\rangle \geqq 0$ and $\langle \nabla\alpha(\bar{x}), h\rangle \in P[B, \alpha(\bar{x})]$, e.g., [3, p. 236]. We will consider two possible cases and show that a contradiction arises from each of them.

(a) Suppose that

$$\langle \nabla\alpha(\bar{x}), h\rangle \in P[B, \alpha(\bar{x})]^{\setminus} - P[B, \alpha(\bar{x})].$$

If $\lim_{i\to\infty} \langle u^i \cdot \nabla\alpha(\bar{x}), h\rangle \neq 0$, then it is positive and

$$\langle \nabla\psi(\bar{x}), h\rangle \leqq -\lim_{i\to\infty} \langle u^i \cdot \nabla\alpha(\bar{x}), h\rangle < 0.$$

This is impossible. Therefore a contradiction must be obtained only for

$$\lim_{i\to\infty} \langle u^i \cdot \nabla\alpha(\bar{x}), h\rangle = 0.$$

(b) Suppose that

$$\langle \nabla\alpha(\bar{x}), h\rangle \in -P[B, \alpha(\bar{x})] \cap P[B, \alpha(\bar{x})].$$

If $\langle \nabla\psi(\bar{x}), h\rangle \neq 0$, then it is positive. Since $\langle u^i \nabla\alpha(\bar{x}), h\rangle = 0$, we have

$$0 \geqq \lim_{i\to\infty} \langle \nabla\psi(\bar{x}) + u^i \cdot \nabla\alpha(\bar{x}), h\rangle = \langle \nabla\psi(\bar{x}), h\rangle.$$

This is impossible. Therefore we have to contradict only $\langle \nabla\psi(\bar{x}), h\rangle = 0$, in this case.

Let us define

$$\xi^i(x) = \psi(x) + u^i \cdot \alpha(x).$$

We have

$$\xi^i(x^k) - \xi^i(\bar{x}) = \langle \nabla\xi^i(\bar{x}), x^k - \bar{x}\rangle + \tfrac{1}{2}\langle \nabla^2\xi^i(\bar{x}), (x^k - \bar{x}, x^k - \bar{x})\rangle + O(\|x^k - \bar{x}\|^2)$$

for all i. Now $\lim_{i\to\infty}\langle \nabla\xi^i(\bar{x}), x^k - \bar{x}\rangle \leqq 0$, by (v).

$$\lim_{i\to\infty} \left\langle \nabla^2\xi^i(\bar{x}), \left(\frac{x^k - \bar{x}}{\|x^k - \bar{x}\|}, \frac{x^k - \bar{x}}{\|x^k - \bar{x}\|} \right) \right\rangle$$

exists for all $k \geqq k_0$, where k_0 is some positive integer, by (vi),

$$x^k \in \{A \cap \beta(\bar{x}, r)\} \setminus \bar{x} \quad \text{and} \quad \frac{x^k - \bar{x}}{\|x^k - \bar{x}\|} = w^k \in W.$$

$$\lim_{k\to\infty}\lim_{i\to\infty} \left\langle \nabla^2\xi^i(\bar{x}), \left(\frac{x^k - \bar{x}}{\|x^k - \bar{x}\|}, \frac{x^k - \bar{x}}{\|x^k - \bar{x}\|} \right) \right\rangle = \lim_{i\to\infty} \langle \nabla^2\xi^i(\bar{x}), (h, h)\rangle,$$

by the continuity assumption in (vi) and $h \in \overline{W}$. $\lim_{i\to\infty}\langle \nabla^2\xi^i(\bar{x}), (h, h)\rangle < 0$, by (vii). Therefore,

$$\lim_{k\to\infty}\lim_{i\to\infty} \frac{\xi^i(x^k) - \xi^i(\bar{x})}{\|x^k - \bar{x}\|^2} < 0.$$

This implies

$$\lim_{i\to\infty} [\xi^i(x^k) - \xi^i(\bar{x})] < 0$$

for all $k \geqq k_0$, where k_0 is some positive integer. But

$$\lim_{i\to\infty} \langle u^i, \alpha(\bar{x}) - \alpha(x^k)\rangle \leqq 0,$$

by (iv) and continuity of u^i. Thus $\psi(x^k) < \psi(\bar{x})$, which contradicts the assumption on $\{x^k\}$.

COROLLARY 2 (Theorem 3 in [3]).

Assumptions.

(i), (ii), (iii) *and* (iv) *are as in Theorem* 2.

(v) *There exists* $u \in P^+[B, \alpha(\bar{x})]$ *such that*

$$\nabla\psi(\bar{x}) + u \cdot \nabla\alpha(\bar{x}) \in G^-.$$

(vi) *For all nontrivial* $h \in X$ *such that*

$$\langle \nabla\psi(\bar{x}), h\rangle = 0 \quad \text{and} \quad \langle \nabla\alpha(\bar{x}), h\rangle \in -P[B, \alpha(\bar{x})] \cap P[B, \alpha(\bar{x})],$$

it follows that

$$\langle \nabla^2 \psi(\bar{x}) + u \cdot \nabla^2 \alpha(\bar{x}), (h, h) \rangle < 0.$$

Conclusion. $\bar{x}$ *is an isolated maximum for* ψ *over* A.

4. Examples.

Example 1. Maximize $-x_1 - x_2^2$ subject to

$$\begin{pmatrix} -x_1^2 \\ x_1 \\ x_2 \end{pmatrix} \in B,$$

where

$$B = \left\{ \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} : 2y_1 y_3 \geqq y_2^2,\ y_1 \geqq 0,\ y_3 \geqq 0 \right\}.$$

B is an "ice cream" cone in R^3, which consists of all vectors forming an angle $\leqq 45°$ with the vector

$$\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}.$$

Let $\bar{x} = 0$ and $C \equiv G \equiv R^2$. Therefore $G^- = \{0\}$. $\psi = -x_1 - x_2^2$ is pseudoconcave over

$$A = \left\{ \begin{pmatrix} 0 \\ x_2 \end{pmatrix} : x_2 \geqq 0 \right\}$$

at $\bar{x}$. $A \equiv \Delta$ is pseudoconvex at $\bar{x}$. Here $B^+ = B$, and therefore $P^+[B, \alpha(\bar{x})] = P^+[B, 0] = P[B, 0] = B$. Further,

$$\nabla \psi(\bar{x}) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \qquad \nabla \alpha(\bar{x}) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Note that $N\{\nabla^T \alpha(\bar{x})\} + B$ is not closed.

First order sufficiency conditions. There exists no u satisfying

$$\nabla \psi(\bar{x}) + \nabla^T \alpha(\bar{x}) \cdot u = 0 \tag{1}$$

and lying in B. The vectors u which satisfy (1) are of the form

$$u = \begin{pmatrix} u_1 \\ 1 \\ 0 \end{pmatrix}, \qquad u_1 \text{ arbitrary},$$

and obviously are not in B. Therefore Corollary 1 cannot be applied here.

Since the sequence

$$u^i = \begin{pmatrix} i \\ 1 \\ 1/(2i) \end{pmatrix}, \qquad i = 1, 2, \cdots \tag{2}$$

is in B, and

$$\lim_{i \to \infty} [\nabla \psi(\bar{x}) + \nabla^T \alpha(\bar{x}) \cdot u^i] = \lim_{i \to \infty} \begin{pmatrix} 0 \\ 1/(2i) \end{pmatrix} = 0.$$

Theorem 1 is applicable and we conclude that $\bar{x} = 0$ is a maximizing point.

Second order sufficiency conditions. Since there exists no u in B satisfying (1), the assumption (v) in Corollary 2 is not satisfied. Therefore Corollary 2 cannot be applied here.

Take the sequence $\{u^i\}$, defined by (2). Then (v), in Theorem 2, is satisfied. For the closed ball $\beta(0, 1)$, the set W consists of the single point

$$w = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The limit

$$\lim_{i \to \infty} \langle \nabla^2 \psi(\bar{x}) + u^i \cdot \nabla^2 \alpha(\bar{x}), (w, w) \rangle$$

$$= \lim_{i \to \infty} \left\langle \begin{pmatrix} -2i & 0 \\ 0 & -2 \end{pmatrix}, \left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \right\rangle = -2$$

exists. Thus (vi) in Theorem 2 is satisfied, too.

All $h \in R^2$, such that

$$\lim_{i \to \infty} \langle u^i \cdot \nabla \alpha(\bar{x}), h \rangle = 0 \quad \text{and} \quad \langle \nabla \alpha(\bar{x}), h \rangle \in P[B, \alpha(\bar{x})] \setminus - P[B, \alpha(\bar{x})],$$

are of the form

$$h = \begin{pmatrix} 0 \\ h_2 \end{pmatrix}, \qquad h_2 > 0.$$

Since

$$\lim_{i \to \infty} \langle \nabla^2 \psi(\bar{x}) + u^i \cdot \nabla^2 \alpha(\bar{x}), (h, h) \rangle = -2h_2^2,$$

and the second requirement in (vii) is here redundant, we have all the assumptions in Theorem 2 satisfied. Theorem 2 is therefore applicable, and we conclude that $\bar{x} = 0$ is an isolated maximizing point.

Example 2. The purpose of this example is to show that one can have $\nabla \alpha(\bar{x})$ with closed range and $H = \{h \in X^* : h = u \cdot \nabla \alpha(\bar{x}),\ u \in P^-[B, \alpha(\bar{x})]\}$ not closed. This is a counterexample to Remark 3 in [3, p. 234].

Consider again the programming problem from Example 1. The Fréchet derivative $\nabla\alpha(\bar{x})$ at the point $\bar{x} = 0$ is

$$\nabla\alpha(\bar{x}) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and it has closed range. Take

$$u^i = \begin{pmatrix} -i \\ -1 \\ -1/(2i) \end{pmatrix}, \qquad i = 1, 2, \cdots.$$

Since here $P[B, \alpha(\bar{x})] = P[B, 0] = B$, and $u^i \in B^-$, we conclude that $u^i \in P^-[B, \alpha(\bar{x})]$. The sequence

$$h^i = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -i \\ -1 \\ -1/(2i) \end{pmatrix} = \begin{pmatrix} -1 \\ -1/(2i) \end{pmatrix}, \qquad i = 1, 2, \cdots$$

is in H, but its limit

$$\lim_{i \to \infty} h^i = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

is not in H. Therefore H is not closed.

Acknowledgment. The author is indebted to Professor A. Ben-Israel for suggestions and helpful discussions. This paper is part of the author's doctoral dissertation at Northwestern University.

REFERENCES

[1] A. Ben-Israel, *Linear equations and inequalities on finite dimensional, real or complex, vector spaces: A unified theory*, J. Math. Anal. Appl., 27 (1969), pp. 367–389.

[2] A. V. Fiacco and G. P. McCormick, *Asymptotic conditions for constrained minimization*, Technical Paper RAC-TP-340, Research Analysis Corporation, McLean, Virginia, 1968.

[3] M. Guignard, *Generalized Kuhn–Tucker conditions for mathematical programming problems in a Banach space*, this Journal, 7 (1969), pp. 232–241.

[4] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[5] K. O. Kortanek and J. P. Evans, *Asymptotic Lagrange regularity for pseudoconcave programming with weak constraint qualification*, Operations Res., 16 (1968), pp. 849–857.

[6] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[7] O. L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[8] R. Penrose, *A generalized inverse for matrices*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406–413.

[9] K. Yosida, *Functional Analysis*, Springer-Verlag, New York, 1965.

[10] S. Zlobec, *Note on generalized Kuhn–Tucker conditions in mathematical programming*, Rep. 70-10, Series in Applied Mathematics, Northwestern University, Evanston, Illinois, 1970.

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

THE MINIMAL BOUND ON THE ESTIMATION ERROR COVARIANCE MATRIX IN THE PRESENCE OF CORRELATED DRIVING NOISE*

JAMES E. POTTER† AND JAMES C. DECKERT‡

Abstract. The application of optimal linear filter theory to situations in which the state forcing function is correlated with the state in an unknown way can present serious problems. In many instances, the cross-correlation and forcing function must be modeled, for if they are ignored the filter gains tend to sink below their optimal levels and useful measurement information is discarded. This paper presents a conservative and minimal formula for bounding cross-correlation between a random forcing function and the state error when this correlation is unknown. The bound is conservative in the sense that its use always results in overestimating the estimation error covariance, and it is minimal in the sense that given any conservative cross-correlation estimate, a bound of the minimal form can always be found which is no more conservative than the given estimate.

When this minimal bound is used to approximate the differential equation for the estimation error covariance matrix, there remains the problem of finding the free parameter associated with the minimal bound. This paper presents a noniterative expression for this parameter as the solution to an optimal control problem in which the cost function is a linear combination of the elements of the covariance matrix at a final time of interest. Simulation results are given for a satellite in orbit around a model earth.

Introduction. Because of its computational simplicity, it is of considerable practical interest to apply recursive filtering with the same form as the Kalman filter [2] to problems in which the random forcing function driving the state is not white, but has correlation between its values at different times. In many situations, the mean square value of the forcing function can be estimated with some accuracy, but its frequency spectrum is poorly known. For example, although the spatial distribution of the deflection of the vertical at the earth's surface is not known, the range of deflections is well documented. Thus, in the absence of more accurate information, its seems feasible to conservatively approximate the covariance of the forcing function with a diagonal matrix whose elements are the squares of the bounds of the individual elements of the forcing function. Attention is focused in this paper on the problem of approximating between discrete measurements the increase in estimation error covariance due to uncertainty introduced by the random forcing function under these conditions.

1. Problem formulation. The problem is as follows: Given the n-dimensional state equation

$$\dot{\mathbf{x}} = F\mathbf{x} + \mathbf{f},$$

where $\mathbf{f}$ is a random forcing function. The estimate of the state, $\hat{\mathbf{x}}$, is extrapolated in the usual way:

$$\dot{\hat{\mathbf{x}}} = F\hat{\mathbf{x}}.$$

* Received by the editors September 30, 1969, and in revised form February 24, 1970. This work was supported by the National Aeronautics and Space Administration under Contract NAS-9-9024.

† Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

‡ Charles Stark Draper Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

Thus the error in the state estimate, $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$, propagates by

$$\dot{\mathbf{e}} = F\mathbf{e} + \mathbf{f}.$$

Forming the covariance matrix of state estimation errors yields

$$P = \overline{\mathbf{e}\mathbf{e}^T},$$

$$\dot{P} = \overline{\mathbf{e}\dot{\mathbf{e}}^T} + \overline{\dot{\mathbf{e}}\mathbf{e}^T} = FP + PF^T + \overline{\mathbf{e}\mathbf{f}^T} + \overline{\mathbf{f}\mathbf{e}^T},$$

where the overbar indicates ensemble average. Since the cross-correlations of $\mathbf{e}$ and $\mathbf{f}$ are often unavailable in practice, it is desired to choose some matrix upper bound $M \geqq (\overline{\mathbf{e}\mathbf{f}^T} + \overline{\mathbf{f}\mathbf{e}^T})$, i.e., $M - (\overline{\mathbf{e}\mathbf{f}^T} + \overline{\mathbf{f}\mathbf{e}^T})$ is positive semidefinite, and propagate the covariance matrix using the equation

$$\dot{P} = FP + PF^T + M.$$

This will ensure that the approximated covariance matrix will be greater than the actual covariance matrix, resulting in a conservative error analysis and high filter gains corresponding to the actual system errors. It is also desirable that it be demonstrated that there is no other matrix M' which satisfies the equations

$$M' \geqq (\overline{\mathbf{e}\mathbf{f}^T} + \overline{\mathbf{f}\mathbf{e}^T}), \qquad M' \leqq M,$$

for this would indicate that the approximated covariance matrix is overly large. Furthermore, the only information to be employed in calculating M is the covariance $\overline{\mathbf{f}\mathbf{f}^T}$ of $\mathbf{f}$, it being assumed that the cross-correlation of $\mathbf{e}$ and $\mathbf{f}$ is completely unknown.

This paper demonstrates that in order to satisfy the above conditions M must have the form

$$M(\lambda) = \lambda P + \overline{\mathbf{f}\mathbf{f}^T}/\lambda$$

with $\lambda > 0$ and lying between the square roots of the maximum and minimum eigenvalues of the matrix $P^{-1}\overline{\mathbf{f}\mathbf{f}^T}$. Pertinent definitions and conventions will now be presented, followed by the theorem statements and a discussion of their importance. Two final sections describe a method of choosing λ and simulation results.

2. Conventions. The shorthand matrix equation for $n \times n$ matrices A and B

$$A \geqq B$$

will be used to indicate that $A - B$ is positive semidefinite, i.e., that the scalar equation

$$\mathbf{x}^T A\mathbf{x} \geqq \mathbf{x}^T B\mathbf{x}$$

holds for all n-vectors $\mathbf{x}$. The $n \times n$ identity matrix will be denoted by I.

3. Definitions.

DEFINITION 1. The covariance of the state estimation error, $\overline{\mathbf{e}\mathbf{e}^T}$, will be denoted by P, $P \geqq 0$, $P \neq 0$.

DEFINITION 2. The covariance of the driving noise, $\overline{\mathbf{f}\mathbf{f}^T}$, will be denoted by Q, $Q \geqq 0$, $Q \neq 0$.

DEFINITION 3. The matrix $\lambda P + Q/\lambda$ will be called $M(\lambda)$.

DEFINITION 4. A symmetric matrix C will be called *conservative* if $C \geqq \overline{\mathbf{e}\mathbf{f}^T} + \overline{\mathbf{f}\mathbf{e}^T}$ for all random variables $\mathbf{e}$ and $\mathbf{f}$ satisfying $\overline{\mathbf{e}\mathbf{e}^T} = P$ and $\overline{\mathbf{f}\mathbf{f}^T} = Q$ with P and Q given.

DEFINITION 5. A symmetric matrix M will be called *minimal* if it is conservative and if there is no other conservative matrix M' such that $M' \leqq M$.

DEFINITION 6. $L(C, \mathbf{x})$ is defined as the scalar quantity

$$L(C, \mathbf{x}) = \frac{\mathbf{x}^T C \mathbf{x}}{2((\mathbf{x}^T P \mathbf{x})(\mathbf{x}^T Q \mathbf{x}))^{1/2}}.$$

DEFINITION 7. $L(C)$ is defined as the infimum or greatest lower bound of $L(C, \mathbf{x})$ given C, over the values of $\mathbf{x}$ for which the denominator is nonzero.

DEFINITION 8. T is the set of n-vectors defined by

$$T = \{\mathbf{x} | \mathbf{x}^T P \mathbf{x} > 0,\ \mathbf{x}^T Q \mathbf{x} > 0\}.$$

Thus T is the set of vectors for which the denominator of $L(C, \mathbf{x})$ is nonzero. T is nonempty because P and Q are nonzero.

DEFINITION 9. R is the subset of the real line satisfying

$$R = \{\lambda | \lambda = \sqrt{\mathbf{x}^T Q \mathbf{x} / \mathbf{x}^T P \mathbf{x}},\ \mathbf{x} \in T\}.$$

Unfortunately, because the relation "$\geqq$" used to define minimality is a partial ordering (i.e., there are symmetric matrices U and V such that neither $U \geqq V$ nor $V \geqq U$ holds), the minimal bound is not unique. However, a formula for all minimal bounds can be determined so that the one most appropriate to the given problem objective may be employed. In view of Theorem 2, however, a minimal bound can always be found which is at least as good as a given conservative bound, so that the use of a conservative but nonminimal bound is never justified. (It is possible that a nonconservative cross-correlation bound might lead to a conservative estimation error covariance estimate, but this is a much more complex problem.)

THEOREM 1. *The symmetric matrix C is minimal if and only if $C = M(\lambda)$ for some λ in R.*

THEOREM 2. *If the symmetric matrix C is conservative, then there exists a λ in R such that $C \geqq M(\lambda)$.*

If both P and Q are singular matrices, the determination of the set R, while straightforward, is somewhat complicated by the enumeration of special cases, e.g., if P and Q have common null vectors. However, if P is nonsingular, the following simple determination of R results.

THEOREM 3. *If P is nonsingular, it follows that*:

(a) *The eigenvalues of $P^{-1}Q$ are nonnegative.*

(b) *Let a and b be the minimum and maximum eigenvalues of $P^{-1}Q$ respectively. If a is nonzero, R is the closed interval $[\sqrt{a}, \sqrt{b}]$ and if a is zero, R is the half-open interval $(0, \sqrt{b}]$.*

These theorems will be proved as a series of lemmas which are listed below so that the reader can visualize the structure of the argument.

LEMMA 1. *$M(\lambda)$ is conservative if $\lambda > 0$.*

This lemma serves to introduce $M(\lambda)$ and demonstrate its conservative nature.

LEMMA 2. *A symmetric matrix C is conservative if and only if* $L(C) \geqq 1$.

This lemma presents a nonprobabilistic condition which any conservative matrix must satisfy.

LEMMA 3. *For any conservative matrix C there exists an* $\mathbf{x}_0$ *in T such that* $L(C, \mathbf{x}_0) = L(C)$.

This lemma demonstrates that $L(C)$ is a true minimum and not just an infimum on the set T, a result required to assure applicability of Lemma 4.

LEMMA 4. *If C is conservative with* $L(C, \mathbf{x}_0) = L(C)$, *then* $C \geqq M(\lambda)$ *with*

$$\lambda = \sqrt{\mathbf{x}_0^T Q \mathbf{x}_0 / \mathbf{x}_0^T P \mathbf{x}_0}.$$

This lemma together with Lemma 3 implies Theorem 2 as well as showing that only matrices of the form $M(\lambda)$ with λ in R can be minimal. This is a key result.

LEMMA 5. $M(\lambda)$ *is minimal if* λ *is in R.*

This lemma is the "if" portion of Theorem 1. The "only if" portion is proved by Lemmas 3 and 4.

The five lemmas will now be proved, together with Theorem 3. Theorems 1 and 2 follow as indicated above.

LEMMA 1. $M(\lambda)$ *is conservative if* $\lambda > 0$.

Proof. The proof rests on the fact that the covariance matrix of any random vector is positive semidefinite.

$$\begin{aligned} &\overline{(\sqrt{\lambda}\mathbf{e} - \mathbf{f}/\sqrt{\lambda})(\sqrt{\lambda}\mathbf{e} - \mathbf{f}/\sqrt{\lambda})^T} \geqq 0, \\ &\lambda\overline{\mathbf{e}\mathbf{e}^T} - \overline{\mathbf{f}\mathbf{e}^T} - \overline{\mathbf{e}\mathbf{f}^T} + \overline{\mathbf{f}\mathbf{f}^T}/\lambda \geqq 0, \\ &\lambda\overline{\mathbf{e}\mathbf{e}^T} + \overline{\mathbf{f}\mathbf{f}^T}/\lambda \geqq \overline{\mathbf{e}\mathbf{f}^T} + \overline{\mathbf{f}\mathbf{e}^T}, \\ &M(\lambda) = \lambda P + Q/\lambda \geqq \overline{\mathbf{e}\mathbf{f}^T} + \overline{\mathbf{f}\mathbf{e}^T}. \end{aligned} \tag{1}$$

LEMMA 2. *A symmetric matrix C is conservative if and only if* $L(C) \geqq 1$.

Proof. Sufficiency (assume $L(C) \geqq 1$). Define the scalars

$$\xi = \mathbf{x}^T\mathbf{e}, \qquad \eta = \mathbf{x}^T\mathbf{f}.$$

Then by assumption it follows that

$$\mathbf{x}^T C \mathbf{x} \geqq 2\sqrt{\overline{\xi^2}\,\overline{\eta^2}}. \tag{2}$$

Substituting the Schwarz inequality

$$\sqrt{\overline{\xi^2}\,\overline{\eta^2}} \geqq \overline{\xi\eta} \tag{3}$$

into (2) yields

$$\mathbf{x}^T C \mathbf{x} \geqq 2\overline{\xi\eta} = 2\mathbf{x}^T\overline{\mathbf{e}\mathbf{f}^T}\mathbf{x} = \mathbf{x}^T(\overline{\mathbf{e}\mathbf{f}^T} + \overline{\mathbf{f}\mathbf{e}^T})\mathbf{x}.$$

This yields by definition $C \geqq \overline{\mathbf{e}\mathbf{f}^T} + \overline{\mathbf{f}\mathbf{e}^T}$.

Necessity (assume C is conservative). Construct $\mathbf{e}$ and $\mathbf{f}$ as follows. Let $\mathbf{u}$ be a random n-vector with covariance given by $\overline{\mathbf{u}\mathbf{u}^T} = I$, and let $\mathbf{e} = \sqrt{P}\mathbf{u}$,

$\mathbf{f} = \sqrt{Q}W\mathbf{u}$, where $\sqrt{P}$ is a square matrix with the property that $\sqrt{P}(\sqrt{P})^T = P$ (Halmos [1]) and W is an orthogonal matrix (i.e., $WW^T = I$) which will be chosen below. Then for an arbitrary n-vector $\mathbf{x}$,

$$\mathbf{x}^T(\overline{\mathbf{ef}^T} + \overline{\mathbf{fe}^T})\mathbf{x} = 2\mathbf{x}^T\sqrt{P}W^T(\sqrt{Q})^T\mathbf{x}. \tag{4}$$

The proof will be completed by choosing W such that

$$\mathbf{x}^T\sqrt{P}W^T(\sqrt{Q})^T\mathbf{x} = \sqrt{(\mathbf{x}^TP\mathbf{x})(\mathbf{x}^TQ\mathbf{x})}, \tag{5}$$

since by assumption

$$\mathbf{x}^TC\mathbf{x} \geqq \mathbf{x}^T(\overline{\mathbf{ef}^T} + \overline{\mathbf{fe}^T})\mathbf{x} \tag{6}$$

and the result that $L(C) \geqq 1$ follows from (4), (5) and (6).

Construction of W. In terms of the vectors $\mathbf{y} = (\sqrt{P})^T\mathbf{x}$, $\mathbf{z} = (\sqrt{Q})^Tx$, (5) requires that $\mathbf{z}^TW\mathbf{y} = |\mathbf{z}|\,|\mathbf{y}|$ which will in turn be satisfied if

$$\frac{W\mathbf{y}}{|\mathbf{y}|} = \frac{\mathbf{z}}{|\mathbf{z}|}. \tag{7}$$

Thus (7) is a sufficient condition on W to ensure the proof of Lemma 2.

Define the unit vectors $\mathbf{u}_1 = \mathbf{y}/|\mathbf{y}|$, $\mathbf{u}_2 = \mathbf{z}/|\mathbf{z}|$. Now let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ be an orthonormal set of n-vectors with $\mathbf{x}_1 = \mathbf{u}_1$, and let $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n$ be an orthonormal set of n-vectors with $\mathbf{y}_1 = \mathbf{u}_2$. Then the matrix X, whose columns are the column vectors $\mathbf{x}_1$ to $\mathbf{x}_n$ (i.e., $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$), and the matrix Y, whose columns are the column vectors $\mathbf{y}_1$ to $\mathbf{y}_n$ (i.e., $Y = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]$), are orthogonal. If $W = YX^T$, W is orthogonal and transforms $\mathbf{u}_1$ into $\mathbf{u}_2$, satisfying (7).

LEMMA 3. *For any conservative matrix C there exists an* $\mathbf{x}_0$ *in T such that* $L(C, \mathbf{x}_0) = L(C)$.

Proof. Define an infinite sequence of vectors in T, $\{\mathbf{x}_n\}$, such that

$$L(C, \mathbf{x}_n) \to L(C) \quad \text{as} \quad n \to \infty. \tag{8}$$

Let V_0 be the orthogonal complement of the intersection of the null spaces of P, Q and C. V_0 is a linear subspace of Euclidean n-space and is therefore a closed set. Define $\mathbf{y}_n$ as the unit vector parallel to the orthogonal projection of $\mathbf{x}_n$ into V_0. Then

$$\mathbf{x}_n = \gamma_n\mathbf{y}_n + \mathbf{j}_n,$$

where $\mathbf{j}_n$ is a null vector of the matrices P, Q and C, and

$$L(C, \mathbf{x}_n) = \frac{\gamma_n^2(\mathbf{y}_n^TC\mathbf{y}_n)}{2\sqrt{\gamma_n^4(\mathbf{y}_n^TP\mathbf{y}_n)(\mathbf{y}_n^TQ\mathbf{y}_n)}} = L(C, \mathbf{y}_n).$$

By (8) it follows that

$$L(C, \mathbf{y}_n) \to L(C) \quad \text{as} \quad n \to \infty. \tag{9}$$

Let S be the intersection of the unit sphere in Euclidean n-space with V_0. S is closed since the intersection of two closed sets is itself closed, and S is bounded because it is a subset of the unit sphere. Therefore, S is a compact set, and by the

Bolzano–Weierstrass theorem the sequence $\{\mathbf{y}_n\}$ lying in S has a limit point $\mathbf{x}_0$. $\mathbf{x}_0 \in V_0$ and since $L(C, \mathbf{x})$ is continuous on T, the theorem will follow by (9) if it can be shown that $\mathbf{x}_0 \in T$.

Define $\mathbf{x}_0^T P \mathbf{x}_0 = a$, $\mathbf{x}_0^T Q \mathbf{x}_0 = b$, $\mathbf{x}_0^T C \mathbf{x}_0 = c$. Now, $\mathbf{x}_0 \in T$ if and only if

$$a > 0, \qquad b > 0. \tag{10}$$

Inequality (10) will be proved by contradiction. Assume $a = 0$. By the definition of $L(C, \mathbf{y}_n)$,

$$2L(C, \mathbf{y}_n)\sqrt{(\mathbf{y}_n^T P \mathbf{y}_n)(\mathbf{y}_n^T Q \mathbf{y}_n)} = \mathbf{y}_n^T C \mathbf{y}_n,$$

and taking the limit over the subsequence of $\{\mathbf{y}_n\}$ which approaches $\mathbf{x}_0$ yields

$$2L(C)\sqrt{ab} = c.$$

Since $L(C)$ is finite, $c = 0$, and because $P, C \geqq 0$, it follows that $\mathbf{x}_0$ is a null vector of both P and C. Now $\mathbf{x}_0$ lies on the unit sphere and is thus nonzero and $\mathbf{x}_0$ also lies in V_0 and cannot be a null vector of all three of the matrices P, Q and C. Thus $b \neq 0$. Since $P \neq 0$, a vector $\mathbf{v}$ may be chosen such that $\mathbf{v}^T P \mathbf{v} > 0$. Then for ε sufficiently small, $(\mathbf{x}_0 + \varepsilon \mathbf{v}) \in T$ and

$$L(C, \mathbf{x}_0 + \varepsilon \mathbf{v}) = \frac{\varepsilon \mathbf{v}^T C \mathbf{v}}{2\sqrt{\mathbf{v}^T P \mathbf{v}(\mathbf{x}_0 + \varepsilon \mathbf{v})^T Q(\mathbf{x}_0 + \varepsilon \mathbf{v})}} \to 0 \quad \text{as} \quad \varepsilon \to 0.$$

This contradicts the assumption that C is conservative and the same result ensues if b is assumed to be zero at the outset. Thus (10) is verified and the lemma follows.

LEMMA 4. *If C is conservative and $L(C, \mathbf{x}_0) = L(C)$, then $C \geqq M(\lambda)$ with*

$$\lambda = \sqrt{\mathbf{x}_0^T Q \mathbf{x}_0 / \mathbf{x}_0^T P \mathbf{x}_0}.$$

Proof. Define $D = C/L(C)$. Since $L(C) \geqq 1$, it is sufficient for the proof to show that $\mathbf{x}^T D \mathbf{x} \geqq \mathbf{x}^T M(\lambda) \mathbf{x}$ for any n-vector $\mathbf{x}$. Define $a = \mathbf{x}_0^T P \mathbf{x}_0$, $b = \mathbf{x}_0^T Q \mathbf{x}_0$, $g = \mathbf{x}_0^T D \mathbf{x}_0$. Then by the definition of D,

$$g = 2\sqrt{ab}, \tag{11}$$

and by the definition of T, the quantities a, b and g are greater than zero. Let $\mathbf{x}$ be any n-vector in T and define

$$\mathbf{y} = \mathbf{x} - r\mathbf{x}_0, \tag{12}$$

where $r = \mathbf{x}_0^T D \mathbf{x}/g$. Then $\mathbf{x}_0^T D \mathbf{y} = 0$. Define

$$\tilde{D} = \begin{bmatrix} g & 0 \\ 0 & g\delta \end{bmatrix} = [\mathbf{x}_0, \mathbf{y}]^T D [\mathbf{x}_0, \mathbf{y}],$$

$$\tilde{P} = a\begin{bmatrix} 1 & e \\ e & \alpha \end{bmatrix} = [\mathbf{x}_0, \mathbf{y}]^T P [\mathbf{x}_0, \mathbf{y}],$$

$$\tilde{Q} = b\begin{bmatrix} 1 & f \\ f & \beta \end{bmatrix} = [\mathbf{x}_0, \mathbf{y}]^T Q [\mathbf{x}_0, \mathbf{y}].$$

Let a vector $\mathbf{z}$ be defined by $\mathbf{z} = \gamma\mathbf{x}_0 + \mu\mathbf{y}$. By the definitions of $L(C, \mathbf{x})$ and $L(C)$ it follows that

$$\mathbf{z}^T C\mathbf{z} \geqq 2L(C)\sqrt{(\mathbf{z}^T P\mathbf{z})(\mathbf{z}^T Q\mathbf{z})} \quad \text{or}$$

$$\mathbf{z}^T D\mathbf{z} \geqq 2\sqrt{(\mathbf{z}^T P\mathbf{z})(\mathbf{z}^T Q\mathbf{z})}\,. \tag{13}$$

The following lemma is merely stated here and will be proven subsequently.

LEMMA. *The validity of* (13) *for all* μ *and* γ *implies that the following relationships hold*:

$$f = -e, \tag{14}$$

$$2\delta \geqq \alpha + \beta. \tag{15}$$

Now, rewriting (12) yields $\mathbf{x} = \mathbf{y} + r\mathbf{x}_0$ and thus

$$\mathbf{x}^T D\mathbf{x} = g(r^2 + \delta). \tag{16}$$

By definition, $\lambda = \sqrt{b/a}$ and

$$\mathbf{x}^T(\lambda P + Q/\lambda)\mathbf{x} = \sqrt{b/a}\,(r^2 + 2er + \alpha)a + \sqrt{a/b}\,(r^2 + 2rf + \beta)b.$$

By (11) and (14) this becomes

$$\mathbf{x}^T\left(\lambda P + \frac{Q}{\lambda}\right)\mathbf{x} = g\left(r^2 + \frac{\alpha + \beta}{2}\right),$$

and by (15) and (16) there results

$$\mathbf{x}^T D\mathbf{x} \geqq \mathbf{x}^T(\lambda P + Q/\lambda)\mathbf{x}.$$

Proof of lemma. Substitution of the definitions of $\tilde{P}$, $\tilde{Q}$, $\tilde{D}$, $\mathbf{y}$ and $\mathbf{z}$ into (13) yields

$$\gamma^2\mathbf{x}_0^T D\mathbf{x}_0 + \gamma\mu(\mathbf{x}_0^T D\mathbf{y} + \mathbf{y}^T D\mathbf{x}_0) + \mu^2\mathbf{y}^T D\mathbf{y} \geqq 2\sqrt{q},$$

where

$$q = [\gamma^2\mathbf{x}_0^T P\mathbf{x}_0 + \gamma\mu(\mathbf{x}_0^T P\mathbf{y} + \mathbf{y}^T P\mathbf{x}_0) + \mu^2\mathbf{y}^T P\mathbf{y}]$$
$$\cdot\,[\gamma^2\mathbf{x}_0^T Q\mathbf{x}_0 + \gamma\mu(\mathbf{x}_0^T Q\mathbf{y} + \mathbf{y}^T Q\mathbf{x}_0) + \mu^2\mathbf{y}^T Q\mathbf{y}]$$

or

$$g(\gamma^2 + \mu^2\delta) \geqq 2\sqrt{ab}\sqrt{(\gamma^2 + 2\gamma\mu e + \mu^2\alpha)(\gamma^2 + 2\gamma\mu f + \mu^2\beta)}.$$

Substitution of (11) yields

$$\gamma^2 + \mu^2\delta \geqq \sqrt{(\gamma^2 + 2\gamma\mu e + \mu^2\alpha)(\gamma^2 + 2\gamma\mu f + \mu^2\beta)}.$$

Squaring both sides and collecting terms yields

$$\mu^4(\delta^2 - \alpha\beta) - 2\gamma\mu^3(\alpha f + e\beta) + \gamma^2\mu^2(2\delta - 4ef - \alpha - \beta) - 2\gamma^3\mu(e + f) \geqq 0. \tag{17}$$

As a function of μ, the left-hand side of (17) must have a local minimum at the point $\mu = 0$.

Setting the first derivative to zero at the point $\mu = 0$ yields

$$f = -e. \tag{14}$$

Substitution of (14) into (17) and division by μ^2 yields

$$\mu^2(\delta^2 - \alpha\beta) + 2\mu\gamma e(\alpha - \beta) + \gamma^2(2\delta - 4e^2 - \alpha - \beta) \geqq 0. \tag{18}$$

Define

$$s = 2\delta - \alpha - \beta. \tag{19}$$

Since $\delta = \mathbf{y}^T D\mathbf{y}/\mathbf{x}_0^T D\mathbf{x}_0$ implies that $\delta \geqq 0$,

$$s \geqq -(\alpha + \beta). \tag{20}$$

Define

$$m_1(s) = s + 4e^2, \tag{21}$$

$$m_2(s) = 2e(\alpha + \beta), \tag{22}$$

$$m_3(s) = [(\alpha - \beta)^2 + 2(\alpha + \beta)s + s^2]/4. \tag{23}$$

Inequality (18) may be rewritten

$$\gamma^2 m_1(s) + m_2(s)\gamma\mu + m_3(s)\mu^2 \geqq 0. \tag{24}$$

In order for (24) to hold for all μ and γ the following must be true:

$$m_1(s) \geqq 0, \tag{25}$$

$$m_3(s) \geqq 0, \tag{26}$$

$$4m_1(s)m_3(s) \geqq m_2^2(s). \tag{27}$$

Substitution of (21), (22) and (23) into (27) yields

$$4sm_4(s) \geqq 0, \tag{28}$$

where

$$m_4(s) = m_3(s) + e^2[2(\alpha + \beta) + s]. \tag{29}$$

By (20) and (26) it follows that $m_4(s) \geqq e^2(\alpha + \beta)$. Since $\alpha, \beta \geqq 0$, it follows that $m_4(s) \geqq 0$, and by (28), either

$$s \geqq 0 \tag{30}$$

or $m_4(s) = 0$, which implies that

$$e^2(\alpha + \beta) = 0. \tag{31}$$

If (30) holds, then by (19) there follows

$$2\delta \geqq \alpha + \beta. \tag{15}$$

If $e = 0$, $m_1(s) = s$ by (21), $s \geqq 0$ by (25), and (15) holds by (19). If $\alpha + \beta = 0$, then (15) holds since $\delta \geqq 0$. Thus (15) always holds and the lemma is proved.

LEMMA 5. *$M(\lambda)$ is minimal if λ is in R.*

Proof. Let C be a conservative matrix such that $C \leqq M(\lambda)$. Since λ is in R there is an $\mathbf{x}_0$ in T such that

$$\lambda = \sqrt{\mathbf{x}_0^T Q \mathbf{x}_0 / \mathbf{x}_0^T P \mathbf{x}_0}.$$

By the definition of $M(\lambda)$, $\mathbf{x}_0^T M(\lambda)\mathbf{x}_0 = 2\sqrt{(\mathbf{x}_0^T P \mathbf{x}_0)(\mathbf{x}_0^T Q \mathbf{x}_0)}$. Also, by Lemma 2, $L(C, \mathbf{x}_0) \geqq 1$ and hence

$$\mathbf{x}_0^T C \mathbf{x}_0 \geqq 2\sqrt{(\mathbf{x}_0^T P \mathbf{x}_0)(\mathbf{x}_0^T Q \mathbf{x}_0)}.$$

Thus $\mathbf{x}_0^T C \mathbf{x}_0 \geqq \mathbf{x}_0^T M(\lambda)\mathbf{x}_0$. Since, by assumption, $C \leqq M(\lambda)$ it must be true that $\mathbf{x}_0^T C \mathbf{x}_0 = \mathbf{x}_0^T M(\lambda)\mathbf{x}_0$, which yields $\mathbf{x}_0^T C \mathbf{x}_0 = 2\sqrt{(\mathbf{x}_0^T P \mathbf{x}_0)(\mathbf{x}_0^T Q \mathbf{x}_0)}$. Thus $L(C, \mathbf{x}_0) = 1$ and since $L(C) \geqq 1$ and $L(C)$ is the infimum of $L(C, \mathbf{x})$ over T, it follows that $L(C, \mathbf{x}_0) = L(C)$. By Lemma 4, $C \geqq M(\lambda)$ with $\lambda = \sqrt{\mathbf{x}_0^T Q \mathbf{x}_0 / \mathbf{x}_0^T P \mathbf{x}_0}$. Thus $M(\lambda) \leqq C \leqq M(\lambda)$ or $C = M(\lambda)$, which implies that $M(\lambda)$ is minimal.

THEOREM 3. *If P is nonsingular, it follows that:*

(a) *The eigenvalues of $P^{-1}Q$ are nonnegative.*

(b) *Let a and b be the minimum and maximum eigenvalues of $P^{-1}Q$ respectively. If a is nonzero, R is the closed interval $[\sqrt{a}, \sqrt{b}]$ and, if a is zero, R is the half-open interval $(0, \sqrt{b}]$.*

Proof. Let $E = P^{-1}Q$, and define the inner product for n-vectors $\mathbf{x}$ and $\mathbf{y}$, $\langle \mathbf{x}, \mathbf{y} \rangle$, to be $\mathbf{x}^T P \mathbf{y}$. Thus the inner product $\langle \mathbf{x}, E\mathbf{y} \rangle$ is equal to $\mathbf{x}^T Q \mathbf{y}$. The operator E^*, adjoint to E relative to the inner product given above, is defined by the identity $\langle E^*\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, E\mathbf{y} \rangle$. Thus the equation $\mathbf{x}^T E^{*T} P \mathbf{y} = \mathbf{x}^T Q \mathbf{y}$ holds for all $\mathbf{x}$ and $\mathbf{y}$. Since P and Q are symmetric, it follows that $E^* = P^{-1}Q = E$ and E represents a positive semidefinite self-adjoint operator relative to the given inner product.

Now (a) follows since the eigenvalues of a positive semidefinite self-adjoint operator are nonnegative. Let

$$\tilde{R} = \{\lambda | \lambda = \sqrt{\mathbf{x}^T Q \mathbf{x} / \mathbf{x}^T P \mathbf{x}}, |\mathbf{x}| \neq 0\}.$$

The definition of $\tilde{R}$ is the same as that of the set R above, except that $\mathbf{x} \in T$ for the set R. This condition prevents $\mathbf{x}$ from being a null vector of Q, and hence R equals $\tilde{R}$ with the point $\lambda = 0$ deleted if present. In terms of the inner product, $\tilde{R}$ equals the set of λ such that

$$\lambda = \sqrt{\langle \mathbf{u}, E\mathbf{u} \rangle}, \qquad \langle \mathbf{u}, \mathbf{u} \rangle = 1.$$

Since the unit sphere determined by the inner product is closed and connected, and $\tilde{R}$ is a continuous image of the unit sphere, $\tilde{R}$ is a closed interval. Applied to the maximum and minimum eigenvalues of E, b and a, the minimax principle [3] states that

$$b = \max_{\langle \mathbf{u}, \mathbf{u} \rangle = 1} \langle \mathbf{u}, E\mathbf{u} \rangle,$$

$$a = \min_{\langle \mathbf{u}, \mathbf{u} \rangle = 1} \langle \mathbf{u}, E\mathbf{u} \rangle.$$

Thus $\tilde{R} = [\sqrt{a}, \sqrt{b}]$ and the theorem is proved.

4. A technique for choosing λ. With the minimal approximation to the covariance between the estimation error and the forcing function, the differential equation for the estimation error covariance matrix takes the form

$$\dot{P} = FP + PF^T + \lambda P + Q/\lambda. \tag{32}$$

There remains the problem of choosing λ in some logical way. In this section an optimal control problem is formulated with λ as the control variable, the elements of P as the state variables, and the cost function given by some linear combination of the elements of P at a final time of interest T. This optimization problem may be solved without iteration.

The optimization problem is to choose the bounded measurable function $\lambda(t)$ to minimize the cost function

$$J = \text{trace}\,[LP(T)], \tag{33}$$

where L is a given positive semidefinite symmetric matrix and $P(t)$ is the absolutely continuous solution of (32) with the given initial condition $P(t_0) = P_0 \geqq 0$. In (32), the scalar function λ is required to be positive and the matrices $F(t)$ and $Q(t)$ are given continuously differentiable functions of time with $Q(t) \geqq 0$. (Note that $J = \overline{\mathbf{e}^T L \mathbf{e}}$ if $\mathbf{e}$ is the estimation error and $P = \overline{\mathbf{e}\mathbf{e}^T}$.) The solution for λ is expressed in terms of the state transition matrix from t to T, $\Phi(T, t)$, which is defined by

$$\frac{\partial \Phi(T, t)}{\partial t} = -\Phi(T, t)F(t), \qquad \Phi(T, T) = I. \tag{34}$$

The form of the optimal λ function and a proof of its minimality will be presented as two theorems. The theorems are listed together without proof, followed by the individual proofs.

THEOREM 4. *In order that the function $\lambda(t)$ minimize the cost functional J given by* (33), *it is necessary and sufficient that*

$$\lambda(t) = \left\{\frac{\text{trace}\,[K(t)Q(t)]}{\text{trace}\,[K(t)P(t)]}\right\}^{1/2}, \tag{35}$$

where $K(t) = \Phi^T(T, t)L\Phi(T, t)$ and $\Phi(T, t)$ is the state transition matrix defined above.

(Note that an optimum solution does not exist if trace $[K(t)Q(t)]$ or trace $[K(t)P(t)]$ vanishes on a set of positive measure. However if both traces vanish simultaneously on a set of positive measure, an optimum solution does exist and any control may be used.)

THEOREM 5. *If the minimizing function $\lambda(t)$ exists, then for almost every t for which the denominator of* (35) *is nonzero, $\lambda \in R$, where R is the time varying region defined above for which $M(\lambda)$ is minimal.*

Proof of Theorem 4. *Necessity.* Employing the Pontryagin maximum principle [4] with the costate variables C_{ij} for P_{ji}, the Hamiltonian function becomes

$$H = \text{trace}\,\{C[FP + PF^T + \lambda P + Q/\lambda]\}.$$

The costate differential equation, $\dot{C}_{ij} = -\partial H/\partial P_{ji}$, in matrix form yields

$$\dot{C} = -F^T C - CF - \lambda C,$$

and the terminal condition, $C_{ij}(T) = \partial J/\partial P_{ji}(T) = L_{ij}$, in matrix form becomes

$$C(T) = L.$$

It may be verified by direct substitution that the solution of the costate equation is

$$C(t) = h(t)K(t),$$

where $h(t) = \exp\left[\int_t^T \lambda(s)\,ds\right]$. Note that $h(t) > 0$, $K(t) \geqq 0$.

It may also be verified by direct substitution that the state is given by

$$P(t) = g(t, t_0)\Phi(t, t_0)P_0\Phi^T(t, t_0) + \int_{t_0}^{t} [g(t, s)\Phi(t, s)Q(s)\Phi^T(t, s)/\lambda(s)]\,ds,$$

where

$$g(t, s) = \exp\left[\int_s^t \lambda(u)\,du\right],$$

$$\frac{\partial\Phi(t, s)}{\partial t} = F(t)\Phi(t, s), \quad \Phi(s, s) = I.$$

Since $g(t, s) > 0$, P_0, $Q(t) \geqq 0$, it follows that $P(t) \geqq 0$.

Since $C(t)$, $P(t)$, $Q(t) \geqq 0$ and the trace of the product of two positive semi-definite matrices is nonnegative, it follows that $\text{trace}[C(t)P(t)] \geqq 0$ and $\text{trace}[C(t)Q(t)] \geqq 0$. With $\lambda \geqq 0$, the Hamiltonian may be written

$$H = \text{trace}\,[C(FP + PF^T)] + 2\sqrt{\text{trace}\,(CP)\,\text{trace}\,(CQ)} + [\sqrt{\lambda\,\text{trace}\,(CP)} - \sqrt{\text{trace}\,(CQ)/\lambda}]^2.$$

Since the radicands are nonnegative, the Hamiltonian is minimized with respect to $\lambda > 0$ when the last term vanishes, i.e.,

$$\lambda = \sqrt{\text{trace}\,(CQ)/\text{trace}\,(CP)} = \sqrt{\text{trace}\,(KQ)/\text{trace}\,(KP)}.$$

Sufficiency. Let $\lambda(t)$ be a control program satisfying (35), let $\lambda_1(t) > 0$ be another control program, and define $P(t)$ and $P_1(t)$ as the corresponding solutions of (32) with the given initial condition. Defining $\Delta P = P_1 - P$ and $\Delta\lambda = \lambda_1 - \lambda$, we have that

$$\Delta\dot{P} = F\Delta P + \Delta PF^T + \lambda_1\Delta P + \Delta\lambda P + \left(\frac{1}{\lambda_1} - \frac{1}{\lambda}\right)Q.$$

Defining $l = \text{trace}\,[K\Delta P]$, we have that $l(t_0) = 0$ and $l(T)$ is the difference in cost resulting from the use of control program $\lambda_1(t)$ rather than $\lambda(t)$.

From the definition of $K(t)$ and (34) it follows that $K(t)$ satisfies

$$\dot{K}(t) = -KF - F^TK,$$

and thus

$$\dot{l} = \lambda_1 l + \operatorname{trace}\left\{K\left[\Delta\lambda P + \left(\frac{1}{\lambda_1} - \frac{1}{\lambda}\right)Q\right]\right\}$$

and

$$l(T) = \int_{t_0}^{T} j^2(t) \exp\left[\int_t^T \lambda_1(u)\, du\right] dt, \tag{36}$$

where

$$j(t) = \sqrt{\lambda_1 \operatorname{trace}(KP)} - \sqrt{\operatorname{trace}(KQ)/\lambda_1}. \tag{37}$$

Since it was shown in the proof of necessity that $K, P \geqq 0$, and the trace of the product of two positive semidefinite matrices is nonnegative; the radicands above are nonnegative and $j(t)$ is real. Inspection of (37) indicates that $j(t)$ is nonzero whenever λ_1 is unequal to λ. Thus by (36), the difference in cost resulting from the use of the two control programs, $l(T)$, is positive if λ_1 differs from λ on a set of positive measure and $\lambda(t)$ is the global optimum solution.

Proof of Theorem 5. By Theorem 2, it is possible to choose $\lambda_1(t)$ such that $\lambda_1(t) \in R(t)$ for all t and

$$M(\lambda_1) \leqq M(\lambda), \tag{38}$$

where the $P(t)$ matrix used in the calculation of $R(t)$, $M(\lambda)$ and $M(\lambda_1)$ is the $P(t)$ matrix determined by the $\lambda(t)$ program.

Let $P_1(t)$ be the solution of

$$\dot{P}_1 = FP_1 + P_1F^T + \lambda_1 P_1 + Q/\lambda_1 \tag{39}$$

with the program $\lambda_1(t)$ and the initial condition $P_1(t_0) = P_0$. Then by defining $\Delta P = P_1 - P$, ΔP satisfies

$$\dot{\Delta P} = F\,\Delta P + \Delta P F^T + M(\lambda_1) - M(\lambda) + \lambda_1 \Delta P$$

with the initial condition $\Delta P(t_0) = 0$. It may be verified by direct substitution that

$$\Delta P(t) = \int_{t_0}^{t} \Phi(t, s)[M(\lambda_1) - M(\lambda)](s)\Phi^T(t, s)h(t, s)\, ds,$$

where

$$h(t, s) = \exp\left[\int_s^t \lambda_1(u)\, du\right].$$

Thus by (38), $\Delta P(t) \leqq 0$.

Let $J_1 = \operatorname{trace}[LP_1(T)]$ be the cost functional evaluated with the control program $\lambda_1(t)$. Since $L \geqq 0$ and the trace of the product of two positive semidefinite matrices is nonnegative, it follows that $J_1 \leqq J$, where J is the cost functional evaluated with the control program $\lambda(t)$. Since $\lambda(t)$ is the global optimum by Theorem 4, it is also true that $J \leqq J_1$, and thus $J_1 = J$. By (36),

$$0 = J_1 - J = l(T) = \int_{t_0}^{T} j^2(t) \exp\left[\int_t^T \lambda_1(u)\, du\right] dt.$$

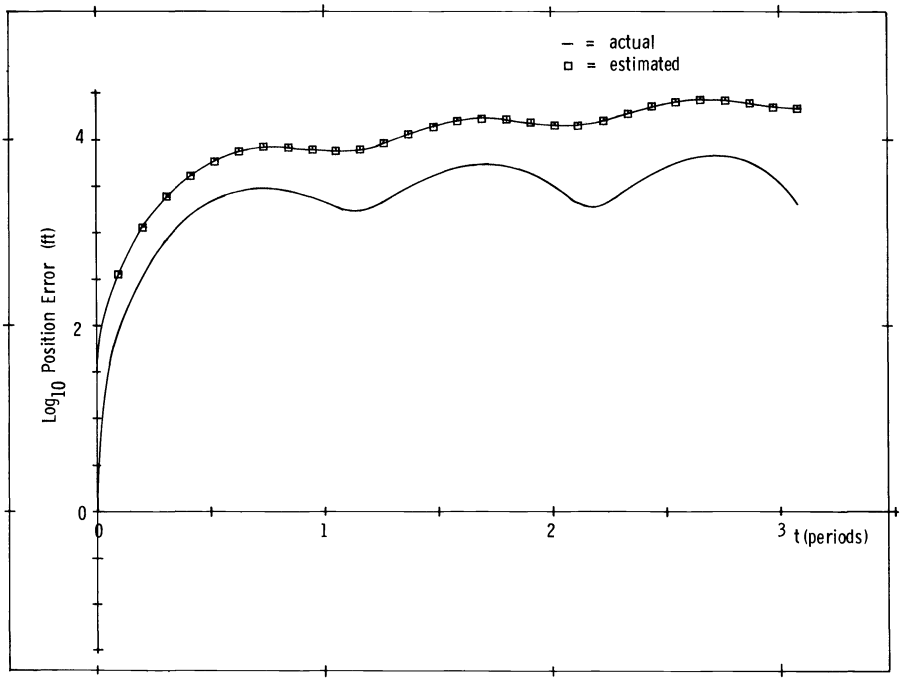


FIG. 1. *Circular orbit*

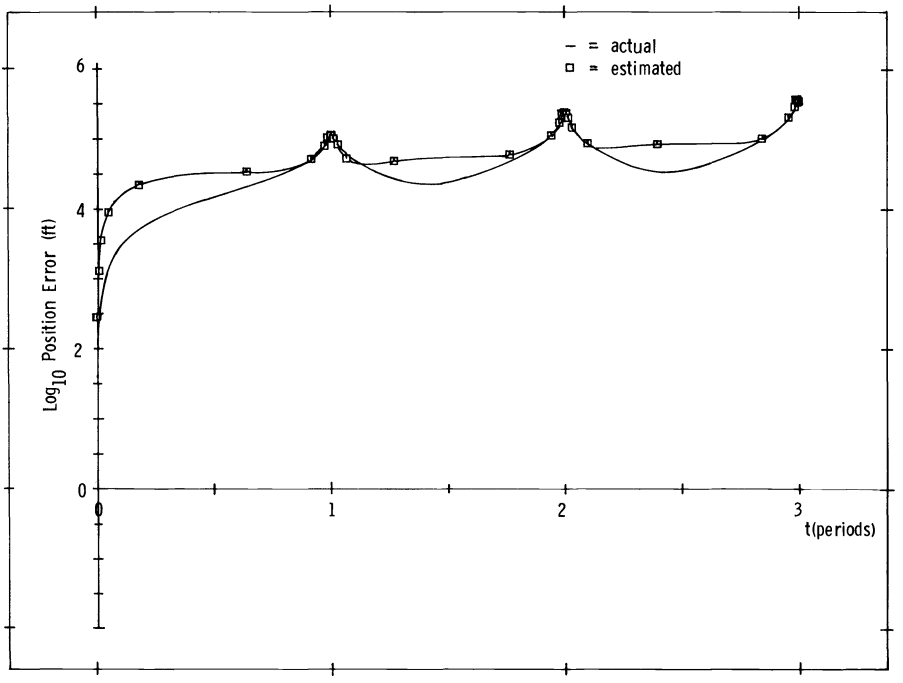


FIG. 2. *Elliptical orbit*

Thus $j(t) = 0$ almost everywhere and by (37),

$$\lambda_1 = \sqrt{\text{trace } (KQ)/\text{trace } (KP)} = \lambda$$

almost everywhere. Since $\lambda_1 \in R$ by definition, $\lambda \in R$ almost everywhere.

5. Simulation results. Figures 1 and 2 illustrate the application of the method described above to the estimation of uncertainties in the orbit of an earth satellite caused by uncertainties in the earth's gravitational field. It was assumed that the inverse square component of the earth's gravitational field was known exactly while the other terms in the spherical harmonic expansion of the earth's gravitational field were unknown perturbations driving the satellite's equations of motion.

Each figure contains two plots, one of which is the RMS position deviation from the nominal conic orbit predicted by the covariance matrix, while the other is the actual magnitude of the position deviation computed using a tenth order spherical harmonic model of the earth's gravitational field with a zero J_2 term. Figure 1 represents a nominal 500,000 ft circular polar orbit starting near Vandenberg Air Force Base while Fig. 2 represents a nominal elliptical orbit, with an eccentricity of 0.8 starting at pericenter 500,000 ft over the North Pole.

Because (32) is essentially a worst case approximation, the state vector chosen must constrain the energy of the orbit, as reflected by the estimated covariance matrix, to lie within the physically realizable bounds dictated by the conservative force field. To this end the energy equation was used to eliminate the component of the deviation from the nominal velocity vector parallel to the nominal velocity vector, and the state vector was defined to consist of the deviation from the nominal energy, the remaining two components of the deviation from the nominal velocity vector, and the three components of the deviation from the nominal position vector. In this case, the forcing function $\mathbf{f}$ was given by $G\mathbf{d}$, where G was a 6×3 matrix determined from the nominal trajectory and $\mathbf{d}$ was a 3-vector consisting of the cross track components of the perturbing gravity acceleration and the perturbation in the massless potential energy at the satellite's present location in space. For the nominal circular orbit, $D = \overline{\mathbf{d}\mathbf{d}^T}$ was approximated by a diagonal matrix whose nonzero elements consisted of the mean square values of the elements of $\mathbf{d}$, computed from the gravity model along the actual trajectory and averaged over three orbit periods. For the nominal elliptical orbit, the D matrix was scaled for radial dependence as though all the components of $\mathbf{d}$ arose from the J_3 term in the spherical harmonic model.

The matrix L was chosen such that trace $[LP(t)]$ was the estimated mean square error in the satellite position. Both trajectories began with zero position and velocity errors, and all elements of the initial P matrix were zero except the diagonal element corresponding to the energy deviation, which was equal to the diagonal element of D corresponding to the initial uncertainty in massless potential energy.

REFERENCES

[1] PAUL R. HALMOS, *A Hilbert Space Problem Book*, Van Nostrand, Princeton, 1967.

[2] R. E. KALMAN AND R. BUCY, *New results in linear filtering and prediction*, Trans. ASME, 83D (1961), pp. 95–108.

[3] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York, 1966.

[4] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

NECESSARY CONDITIONS FOR OPTIMIZATION PROBLEMS WITH OPERATORIAL CONSTRAINTS*

C. VÎRSAN†

Abstract. In this paper, we extend some results of H. Halkin and L. W. Neustadt [1] and [2] to optimization problems in which there is an equality constraint defined in terms of an operator whose range is infinite-dimensional.

We obtain abstract multiplier rules (Theorems 2 and 3) which make it possible to obtain necessary conditions in the form of a maximum principle (Theorem 4) for optimal control problems with equality-type phase constraints.

Introduction. In the last few years it has become clear that the necessary conditions for optimality in control theory, known as the maximum principle, may be obtained by using abstract multiplier rules. Such abstract multiplier rules which make it possible to obtain the maximum principle were given in [1] and [2] by H. Halkin and L. W. Neustadt. The results in [1] and [2] were well-adapted to problems that contain constraints which are given by a finite number of functionals. In order to study optimal control problems with *equality-type* phase constraints, we need abstract multiplier rules for general optimization problems with equality constraints defined by an operator with infinite-dimensional range. It is the purpose of this paper to obtain such abstract results and to apply them to optimal control problems with equality-type phase constraints.

We begin § 1 with the usual case of a finite number of equality constraints, for which we prove an alternate form of the Halkin–Neustadt result; in fact, we replace the explicit condition obtained by Halkin and Neustadt with the aid of a fixed-point theorem by another one (condition II in Theorem 1) in which we require the existence of a map A with certain properties, and then show, with the use of Lemma 1, that under the assumptions made by Halkin, such a map A can be constructed. Then, in § 2, we strengthen condition II in a way which allows us to obtain the corresponding result for operatorial constraints. In this more general case, it does not seem possible to obtain, with the use of fixed-point theorems, explicit conditions for the existence of such a map A which is convenient in all situations. Nevertheless, in specific optimal control problems, such as the one with equality-type phase constraints, the construction of the map A is possible, as we show in § 3, which makes up the main part of the paper. Our final result is Theorem 4, where we obtain the necessary conditions from the abstract multiplier rule in the form of a maximum principle.

The most difficult parts of our proofs are the constructions of all the elements which appear in the abstract Theorem 3, and the verification that the hypotheses of this theorem hold. In this verification, we make use of some general properties of ordinary differential equations and of implicit function theorems, which are of course essentially equivalent to fixed-point theorems.

Our final result has much in common with that of Gamkrelidze [3] and the ones of Guinn [4] and Berkovitz [5]. We point out, however, that Gamkrelidze

* Received by the editors March 5, 1968, and in final revised form January 19, 1970.

† Institute of Mathematics of the Romanian Academy of Sciences, Bucharest, Romania.

did not consider equality-type phase constraints and that our result cannot be obtained by using the same kind of variations that he used, and that Guinn made stronger regularity assumptions than we do.

1. The case of a finite number of equality constraints. Let $\mathscr{X}$ be a real linear space, let $L \subset \mathscr{X}$ be arbitrary, and let φ_i, for $i = 0, 1, \cdots, m + k$, be real-valued functions defined on L. Further, let $\varphi = (\varphi_{m+1}, \cdots, \varphi_{m+k})$, let θ be the zero element in R^k, let $\tilde{L} = \{x \in L : \varphi(x) = \theta\}$ and let $L^- = \{x \in L : \varphi_i(x) \leqq 0,\ i = 1, \cdots, m\}$. We shall say that an element $\tilde{x} \in \tilde{L} \cap L^-$ is optimal if $\varphi_0(\tilde{x}) \leqq \varphi_0(x)$ for all elements $x \in \tilde{L} \cap L^-$.

We assume that an optimal element $\tilde{x}$ exists and, without loss of generality, that $\tilde{x} = 0$, that $\varphi_i(0) = 0$ for $i = 0, \cdots, p$, and that $\varphi_i(0) < 0$ for $i = p + 1, \cdots, m$.

THEOREM 1. *Let 0 be an optimal element. Suppose that there exists a convex set $M \subset \mathscr{X}$ such that $0 \in M$ and that there exist functions $h_i : M \to R$ for $i = 0, 1, \cdots, m + k$, such that:*

I. *the h_i, for $i = 0, 1, \cdots, m$, are convex, $h_i(0) = \varphi_i(0)$ for $i = 0, 1, \cdots, m$, and $h = (h_{m+1}, \cdots, h_{m+k}) : \mathscr{X} \to R^k$ is linear;*

II. *either*[1] $\theta \notin \operatorname{int} h(M)$, *or, whenever $x_1, \cdots, x_{k+1}$ are linearly independent elements of M with the property that*[1] $\theta \in \operatorname{int} \operatorname{co} \{h(x_1), \cdots, h(x_{k+1})\}$, *there exist functions $\delta : (0, 1) \to (0, 1)$ and $A : \delta(0, 1) \to \tilde{L}$ such that*

$$\lim_{\varepsilon \to 0^+} \frac{\varphi_i(A(\delta(\varepsilon))) - h_i(\delta(\varepsilon)\bar{x})}{\delta(\varepsilon)} \leqq 0 \quad \text{for } i = 0, 1, \cdots, m,$$

$$\lim_{\varepsilon \to 0^+} \delta(\varepsilon) = 0,$$

where $\bar{x}$ is the unique element in $\operatorname{co} \{x_1, \cdots, x_{k+1}\}$ such that $h(\bar{x}) = \theta$. Then there exist constants α_i, $i = 0, 1, \cdots, m + k$, such that

$$\text{(a)} \quad \sum_0^{m+k} \alpha_i h_i(x) \leqq \sum_0^{m+k} \alpha_i h_i(0) = 0 \quad \text{for all } x \in M,$$

$$\text{(b)} \quad \sum_0^{m+k} |\alpha_i| > 0, \quad \alpha_i \leqq 0, \quad \text{for } i = 0, 1, \cdots, m,$$

$$\alpha_i \varphi_i(0) = 0 \quad \text{for } i = 1, \cdots, m.$$

Proof. Let $g = (h_0, h_1, \cdots, h_m, h)$, and let $\mathscr{A}$ denote the set $g(M) + P \subset R^{m+k+1}$, where

$$P = \{(\varepsilon_0, \varepsilon_1, \cdots, \varepsilon_{m+k}) : \varepsilon_i \geqq 0 \text{ for } i = 0, 1, \cdots, m,$$
$$\varepsilon_i = 0 \text{ for } i = m + 1, \cdots, m + k\}.$$

The set $\mathscr{A}$ is convex. Indeed, let $a_1, a_2 \in \mathscr{A}$ so that there exist elements $x_1, x_2 \in M$ and $\varepsilon', \varepsilon'' \in P$ such that $g(x_1) + \varepsilon' = a_1$ and $g(x_2) + \varepsilon'' = a_2$. Let $0 \leqq \alpha \leqq 1$. Since the functions h_i for $i = 0, 1, \cdots, m$ are convex, it follows that

$$\bar{\varepsilon} = \alpha g(x_1) + (1 - \alpha) g(x_2) - g(\alpha x_1 + (1 - \alpha) x_2) \in P.$$

Since M is convex, we have that $x = \alpha x_1 + (1 - \alpha) x_2 \in M$ and

$$\alpha a_1 + (1 - \alpha) a_2 = g(x) + \alpha \varepsilon' + (1 - \alpha) \varepsilon'' = \bar{\varepsilon}.$$

[1] If A is a set in a linear vector space, then co A will denote the convex hull of A; if the space is also topological, then int A will denote the interior of A.

It is clear that $\alpha\varepsilon' + (1 - \alpha)\varepsilon'' + \bar{\varepsilon} \in P$, and therefore $\mathscr{A}$ is convex.

We shall first consider the case where $\operatorname{int} \mathscr{A} \neq \varnothing$.

Consider the set

$$B = \{(\xi_0, \cdots, \xi_p, \xi_{p+1}, \cdots, \xi_m, \theta): \xi_i < 0 \text{ for } i = 0, 1, \cdots, p\} \subset R^{m+k+1}.$$

Recall that $\varphi_i(0) = 0$ for $i = 0, 1, \cdots, p$ and that $\varphi_i(0) < 0$ for $i = p + 1, \cdots, m$. It is easy to see that the set B is convex. Since $\mathscr{A}$ is convex, it follows that $\operatorname{int} \mathscr{A}$ is convex. We shall show that $(\operatorname{int} \mathscr{A}) \cap B = \varnothing$. Suppose that $(\operatorname{int} \mathscr{A}) \cap B \neq \varnothing$. Then there exist real numbers $\xi_i^*, i = 0, 1, \cdots, m$, such that $\xi_i^* < 0$ for $i = 0, 1, \cdots, p$, and an open ball $V \subset R^k$ centered at θ such that

$$(\xi_0^*, \cdots, \xi_m^*) \times V \subset \operatorname{int} \mathscr{A}.$$

Let $M^* = \{x \in M: h_i(x) \leqq \xi_i^* \text{ for } i = 0, 1, \cdots, m, h(x) \in V\}$. Since the h_i are convex and h is linear, it follows that the set M^* is convex. Because

$$(\xi_0^*, \cdots, \xi_m^*) \times V \subset \operatorname{int} \mathscr{A},$$

it follows that $h(M^*) = V \subset R^k$. Therefore, there exist elements $x_1, \cdots, x_{k+1} \in M^*$ such that $h(x_1), \cdots, h(x_{k+1})$ are in general position[2] and such that

$$\theta \in \operatorname{int} \operatorname{co} \{h(x_1), \cdots, h(x_{k+1})\}.$$

It follows from the linearity of h that $x_1, \cdots, x_{k+1}$ are in general position and that

$$\operatorname{co} \{h(x_1), \cdots, h(x_{k+1})\} = h(\operatorname{co} \{x_1, \cdots, x_{k+1}\}).$$

Since $h_i(0) = 0 > \xi_i^*$ for $i = 0, 1, \cdots, p$, $0 \notin M^*$. We shall show that $x_1, \cdots, x_{k+1}$ are linearly independent. Suppose that $x_1, \cdots, x_{k+1}$ are linearly dependent. Since they are in general position, there exist real numbers α_i such that $\sum_1^{k+1} \alpha_i = 1$ and $\sum_1^{k+1} \alpha_i x_i = 0$. Therefore,

$$\theta = h\left(\sum_1^{k+1} \alpha_i x_i\right) = \sum_1^{k+1} \alpha_i h(x_1)$$

with $\sum_1^{k+1} \alpha_i = 1$. Since $\theta \in \operatorname{int} \operatorname{co} \{h(x_1), \cdots, h(x_{k+1})\}$, $\alpha_i > 0$ for $i = 1, \cdots, k + 1$. Hence $0 \in \operatorname{co} \{x_1, \cdots, x_{k+1}\}$, contradicting the fact that $0 \notin M^*$.

Let $\bar{x} \in \operatorname{co} \{x_1, \cdots, x_{k+1}\}$ be such that $h(\bar{x}) = \theta$. Since the functions h_i for $i = 0, 1, \cdots, m$ are convex and satisfy condition II, we have that

$$\varphi_i(A(\delta(\varepsilon))) \leqq (1 - \delta(\varepsilon))h_i(0) + \delta(\varepsilon)h_i(\bar{x}) + \varepsilon_i\delta(\varepsilon),$$

where $\lim_{\varepsilon \to 0^+} s_i = 0$. Since $h_i(0) < 0$ for $i = p + 1, \cdots, m$ and $\lim_{\varepsilon \to 0^+} \delta(\varepsilon) = 0$, we conclude that there exists an $\varepsilon' \in (0, 1)$ such that

$$\varphi_i(A(\delta(\varepsilon))) < 0 \quad \text{for } i = p + 1, \cdots, m \quad \text{for all } \varepsilon \in (0, \varepsilon').$$

Since $h_i(\bar{x}) \leqq \xi_i^* < 0$ and $h_i(0) = 0$ for $i = 0, 1, \cdots, p$, there exists an $\varepsilon'' \in (0, \varepsilon')$ such that

$$\varphi_i(A(\delta(\varepsilon))) < 0 \quad \text{for } i = 0, 1, \cdots, m \quad \text{and} \quad \varepsilon \in (0, \varepsilon'').$$

Hence $A(\delta(\varepsilon)) \in \tilde{L} \cap L^-$ and $\varphi_0(A(\delta(\varepsilon))) < 0$ for all $\varepsilon \in (0, \varepsilon'')$, contradicting the

[2] The elements $h(x_1), \cdots, h(x_{k+1})$ are said to be in general position if the elements $[h(x_i) - h(x_{k+1})]$ for $i = 1, \cdots, k$ are linearly independent.

optimality of the zero element. This shows that

$$(\text{int } \mathscr{A}) \cap B = \varnothing.$$

Hence, there exist constants c, α_i, $i = 0, 1, \cdots, m + k$, such that

$$\sum_0^{m+k} |\alpha_i| > 0 \quad \text{and} \quad \sum_0^{m+k} \alpha_i a_i \leqq c \leqq \sum_0^m \alpha_i \xi_i$$

for all $a = (a_0, \cdots, a_{m+k}) \in \mathscr{A}$ and all $\xi = (\xi_0, \cdots, \xi_m, \theta) \in \bar{B}$ (the closure of B). Since $0 \in M$ and $h_i(0) \leqq 0$, the zero element of R^{m+k+1} is in $\mathscr{A}$. Since $\bar{B}$ contains the zero element, $c = 0$. By definition of the sets $\mathscr{A}$ and B, we see that

$$\sum_0^{m+k} \alpha_i h_i(x) + \sum_0^m \alpha_i \varepsilon_i \leqq 0 \quad \text{whenever } x \in M \quad \text{and} \quad \varepsilon_i \geqq 0, \tag{1}$$

$$\sum_0^m \alpha_i \xi_i \geqq 0 \tag{2}$$

whenever $\xi_i \leqq 0$ for $i = 0, 1, \cdots, p$ and $\xi_i \in R$ for $i = p + 1, \cdots, m$. From relation (2), we obtain that $\alpha_i \leqq 0$ for $i = 0, 1, \cdots, p$ and that $\alpha_i = 0$ for $i = p + 1, \cdots, m$. Setting $\varepsilon_i = 0$ for $i = 0, 1, \cdots, m$, we conclude that

$$\sum_0^{m+k} \alpha_i h_i(x) \leqq 0 \quad \text{for all } x \in M, \tag{1'}$$

$$\alpha_i \varphi_i(0) = 0, \quad \alpha_i \leqq 0 \quad \text{for } i = 0, 1, \cdots, m, \tag{2'}$$

and we have obtained the desired conclusions of our theorem.

Now suppose that int $\mathscr{A} = \varnothing$. Since $\mathscr{A}$ is convex, it follows that there exist constants α_i, $i = 0, 1, \cdots, m + k$, and c such that

$$\sum_0^{m+k} |\alpha_i| > 0, \qquad \sum_0^{m+k} \alpha_i a_i = c \quad \text{for all } a = (a_0, \cdots, a_{m+k}) \in \mathscr{A}. \tag{3}$$

Since $h_i(0) = \varphi_i(0) \leqq 0$ for $i = 0, 1, \cdots, m$ (see assumption I) and $h(0) = \theta$, we obtain, after setting $\varepsilon_i = -h_i(0)$, that the zero element of R^{m+k+1} is in $\mathscr{A}$. Hence, $c = 0$.

It follows from the definition of $\mathscr{A}$ that

$$\sum_0^{m+k} \alpha_i h_i(x) + \sum_0^m \alpha_i \varepsilon_i = 0 \quad \text{whenever } x \in M \quad \text{and} \quad \varepsilon_i \geqq 0. \tag{3'}$$

Set $\varepsilon_i \geqq -h_i(0)$ for $i = 0, 1, \cdots, m$; it then follows from (3'), since $0 \in M$, that $\sum_0^m \alpha_i \varepsilon_i' = 0$, where $\varepsilon_i' = \varepsilon_i + h_i(0) \geqq 0$. Hence, $\alpha_i = 0$ for $i = 0, 1, \cdots, m$, and we see, by virtue of (3'), that

$$\sum_{m+1}^{m+k} \alpha_i h_i(x) = 0 \quad \text{for all } x \in M, \qquad \sum |\alpha_i| > 0.$$

If we set $\alpha_i = 0$ for $i = 0, 1, \cdots, m$, we have again obtained our desired conclusion.

Remark 1. The basic element in the proof of Theorem 1 is the construction of comparison elements, starting with elements of M as first approximations.

The role of the maps in A and δ is to make these elements satisfy the equality constraints in L. If X is a normed linear space, L is the kernel of an operator T, and (φ, T) has a Fréchet derivative at $x = 0$ which is equal to (h, T_0), then we may choose M to be the kernel of T, and the maps A and δ in hypothesis II allow us to construct, starting with elements in the kernel of (h, T_0) as first approximations, elements in the kernel of (φ, T).

LEMMA 1. *Let there be given subsets L and M of $\mathcal{X}$, a function $\varphi: L \to R^k$, a linear map $h: \mathcal{X} \to R^k$, and linearly independent elements $x_1, \cdots, x_{k+1} \in M$ such that*

$$\theta \in \operatorname{int} \operatorname{co} \{h(x_1), \cdots, h(x_{k+1})\}.$$

Suppose that there exists a map $\zeta: \operatorname{co} \{0, x_1, \cdots, x_{k+1}\} \to L$ such that

(i) *the map $\tilde{\varphi}_\delta: \operatorname{co}\{x_1, \cdots, x_{k+1}\} \to R^k$ defined by*

$$\tilde{\varphi}_\delta(x) = \varphi(\zeta(\delta x))$$

is continuous[3] *for every $\delta \in (0, 1)$, and*

(ii)
$$\lim_{\delta \to 0^+} \frac{|\tilde{\varphi}_\delta(x) - \delta h(x)|}{\delta} = 0$$

uniformly with respect to $x \in \operatorname{co} \{x_1, \cdots, x_{k+1}\}$. Then there exist maps $\delta: (0, 1) \to (0, 1)$ and $x: (0, 1) \to \operatorname{co} \{x_1, \cdots, x_{k+1}\}$ such that $\zeta(\delta(\varepsilon)x(\varepsilon)) \in L$, $\varphi(\zeta(\delta(\varepsilon)x(\varepsilon))) = \theta$ for all $\varepsilon \in (0, 1)$ and

$$\lim_{\varepsilon \to 0^+} \delta(\varepsilon) = 0, \qquad \lim_{\varepsilon \to 0^+} x(\varepsilon) = \bar{x}$$

in $\operatorname{co} \{x_1, \cdots, x_{k+1}\}$, where $\bar{x}$ is the unique element such that $h(\bar{x}) = \theta$.

Proof. Our aim is to construct a map which satisfies the hypotheses of the Brouwer fixed-point theorem.

Since $x_1, \cdots, x_{k+1}$ are linearly independent, they are in general position. Because $\theta \in \operatorname{int} \operatorname{co} \{h(x_1), \cdots, h(x_{k+1})\}$, it follows from the linearity of h that $h(x_1), \cdots, h(x_{k+1})$ are in general position. For each $\varepsilon \in (0, 1)$, let $S(\varepsilon) = \{x_1(\varepsilon), \cdots, x_{k+1}(\varepsilon)\}$, where $x_i(\varepsilon) = (1 - \varepsilon)\bar{x} + \varepsilon x_i$ for $i = 1, \cdots, k + 1$. Since $h(\bar{x}) = \theta$,

$$h(\operatorname{co} S(\varepsilon)) = \varepsilon h(\operatorname{co} S(1)) = \varepsilon \operatorname{co} h(S(1)).$$

Hence, there exist an $\varepsilon_0 > 0$ and a closed ball $V(\theta, \varepsilon_0)$ centered at θ such that $V(\theta, \varepsilon_0) \subset h(\operatorname{co} S(1))$. Since $h(\operatorname{co} S(\varepsilon)) = \varepsilon \operatorname{co} h(S(1))$,

$$V(\theta, \varepsilon\varepsilon_0) \subset h(\operatorname{co} S(\varepsilon)) \quad \text{for all } \varepsilon \in (0, 1).$$

To each $z \in h(\operatorname{co} S(\varepsilon))$, we correspond the element $x(\varepsilon; z) \in \operatorname{co} S(\varepsilon)$ whose barycentric coordinates coincide with those of z, so that $h(x(\varepsilon; z)) = z$. The map $x(\varepsilon, \cdot): h(\operatorname{co} S(\varepsilon)) \to \operatorname{co} S(\varepsilon)$ is continuous, and

$$(\operatorname{co} S(\varepsilon) \subset \operatorname{co} S(1) = \operatorname{co} \{x_1, \cdots, x_{k+1}\}) \quad \text{for all } \varepsilon \in (0, 1).$$

Let $\zeta: \operatorname{co} \{0, x_1, \cdots, x_{k+1}\} \to L$ be the function whose existence was assumed. For each $\delta \in (0, 1)$, we define the map $\Phi_\delta^\varepsilon(z): h(\operatorname{co} S(\varepsilon)) \to R^k$ through the relation

$$\Phi_\delta^\varepsilon(z) = z - \frac{\tilde{\varphi}_\delta(x(\varepsilon; z))}{\delta}.$$

[3] On the set $\operatorname{co} \{x_1, \cdots, x_{k+1}\}$, we take the ordinary Euclidean finite-dimensional topology.

By property (i) of ζ, Φ_δ^ε is continuous, and it follows from (ii) that

$$\lim_{\delta \to 0^+} |\Phi_\delta^\varepsilon(z)| = \lim_{\delta \to 0^+} \frac{|\tilde{\varphi}(x(\varepsilon; z)) - \delta h(x(\varepsilon; z))|}{\delta} = 0$$

uniformly over $h(\text{co } S(\varepsilon))$, for all $\varepsilon \in (0, 1)$.

Since $V(\theta, \varepsilon\varepsilon_0) \subset h(\text{co } S(\varepsilon))$, there exists, for each $\varepsilon \in (0, 1)$, a number $\delta(\varepsilon)$, with $0 < \delta(\varepsilon) \leqq \varepsilon$, such that $\Phi_{\delta(\varepsilon)}^\varepsilon$ maps $V(\theta, \varepsilon\varepsilon_0)$ into itself. It is a consequence of the Brouwer fixed-point theorem that, for each $\varepsilon \in (0, 1)$, there exists an element $z(\varepsilon) \in V(\theta, \varepsilon\varepsilon_0)$ such that $\Phi_{\delta(\varepsilon)}^\varepsilon(z(\varepsilon)) = z(\varepsilon)$. Hence, $\varphi(\zeta(\delta(\varepsilon)x(\varepsilon); z(\varepsilon))) = \theta$ for all $\varepsilon \in (0, 1)$. We denote $x(\varepsilon; z(\varepsilon))$ by $x(\varepsilon)$. Then, by (i), we have that $\zeta(\delta(\varepsilon)x(\varepsilon)) \in L$ for all $\varepsilon \in (0, 1)$. Since $0 < \delta(\varepsilon) \leqq \varepsilon$, $\lim_{\varepsilon \to 0^+} \delta(\varepsilon) = 0$. Since $x(\varepsilon; z(\varepsilon)) \in \text{co } S(\varepsilon) = (1 - \varepsilon)\bar{x} + \varepsilon \text{ co } \{x_1, \cdots, x_{k+1}\}$, we conclude that $\lim x(\varepsilon) = \bar{x}$, completing the proof of the lemma.

The conclusions of Theorem 1 are the same as those of Theorem II in [1].

Using the preceding lemma, we can show that the hypotheses of Theorem 1 are satisfied under the assumption stated in [1]. The following assumptions were made in [1].

ASSUMPTION H_1.

1. H is a normed linear space, and the functions $f = (\varphi_0, \cdots, \varphi_{m+k})$ and $g = (h_0, \cdots, h_{m+k})$ are continuous.

2. $$\lim_{\|x\| \to 0} \frac{|f(x) - g(x)|}{\|x\|} = 0.$$

3. The functions h_i are convex for $i = 0, 1, \cdots, m$, and linear for $i = m + 1, \cdots, m + k$.

ASSUMPTION H_2.

1. The set M is convex, and $0 \in M$.

2. For every set $S = \{x_1, \cdots, x_{k+1}\} \subset M$ made up of $k + 1$ linearly independent elements, there exists a function $\zeta : \text{co } \{0, x_1, \cdots, x_{k+1}\} \to L$ such that:

(i) for every $\delta \in [0, 1]$, $\zeta(\delta \cdot x)$ is continuous in x over co S,

(ii) $\lim_{\delta \to 0^+} (\zeta(\delta \cdot x)/\delta) = x$ uniformly in x over co S.

The following theorem was stated in [1].

THEOREM. *If* 0 *is an optimal element and assumptions* H_1 *and* H_2 *hold, then there exists a vector* $\alpha = (\alpha_0, \cdots, \alpha_{m+k}) \in R^{m+k+1}$, *with* $|\alpha| \neq 0$, *such that*:

(i) $\alpha g(x) \leqq \alpha g(0) = 0$ *for all* $x \in M$,

(ii) $\alpha_i \leqq 0$ *for* $i = 0, 1, \cdots, m$,

(iii) $\alpha_i h_i(0) = 0$ *for* $i = 1, \cdots, m$.

Conclusion (i) *is identical with conclusion* (a) *of Theorem* 1, *and conclusions* (ii) *and* (iii), *together with* $|\alpha| \neq 0$, *are conclusions* (b) *of Theorem* 1.

PROPOSITION. *If Assumptions* H_1 *and* H_2 *hold, then the hypotheses of Theorem* 1 *also hold.*

Proof. Hypothesis I is obviously satisfied. (See (2) and (3).) We have to verify that hypothesis II holds.

If $\theta \in \text{int } h(M)$, let $x_1, \cdots, x_{k+1} \in M$ be linearly independent and such that

$$\theta \in \text{int co } \{h(x_1), \cdots, h(x_{k+1})\} \subset R^k,$$

where $h = (h_{m+1}, \cdots, h_{m+k}) : \mathcal{X} \to R^k$. The map $\tilde{\varphi}_\delta : \text{co } \{x_1, \cdots, x_{k+1}\} \to R^k$,

defined by $\tilde{\varphi}_\delta(x) = \varphi(\zeta(\delta x))$, where $\varphi = (\varphi_{m+1}, \cdots, \varphi_{m+k})$ and ζ is the function in Assumption H_2, is continuous for each $\delta \in [0, 1]$ inasmuch as φ and ζ are continuous. We see that ζ satisfies condition (i) of Lemma 1. By Assumptions H_1 and H_2 we further conclude that

$$\lim_{\delta \to 0^+} \frac{|\tilde{\varphi}_\delta(x) - \delta h(x)|}{\delta} = \lim_{\delta \to 0^+} \frac{|\varphi(\delta x + o(\delta)) - h(\delta x)|}{\delta}$$

$$\leqq \lim_{\delta \to 0^+} \frac{|\varphi(\delta x + o(\delta)) - h(\delta x + o(\delta))|}{\delta} + \lim_{\delta \to 0^+} \frac{|h(o(\delta))|}{\delta}$$

uniformly in x over co $\{x_1, \cdots, x_{k+1}\}$. Hence, condition (ii) is also satisfied.

By Lemma 1 there exist functions $\delta : (0, 1) \to (0, 1)$ and $x : (0, 1) \to$ co $\{x_1, \cdots, x_{k+1}\}$ such that $\lim_{\varepsilon \to 0} \delta(\varepsilon) = 0$ and $\lim_{\varepsilon \to 0^+} x(\varepsilon) = \bar{x}$, where $\bar{x} \in$ co $\{x_1, \cdots, x_{k+1}\}$ is such that $h(\bar{x}) = 0$.

We denote $\zeta(\delta(\varepsilon)x(\varepsilon))$ by $A(\delta(\varepsilon))$. By Lemma 1, $A(\delta(\varepsilon)) \in \tilde{L}$ for all $\varepsilon \in (0, 1)$. Let

$$\frac{\varphi_i(A(\delta(\varepsilon))) - h_i(\delta(\varepsilon)\bar{x})}{\delta(\varepsilon)} = E_i^1 + E_i^2 + E_i^3 \quad \text{for } i = 0, 1, \cdots, m,$$

where

$$E_i^1 = \frac{\varphi_i(A(\delta(\varepsilon))) - h_i(A(\delta(\varepsilon)))}{\delta(\varepsilon)} \cdot \frac{\|A(\delta(\varepsilon))\|}{\|A(\delta(\varepsilon))\|},$$

$$E_i^2 = \frac{h_i(A(\delta(\varepsilon))) - h_i(\delta(\varepsilon)x(\varepsilon))}{\delta(\varepsilon)}$$

and

$$E_i^3 = \frac{h_i(\delta(\varepsilon)x(\varepsilon)) - h_i(\delta(\varepsilon)\bar{x})}{\delta(\varepsilon)}.$$

It follows from Assumptions H_1 and H_2 that $\lim E_i^1 = 0$ for $i = 0, 1, \cdots, m$, and from Lemma 2.1 in [1] that $|E_i^3| \leqq N_1 \|x(\varepsilon) - \bar{x}\|$ whenever $0 < \varepsilon \leqq \varepsilon_1$, where N_1 is some constant. Since $\lim_{\varepsilon \to 0^+} x(\varepsilon) = \bar{x}$ in co $\{x_1, \cdots, x_{k+1}\}$, $\lim_{\varepsilon \to 0^+} \|x(\varepsilon) - \bar{x}\| = 0$. Hence, $\lim_{\varepsilon \to 0^+} |E_i^3| = 0$. By Assumption H_2, there exists an $\varepsilon_2 \in (0, 1)$ such that $A(\delta(\varepsilon))$ is arbitrarily small for all $\varepsilon \in (0, \varepsilon_2)$. By Lemma 2.1 in [1] we conclude that

$$|E_i^2| \leqq N_2 \|A(\delta(\varepsilon))/\delta(\varepsilon) - x(\varepsilon)\| \quad \text{for } \varepsilon \in (0, \varepsilon_2).$$

It follows from Assumption H_2 that $\lim_{\varepsilon \to 0^+} E_i^2 = 0$. Hence,

$$\lim_{\varepsilon \to 0^+} \frac{\varphi_i(A(\delta(\varepsilon))) - h_i(\delta(\varepsilon)\bar{x})}{\delta(\varepsilon)} = 0,$$

and the proposition is proved.

2. The case of operatorial constraints. If we change the problem described in § 1 by requiring that the equality constraints, instead of being defined by a finite number of real-valued functions, are defined by an operator from L into an infinite-dimensional linear space, then we can obtain a result comparable to Theorem 1 only if we strengthen condition II.

Let $\mathscr{X}$ be a real linear space, let L be an arbitrary subset of $\mathscr{X}$, let φ_i (for $i = 0, 1, \cdots, m$) be real-valued functions defined on L, and let T be a map from L into $\mathscr{Y}$, where $\mathscr{Y}$ is some linear topological vector space. Let Θ denote the zero element in $\mathscr{Y}$, and let $L_0 = \{x \in L : Tx = \Theta\}$, $L^- = \{x \in L : \varphi_i(x) \leqq 0$ for $i = 1, \cdots, m\}$. We shall say that an element $\tilde{x} \in L_0 \cap L^-$ is optimal if $\varphi_0(\tilde{x}) \leqq \varphi_0(x)$ for all $x \in L_0 \cap L^-$.

We shall suppose that an optimal element $\tilde{x}$ exists and, without loss of generality, shall suppose that $\tilde{x} = 0$, that $\varphi_i(0) = 0$ for $i = 0, 1, \cdots, p$, and that $\varphi_i(0) < 0$ for $i = p + 1, \cdots, m$.

THEOREM 2. *Let 0 be an optimal element. Suppose that there exist a convex set $M \subset \mathscr{X}$ such that $0 \in M$, convex functions $h_i : M \to R$ such that $h_i(0) = \varphi_i(0)$ for $i = 0, 1, \cdots, m$, and a linear operator $T_0 : \mathscr{X} \to \mathscr{Y}$ with the following properties:*

I. *There exists a set $M_1 \subset M$ such that $\Theta \in \operatorname{int} T_0 M_1$ and a number $\mu > 0$ such that $h_i(x) \leqq \mu$ for all $x \in M_1$ and each $i = 0, 1, \cdots, m$.*

II. *For each $\bar{x} \in \{x \in M : T_0 x = \Theta\}$, there exist functions $\delta : (0, 1) \to (0, 1)$ and $A : \delta(0, 1) \to L_0$ such that*

$$\lim_{\varepsilon \to 0^+} \frac{\varphi_i(A(\delta(\varepsilon))) - h_i(\delta(\varepsilon)\bar{x})}{\delta(\varepsilon)} \leqq 0 \quad \text{for } i = 0, 1, \cdots, m,$$

$$\lim_{\varepsilon \to 0^+} \delta(\varepsilon) = 0.$$

Then there exist constants $\alpha_i \leqq 0$ and a functional $f \in \mathscr{X}^*$ such that

(a) $\sum_0^m \alpha_i h_i(x) + f(T_0 x) \leqq \sum_0^m \alpha_i h_i(0) = 0$ *for all* $x \in M$,

(b) $\sum_0^m \alpha_i < 0, \quad \alpha_i \varphi_i(0) = 0, \quad i = 1, \cdots, m.$

Proof. Consider the sets $M_0 = \{x \in M : T_0 x = \Theta\}$,

$$\mathscr{A} = \{(h_0(x) + \varepsilon_0, \cdots, h_m(x) + \varepsilon_m) : x \in M_0, \varepsilon_i \geqq 0 \text{ for } i = 0, \cdots, m\} \subset R^{m+1},$$

$$B = \{(\xi_0, \cdots, \xi_p, \xi_{p+1}, \cdots, \xi_m) : \xi_i < 0 \text{ for } i = 0, 1, \cdots, p\} \subset R^{m+1}.$$

Since M_0 is convex, we can show that $\mathscr{A}$ is convex by straightforward arguments similar to those used in the proof of Theorem 1. The set B is obviously convex. Let us show that $\mathscr{A} \cap B = \varnothing$. Indeed, if $\mathscr{A} \cap B \neq \varnothing$, then there exist an element $x_0 \in M_0$, numbers $\varepsilon_i^0 \geqq 0$ for $i = 0, 1, \cdots, m$ and a vector $(\xi_0^0, \cdots, \xi_{p+1}^0, \cdots, \xi_m^0) \in B$ such that $h_i(x_0) + \varepsilon_i^0 = \xi_i^0$ for $i = 0, 1, \cdots, m$. For the element x_0, by condition II, there exist functions $\delta : (0, 1) \to (0, 1)$ and $A : \delta(0, 1) \to L_0$ such that

$$\lim_{\varepsilon \to 0^+} \frac{\varphi(A(\delta(\varepsilon))) - h_i(\delta(\varepsilon)x_0)}{\delta(\varepsilon)} = \lambda_i^0 \leqq 0 \quad \text{for } i = 0, 1, \cdots, m,$$

$$\lim_{\varepsilon \to 0^+} \delta(\varepsilon) = 0.$$

Hence,

$$\varphi_i(A(\delta(\varepsilon))) \leqq h_i(\delta(\varepsilon)x_0) + \tilde{\varepsilon}_i \delta(\varepsilon)$$

with $\tilde{\varepsilon}_i \to 0$ $(\varepsilon \to 0^+)$. Since $h_i(0) = \tilde{\varphi}_i(0)$ for $i = 0, 1, \cdots, p$ and the h_i are convex,

$$\varphi_i(A(\delta(\varepsilon))) \leqq \delta(\varepsilon)(h_i(x_0) + \tilde{\varepsilon}_i) \quad \text{for } i = 0, 1, \cdots, p.$$

We have that $h_i(x_0) = \xi_i^0 - \varepsilon_i^0 < 0$ for $i = 0, 1, \cdots, p$, and there exists an $\varepsilon_1 \in (0, 1)$ such that $\varphi_i(A(\delta(\varepsilon))) < 0$ for $i = 0, 1, \cdots, p$ for all $\varepsilon \in (0, \varepsilon_1)$. Further,

$$\varphi_i(A(\delta(\varepsilon))) \leqq (1 - \delta(\varepsilon))h_i(0) + \delta(\varepsilon)h_i(x_0) + \tilde{\varepsilon}_i\delta(\varepsilon) \quad \text{for } i = p + 1, \cdots, m,$$

and since

$$\lim_{\varepsilon \to 0^+} \delta(\varepsilon) = 0 \quad \text{and} \quad h_i(0) < 0 \quad \text{for } i = p + 1, \cdots, m,$$

it follows that

$$\varphi_i(A(\delta(\varepsilon))) < 0$$

for $i = p + 1, \cdots, m$ for all $\varepsilon \in (0, \varepsilon_2)$, where $\varepsilon_2 \in (0, \varepsilon_1)$. Therefore,

$$A(\delta(\varepsilon)) \in L_0 \cap L^-$$

for all $\varepsilon \in (0, \varepsilon_2)$ and

$$\varphi_0(A(\delta(\varepsilon))) < 0 = \varphi_0(0).$$

This contradicts the assumption that 0 is an optimal element. Hence, $\mathscr{A} \cap B = \varnothing$.

Consequently, there exists a hyperplane which separates $\mathscr{A}$ and B and hence also separates $\mathscr{A}$ and the closure of B; i.e., there exist constants α_i, $i = 0, 1, \cdots, m$, such that $\sum_0^m |\alpha_i| > 0$ and such that

$$\sum_0^m \alpha_i h_i(x) + \sum_0^m \alpha_i \varepsilon_i \leqq 0 \leqq \sum_0^m \alpha_i \xi_i \tag{4}$$

whenever $x \in M_0$, $\varepsilon_i \geqq 0$ and $\xi = (\xi_0, \cdots, \xi_m) \in \bar{B}$. It follows from the inequality

$$\sum_0^m \alpha_i \xi_i \geqq 0 \quad \text{for all } \xi \in \bar{B}$$

that $\alpha_i \leqq 0$ for $i = 0, 1, \cdots, p$, and that $\alpha_i = 0$ for $i = p + 1, \cdots, m$. Setting $\varepsilon_i = 0$ for $i = 0, 1, \cdots, m$, we obtain that

$$\sum_0^m \alpha_i h_i(x) \leqq 0 \quad \text{for all } x \in M_0, \tag{5}$$

$$\sum_0^m \alpha_i < 0, \quad \alpha_i \varphi_i(0) = 0 \quad \text{for } i = 1, \cdots, m, \tag{6}$$

$$\alpha_i \leqq 0 \quad \text{for } i = 0, 1, \cdots, m.$$

Let $c(x) = \sum_0^m \alpha_i h_i(x)$. Since $\alpha_i \leqq 0$ for $i = 0, 1, \cdots, m$ and the h_i are convex, $c(x)$ is a concave function. Let

$$K = \{(c(x) - \varepsilon, T_0 x) : x \in M, \varepsilon > 0\} \subset R^1 \times \mathscr{Y}.$$

The set K is convex. Indeed, let $k_1 = (c(x_1) - \varepsilon_1, T_0 x_1)$ and $k_2 = (c(x_2) - \varepsilon_2, T_0 x_2)$ belong to K. If $0 \leqq \alpha \leqq 1$, then

$$\alpha[c(x_1) - \varepsilon_1] + (1 - \alpha)[c(x_2) - \varepsilon_2] = c(\alpha x_1 + (1 - \alpha)x_2) - \varepsilon,$$

where

$$\varepsilon - \alpha\varepsilon_1 - (1 - \alpha)\varepsilon_2 = \varepsilon' \geqq 0.$$

Since the set M is convex and the operator T_0 is linear,

$$\alpha k_1 + (1 - \alpha)k_2 = (c(x) - \varepsilon, T_0(x)),$$

where

$$x = \alpha x_1 + (1 - \alpha)x_2 \in M \quad \text{and} \quad \varepsilon > 0.$$

Hence, K is convex.

We shall show that $\operatorname{int} K \neq \varnothing$. Since the functions h_i are bounded from above on the set M (see condition I), then if we set $m = \inf_{x \in M_1} c(x)$, $m > -\infty$. Since $\Theta \in \operatorname{int} T_0 M_1$ (see condition I), there exists a neighborhood U of Θ such that $U \subset T_0 M_1$. Let $\delta_0 > 0$. For each

$$(\alpha, y) \in (-\infty, m - \delta_0) \times U,$$

there exists an $x \in M_1$ such that $T_0 x = y$ and such that $c(x) - (m - \delta_0) = \varepsilon > 0$. Hence $(-\infty, m - \delta_0) \times U \subset K$, and $\operatorname{int} K \neq \varnothing$.

We shall show that $(0, \Theta) \notin K$. Indeed, if $(0, \Theta) \in K$, then there exist an $x_0 \in M$ and an $\varepsilon_0 > 0$ such that $c(x_0) = \varepsilon_0$ and such that $T_0 x_0 = \Theta$, contradicting relation (5). It follows that there exists a nonzero functional $F \in (R \times \mathscr{Y})^*$ (see [4]) such that

$$F(k) \leqq 0 \quad \text{for all } k \in K.$$

Since $F \in (R \times \mathscr{Y})^*$, there exist a constant α and a functional $f \in \mathscr{Y}^*$ such that

$$\alpha c(x) - \alpha\varepsilon + f(T_0 x) \leqq 0$$

whenever $x \in M$ and $\varepsilon > 0$. Since $0 \in M$, $-\alpha\varepsilon \leqq 0$, which implies that $\alpha \geqq 0$. Letting $\varepsilon \to 0^+$, we obtain

$$\alpha c(x) + f(T_0 x) \leqq 0 \quad \text{for all } x \in M. \tag{7}$$

If $\alpha = 0$, then $f(T_0 x) \leqq 0$ for all $T_0 x$ with $x \in M_1$, and this implies that $f = 0$, contradicting the fact that $F = (\alpha, f) \neq 0$. Hence, $\alpha > 0$. Dividing by α in relation (7), we obtain, by virtue of (5), the desired conclusions (a) and (b) of Theorem 2.

If, in addition to the inequality constraints and the operatorial constraints, there exists a finite number of equality constraints defined in terms of real-valued functions $\varphi_{m+1}, \cdots, \varphi_{m+k}$, then we can state the following theorem.

THEOREM 3. *With the same notation as in Theorems* 1 *and* 2, *suppose that* 0 *is an optimal element, and suppose that there exist a convex set* $M \subset \mathscr{X}$ *such that* $0 \in M$, *convex functions* $h_i : M \to R$ *such that* $h_i(0) = \varphi_i(0)$ *for* $i = 0, 1, \cdots, m$, *a linear map* $h = (h_{m+1}, \cdots, h_{m+k})$ *from* $\mathscr{X}$ *into* R^k, *and a linear operator* T_0 *from* $\mathscr{X}$ *into* $\mathscr{Y}$ *with the following properties*:

I. *There exist a set* $M_1 \subset M$ *such that* $\Theta \in \operatorname{int} T_0 M_1$ *and a number* $\mu > 0$ *such that* $h_i(x) \leqq \mu$ *for* $i = 0, 1, \cdots, m$ *and* $|h_i(x)| \leqq \mu$ *for* $i = m + 1, \cdots, m + k$ *for every* $x \in M_1$.

II. *Let* $M_0 = \{x \in M : T_0 x = \Theta\}$; *then either* $\theta \notin \operatorname{int} h(M_0)$ *or, for every subset* $\{x_1, \cdots, x_{k+1}\}$ *of* M_0 *consisting of* $k + 1$ *linearly independent elements such that*

$$\theta \in \operatorname{int} \operatorname{co} \{h(x_1), \cdots, h(x_{k+1})\},$$

there exist functions $\delta:(0,1)\to(0,1)$ *and* $A:\delta(0,1)\to L_0\cap\tilde{L}$ *such that*

$$\lim_{\varepsilon\to 0^+}\delta(\varepsilon)=0,$$

$$\lim_{\varepsilon\to 0^+}\frac{\varphi_i(A(\delta(\varepsilon)))-h_i(\delta(\varepsilon)\bar{x})}{\delta(\varepsilon)}\leqq 0\quad \textit{for } i=0,1,\cdots,m,$$

where $\bar{x}$ *is the unique element in* $\operatorname{co}\{x_1,\cdots,x_{k+1}\}$ *such that* $h(\bar{x})=\theta$. *Then there exist numbers* α_i, $i=0,1,\cdots,m+k$, *and a functional* $f\in\mathscr{Y}^*$ *such that*

$$\text{(a)}\quad \sum_0^{m+k}\alpha_i h_i(x)+f(T_0x)\leqq\sum_0^{m+k}\alpha_i h_i(0)=0\quad \textit{for all } x\in M,$$

$$\text{(b)}\quad \sum_0^{m+k}|\alpha_i|>0,\quad \alpha_i\leqq 0,\quad \textit{for } i=0,1,\cdots,m,$$

$$\alpha_i\varphi_i(0)=0\quad \textit{for } i=1,\cdots,m.$$

Proof. In order to derive relations (a) and (b), we use the same method as in the proof of Theorem 2. Consider the set

$$\mathscr{A}=\{(h_0(x)+\varepsilon_0,\cdots,h_m(x)+\varepsilon_m,h(x)):x\in M_0,\ \varepsilon_i\geqq 0\quad\text{for } i=0,1,\cdots,m\}.$$

Arguing as in Theorem 1, we conclude that there exist constants α_i for $i=0,1,\cdots,m+k$ such that

$$c(x)=\sum_0^{m+k}\alpha_i h_i(x)\leqq 0\quad\text{for all } x\in M_0, \tag{8}$$

$$\sum_0^{m+k}|\alpha_i|>0,\qquad \alpha_i\varphi_i(0)=0\quad\text{for } i=1,\cdots,m, \tag{9}$$

$$\alpha_i\leqq 0\quad\text{for } i=0,1,\cdots,m.$$

Further, considering the convex xet K as in Theorem 2, we obtain the conclusions (a) and (b).

3. An application. Theorem 3, together with Lemma 1, have applications in control theory in the case of equality-type phase constraints. In these applications, the most difficult step is the verification of condition II. We shall show that, in the presence of the regularity conditions given by Gamkrelidze in [3], this condition is satisfied.

To begin, let us state the problem and some fundamental assumptions.

(a) Let $f:J\times G_1\times G_2\to R^n$ be continuous and of class C^1 in $(z,u)\in G_1\times G_2$, where $G_1\subset R^n$, $G_2\subset R^p$, $J\subset R$ are open sets, and $J\supset[0,1]$.

Let $g:G_1\to R$ be of class C^2, and let

$$q_i:[t^1,t^2]\times G_1\times G_2\to R\quad\text{for } i=1,\cdots,r<p$$

be continuous functions of class C^1 in $(z,u)\in G_1\times G_2$, where t^1 and t^2 are fixed numbers such that $0<t^1<t^2\leqq 1$.

Let U be some subset of G_2, and let, for each $(t,z)\in[t^1,t^2]\times G_1$,

$$Q(t,z)=\{v\in G_2:q_i(t,z,v)=0\text{ for } i=1,\cdots,r\}.$$

Let $\mathcal{U}$ denote the class of all piecewise-continuous functions from $[0, 1]$ into G_2. Let $\chi_i : G_1 \to R$ for $i = 0, 1, \cdots, k-1$ be given functions which are of class C^1, and let $z_0 \in R^n$.

(b) Let L' denote the set of all pairs (z, u) such that z is an absolutely continuous function from $[0, 1]$ into R^n, $u \in \mathcal{U}$, and such that

$$z(t) = z_0 + \int_0^t f(s, z(s), u(s))\, ds \quad \text{for all } t \in [0,1].$$

(c) A pair (z, u) will be called admissible if $(z, u) \in L'$, if $u(t) \in U$ for all $t \in [0, t^1) \cup (t^2, 1]$, if $u(t) \in Q(t, z(t))$ for all $t \in [t^1, t^2]$, if $g(z(t)) = 0$ for all $t \in [t^1, t^2]$, and if $\chi_i(z(1)) = 0$ for $i = 1, \cdots, k-1$. A pair $(\tilde{z}, \tilde{u})$ will be said to be optimal if it is admissible and if $\chi_0(z(1)) \geqq \chi_0(\tilde{z}(1))$ whenever (z, u) is admissible.

Our aim is to find conditions which optimal pairs must satisfy.

Let us denote $\nabla g(z) \cdot f(t, z, u)$ by $p(t, z, u)$. For any admissible pair (z, u), $g(z(t)) = 0$ for all $t \in [t^1, t^2]$ if and only if $g(z(t_1)) = 0$ and $\nabla g(z(t)) \cdot f(t, z(t), u(t)) = 0$ for all $t \in [t^1, t^2]$. Hence, an optimal pair $(\tilde{z}, \tilde{u})$ satisfies the relations $g(\tilde{z}(t_1)) = 0$ and $p(t, \tilde{z}(t), \tilde{u}(t)) = 0$ for all $t \in [t^1, t^2]$.

DEFINITION. A point $v \in Q(t, \tilde{z}(t))$ will be said to be *regular with respect to* $(t, \tilde{z}(t)) \in R^{n+1}$ if $p(t, \tilde{z}(t), v) = 0$ and if the vectors[4] $\nabla_u q_1(t, \tilde{z}(t), v), \cdots, \nabla_u q_r(t, \tilde{z}(t), v)$, $\nabla_u p(t, \tilde{z}(t), v)$ are linearly independent.

HYPOTHESIS A. The vectors $\nabla \chi_i(\tilde{z}(1))$ for $i = 0, 1, \cdots, k-1$ are linearly independent.

HYPOTHESIS B. Let $R(t) \subset Q(t, \tilde{z}(t))$ be the set of points which are regular with respect to $(t, \tilde{z}(t))$. We suppose that $\tilde{u}(t+0), \tilde{u}(t-0) \in R(t)$ for all $t \in (t^1, t^2)$, and that $\tilde{u}(t^1 + 0) \in R(t^1)$ and $\tilde{u}(t^2 - 0) \in R(t^2)$.

In order to obtain necessary conditions for optimality, we shall appeal to Theorem 3. In order to verify that the hypotheses of Theorem 3 hold, we shall make use of the following lemmas from the general theory of differential equations which are analogous to the ones used by Pallu de la Barrière in [6] and by Gamkrelidze in [3, Chap. VI]. We shall state these lemmas without proofs.

LEMMA 2. *Let (z_i, u_i) for $i = 1, \cdots, k+1$ be such that, for each i, z_i is an absolutely continuous function from $[t^1, t^2]$ into R^n, u_i is a piecewise-continuous function from $[t^1, t^2]$ into R^p, and*

$$(\alpha) \quad \frac{dz_i}{\partial t} = \frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t)) z_i(t) + \frac{\partial f}{\partial u}(t, \tilde{z}(t), \tilde{u}(t)) u_i(t),$$

$$t \in [t^1, t^2], \quad \textit{for} \quad i = 1, \cdots, k+1,$$

$$(\beta) \quad \nabla_z q_1(t, \tilde{z}(t), \tilde{u}(t)) z_i(t) + \nabla_u q_1(t, \tilde{z}(t), \tilde{u}(t)) u_i(t)$$

$$= \quad \vdots$$

$$= \nabla_z q_r(t, \tilde{z}(t), \tilde{u}(t)) z_i(t) + \nabla_u q_r(t, \tilde{z}(t), \tilde{u}(t)) u_i(t)$$

$$= \nabla_z p(t, \tilde{z}(t), \tilde{u}(t)) z_i(t) + \nabla_u p(t, \tilde{z}(t), \tilde{u}(t)) u_i(t), \quad t \in [t^1, t^2],$$

$$\textit{for } i = 1, \cdots, k+1,$$

[4] $\nabla_u q_i(t, \tilde{z}(t), v)$ denotes the vector

$$\left(\frac{\partial q_i}{\partial u_i}(t, \tilde{z}(t), v), \cdots, \frac{\partial q_i}{\partial u_r}(t, \tilde{z}(t), v) \right).$$

where $(\tilde{z}, \tilde{u})$ *is an admissible pair which satisfies Hypothesis* B *and* f, g *and the* q_i *are as previously indicated. Let* τ_0 *be some element in* $(0, 1]$. *We shall denote* $\{(\beta_1, \cdots, \beta_{k+1}) : \beta_i \geqq 0, \sum_1^{k+1} \beta_i = 1\}$ *by* P^{k+1}. *Also, we shall denote by* $o(\cdot, \cdot)$ *any function from* $[0, \tau_0] \times P^{k+1}$ *into* R^n *such that*

(γ) $o(\varepsilon, \cdot) : P^{k+1} \to R^n$ *is continuous for each* ε,

(δ) $\lim_{\varepsilon \to 0^+} \dfrac{|o(\varepsilon, \beta)|}{\varepsilon} = 0$ *uniformly in* $\beta \in P^{k+1}$.

We denote $\sum_1^{k+1} \beta_i z_i$ *by* z_β *and* $\sum_1^{k+1} \beta_i u_i$ *by* u_β *for each* $\beta \in P^{k+1}$. *Then there is an* $\varepsilon_0 \in (0, \tau_0]$ *such that, for each* $(\varepsilon, \beta) \in [0, \varepsilon_0] \times P^{k+1}$, *there exist absolutely continuous functions* $\sigma_{\varepsilon,\beta} : [t^1, t^2] \to R^n$, *and piecewise-continuous functions* $s_{\varepsilon,\beta} : [t^1, t^2] \to R^p$ *with the following properties*:

(a) *The functions* $y_{\varepsilon,\beta}$ *and* $u_{\varepsilon,\beta}$ *defined by the relations*

$$y_{\varepsilon,\beta} = \tilde{z} + \varepsilon z_\beta + \sigma_{\varepsilon,\beta}, \qquad u_{\varepsilon,\beta} = \tilde{u} + \varepsilon u_\beta + s_{\varepsilon,\beta}$$

satisfy the relations

$$\text{(i)} \quad y_{\varepsilon,\beta}(t) = \tilde{z}(t^1) + \varepsilon z_\beta(t^1) + o(\varepsilon, \beta) + \int_{t^1}^{t} f(s, y_{\varepsilon,\beta}(s), u_{\varepsilon,\beta}(s))\, ds,$$

$$\text{(ii)} \quad q_1(t, y_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t)) = \cdots = q^r(t, y_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t))$$
$$= p(t, y_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t)) = 0 \quad \text{for all } t \in [t^1, t^2];$$

(b) $\sigma_{\varepsilon,\cdot}(t) : P^{k+1} \to R^n$ *is continuous for all* $(\varepsilon, t) \in [0, \varepsilon_0] \times [t^1, t^2]$, *and*

$$\lim_{\varepsilon \to 0^+} \frac{|\sigma_{\varepsilon,\beta}(t)|}{\varepsilon} = 0 \quad \text{uniformly in } (t, \beta) \in [t^1, t^2] \times P^{k+1}.$$

LEMMA 3. *Let* p, q_i, $i = 1, \cdots, r$, *be the functions defined previously, and let* $(\tilde{z}, \tilde{u})$ *be an admissible pair which satisfies Hypothesis* B. *Let* $s \in (t^1, t^2]$ *be a point of continuity of* $\tilde{u}$, *and let* $v \in R(s)$. *Then there exist a number* $\varepsilon_0 \in (0, 1)$, *a closed ball* $S(\tilde{z}(s)) \subset G_1$ *centered at* $\tilde{z}(s)$, *and a function* $u : [s - \varepsilon_0, s] \times S(\tilde{z}(s)) \to G_2$ *such that*
(a) $q_1(t, z, u(t, z)) = \cdots = q_r(t, z, u(t, z)) = p(t, z, u(t, z)) = 0$
for all $(t, z) \in [s - \varepsilon_0, s] \times S(\tilde{z}(s))$;
(b) *the function* u *is of class* C^1 *in* z, *and* $u(s, \tilde{z}(s)) = v$.

LEMMA 4. *Let the sets* $G_1 \subset R^n$ *and* $G_2 \subset R^p$ *and the function* f *be as previously indicated, and let* $(\tilde{z}, \tilde{u})$ *be an admissible pair. Let* $s \in (t^1, t^2]$ *be a point of continuity for the function* $\tilde{u}$, *and let* u_i (*for* $i = 1, \cdots, l$) *be continuous functions from* $[s - \tau_0, s] \times S(\tilde{z}(s))$ *into* G_2 *which are of class* C^1 *in* $z \in S(\tilde{z}(s))$, *where* $\tau_0 > 0$, $s - \tau_0 > t^1$, *and* $S(\tilde{z}(s)) \subset G_1$ *is a closed ball centered at* $\tilde{z}(s)$. *For each* $\varepsilon \in [0, \tau_0]$, *let* $s_1, s_2, \cdots, s_{l-1}$ *be such that* $s - \varepsilon \leqq s_1 \leqq \cdots \leqq s_{l-1} \leqq s$. *We define the function*

$$u_\varepsilon : [s - \varepsilon, s] \times S(\tilde{z}(s)) \to G_2$$

in such a way that $u_\varepsilon(t, z) = u_{i+1}(t, z)$ *for every* $(t, z) \in (s_i, s_{i+1}) \times S(\tilde{z}(s))$ *and each* $i = 0, 1, \cdots, l - 1$, *where* $s_0 = s - \varepsilon$ *and* $s_l = s$. *Then there are numbers* $\varepsilon_0 \in (0, \tau_0]$ *and* $\rho_0 > 0$ *with the property that, for every* $\varepsilon \in [0, \varepsilon_0]$ *and every* y *satisfying* $|y - \tilde{z}(s - \varepsilon)| \leqq \rho_0$, *there exists a function*

$$z_{\varepsilon,y}(\cdot) : [s - \varepsilon, s] \to S(\tilde{z}(s))$$

such that

$$\text{(a) } z_{\varepsilon,y}(t) = y + \int_{s-\varepsilon}^{t} f(\sigma, z_{\varepsilon,y}(\sigma), u_\varepsilon(\sigma, z_{\varepsilon,y}(\sigma)))\, d\sigma \text{ for all } t \in [s-\varepsilon, s],$$

(b) $z_{\varepsilon,y}(\cdot)$ *depends continuously on* $s_0, \cdots, s_{l-1}, y$ *in the topology of uniform convergence and*

$$\lim_{\substack{\varepsilon \to 0^+ \\ y \to z(s)}} |z_{\varepsilon,y}(t) - \tilde{z}(s)| = 0.$$

LEMMA 5. *Let* $(\tilde{z}, \tilde{u})$ *be an admissible pair. Let* $s \in (t^1, t^2]$ *be a point of continuity for* $\tilde{u}$, *let* $\tau_0 > 0$ *with* $s - \tau_0 > t^1$, *and let*

$$u_i : [s-\tau_0, s] \times S(\tilde{z}(s)) \to G_2, \qquad i = 1, \cdots, l,$$

be continuous functions of class C^1 *in* $z \in S(\tilde{z}(s))$, *where* $S \subset G_1$ *is a closed ball centered at* $\tilde{z}(s)$.

Let

$$P^{k+1} = \{(\beta_1, \cdots, \beta_{k+1}) = \beta : \beta_i \geqq 0, \textstyle\sum \beta_i = 1\},$$

let $C^{j,r} \geqq 0$ *for* $j = 1, \cdots, l+1$ *and for* $r = 1, \cdots, k+1$, *and let* $C^{l+1,r} = 0$. *Further, let*

$$s_i(\varepsilon, \beta) = s - \varepsilon \sum_{j=i+1}^{l+1} \sum_{1}^{k+1} \beta_r C^{j,r}$$

for all $i = 1, \cdots, l$, $\beta \in P^{k+1}$, *and* $\varepsilon \in (0, \tau_0']$, *where*

$$\tau_0' = \min\{\tau_0, \tau_0 / \textstyle\sum C^{j,r}\}.$$

Let

$$u_{\varepsilon,\beta}(t, z) = u_{i+1}(t, z)$$

for all $t \in (s_i(\varepsilon, \beta), s_{i+1}(\varepsilon, \beta)]$, $i = 0, 1, \cdots, l-1$. *Let the functions*

$$h : [s-\tau_0, s] \times P^{k+1} \to R^n$$

and

$$o_1(\varepsilon, \cdot, \cdot) : [s-\tau_0, s] \times P^{k+1} \to R^n$$

be such that h *and* $o_1(\varepsilon, \cdot, \cdot)$ *are continuous and*

$$\lim_{\varepsilon \to 0^+} \frac{|o_1(\varepsilon, t, \beta)|}{\varepsilon} = 0 \quad \textit{uniformly in } (t, \beta) \in [s = \tau_0, s] \times P^{k+1}.$$

Then there is an $\varepsilon_0 \in (0, \tau_0]$ *such that, for all* $(\varepsilon, \beta) \in [0, \varepsilon_0] \times P^{k+1}$, *there exists a function*

$$z_{\varepsilon,\beta} : [s_0, s] \to S(\tilde{z}(s))$$

with the property that

$$\text{(a) } z_{\varepsilon,\beta}(t) = \tilde{z}(s_0) + \varepsilon h(s_0, \beta) + o_1(\varepsilon, s_0, \beta) + \int_{s_0}^{t} f(\sigma, z_{\varepsilon,\beta}(\sigma), u_{\varepsilon,\beta}(\sigma))\, d\sigma \quad \textit{for all } t \in [s_0, s],$$

(b) $z_{\varepsilon,\beta}(s) = \tilde{z}(s) + \varepsilon h(s,\beta)$

$$+ \varepsilon \sum_{j=1}^{l} \left(\sum_{1}^{k+1} \beta_r C^{j,r} \right) [f(s, \tilde{z}(s), u_j(s, \tilde{z}(s))) f(s, \tilde{z}(s), \tilde{u}(s))] + o(\varepsilon, \beta),$$

where $o(\varepsilon, \cdot)$ *is continuous from* P^{k+1} *into* R^n *and*

$$\lim_{\varepsilon \to 0^+} \frac{|o(\varepsilon, \beta)|}{\varepsilon} = 0 \quad \textit{uniformly in } \beta \in P^{k+1}.$$

Let $\mathscr{X}$ denote the linear space of all pairs of piecewise-continuous functions $(z, u) = x$, where $z: [0, 1] \to R^n$ and $u: [0, 1] \to R^p$. Let $L = L' - (\tilde{z}, \tilde{u})$, where the set L' is defined as before, and $(\tilde{z}, \tilde{u})$ is an optimal pair. Let

$$\varphi_0(x) = \chi_0(\tilde{z}(1) + z(1)) - \chi_0(\tilde{z}(1)),$$

$$\varphi_i(x) = \chi_i(\tilde{z}(1) + z(1)) \quad \text{for } i = 1, \cdots, k-1,$$

$$\varphi_k(x) = g(\tilde{z}(t^1) + z(t^1)) \quad \text{for } x \in L,$$

where g and the χ_i for $i = 0, 1, \cdots, k-1$ are defined as before. Let $\mathscr{Y}$ denote the normed linear space of all piecewise-continuous functions $y: [t^1, t^2] \to R^{r+1}$, with the norm defined by

$$\|y\| = \sum_{1}^{r+1} \int_{t^1}^{t^2} |y_i(t)|\, dt.$$

Define the operator $T: \mathscr{X} \to \mathscr{Y}$ as follows:

$$(T_0 x)(t) = (\nabla_z q_1(t, \tilde{z}(t), \tilde{u}(t))z(t) + \nabla_u q_1(t, \tilde{z}(t), \tilde{u}(t))u(t), \cdots, \nabla_z p(t, \tilde{z}(t), \tilde{u}(t))z(t)$$

Let $h_i(x) = \nabla \chi_i(\tilde{z}(1)) \cdot z(1)$ for $i = 0, 1, \cdots, k-1$ and let $h_k(x) = \nabla g(\tilde{z}(t^1)) \cdot z(t^1)$. Define the operator $T_0: \mathscr{X} \to \mathscr{Y}$ as follows:

$$(T_0 x)(t) = (\nabla_z q_1(t, \tilde{z}(t), \tilde{u}(t))z(t) + \nabla_u q_1(t, \tilde{z}(t), \tilde{u}(t))u(t), \cdots, \nabla_z p(t, \tilde{z}(t), \tilde{u}(t))z(t)$$
$$+ \nabla_u p(t, \tilde{z}(t), \tilde{u}(t))u(t)), \qquad t \in [t^1, t^2].$$

We shall now define the set M.

(i) Let $I_1 \subset (0, t^1)$ denote the set of all points of continuity of the function $\tilde{u}$. Let $s \in I_1$ and $v \in U$. Define the piecewise-continuous function $z_{s,v}: [0, t^1] \to R^n$ by the formula

$$z_{s,v}(t) = \begin{cases} 0 & \text{for } 0 \leqq t < s, \\ \Phi(t, s)[f(s, \tilde{z}(s), v) - f(s, \tilde{z}(s), \tilde{u}(t))] & \text{for } s \leqq t \leqq t^1, \end{cases}$$

where $\Phi(t, s)$ is the resolvent of the equation

$$\frac{dz}{dt} = \frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t))z.$$

For any $v \geqq 1$, let $P^v = \{\beta = (\beta_1, \cdots, \beta_v): \beta_i \geqq 0, \sum \beta_i = 1\}$. Let N_1 denote the set of all functions $x = (z, 0) \in \mathscr{X}$ for which z is of the form

$$z = \sum_{1}^{v} \beta_i z_{s_i, v_i},$$

where $\beta \in P^\nu$, $\nu \geqq 1$, $s_i \in I_1$, $v_i \in U$. It is easy to see that the set N_1 is convex and that $0 \in N_1$.

(ii) Let $R(t)$, for each $t \in [t^1, t^2]$, be the set introduced in Hypothesis B. For each piecewise-continuous function $u : [t^1, t^2] \to R^p$ and each $\xi \in R^n$, let $z_{u,\xi}$ be the absolutely continuous solution of the equation

$$\frac{dz}{dt} = \frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t))z + \frac{\partial f}{\partial u}(t, \tilde{z}(t), \tilde{u}(t))u(t), \qquad t \in [t^1, t^2], \quad z(t^1) = \xi. \tag{10}$$

Let $I_2 \subset (t^1, t^2]$ be the set of all points of continuity of $\tilde{u}$. For each $s \in I_2$, $\xi \in R^n$, $r \in R(s)$, and piecewise-continuous function $u : [t^1, t^2] \to R^n$, we define the function $z_{s,r,u,\xi} : [t^1, t^2] \to R^n$ by the formula

$$z_{s,r,u,\xi}(t) = \begin{cases} z_{u,\xi} & \text{for } t^1 \leqq t < s, \\ z_{u,\xi}(s) + \Phi(t, s)[f(s, \tilde{z}(s), r) - f(s, \tilde{z}(s), \tilde{u}(s))] & \\ \qquad + \int_s^t \Phi(t, \sigma)\frac{\partial f}{\partial u}(\sigma, \tilde{z}(\sigma), \tilde{u}(\sigma))\, d\sigma & \text{for } s \leqq t \leqq t^2. \end{cases}$$

Let P^ν be determined as above for each $\nu \geqq 1$. Let N_2 denote the set of all functions $x = (z, u)$ of the form

$$z = \sum_1^\nu \beta_i z_{s_i, r_i, u_i, \xi_i}, \qquad u = \sum_1^\nu \beta_i u_i,$$

where $\beta \in P^\nu$, $\nu \geqq 1$, $\xi_i \in R^n$, $r_i \in R(s_i)$, and $s_i \in I_2$. Setting $\xi = 0$, $r = \tilde{u}(s)$ and $u = 0$, we observe that the function which vanishes identically on $[t^1, t^2]$ is in the set N_2. It is easily seen that

$$\sum_1^\nu \beta_i z_{s_i, r_i, u_i, \xi_i} = \sum_1^\nu \beta_i z_{s_i, r_i, u, \xi},$$

where $u = \sum_1^\nu \beta_i u_i$ and $\xi = \sum_1^\nu \beta_i \xi_i$ for any $\beta \in P^\nu$. We shall show that the set N_2 is convex. Let (z_1, u_1) and $(z_2, u_2) \in N_2$, and let $0 \leqq \lambda \leqq 1$. We have

$$\lambda z_1 + (1 - \lambda)z_2 = \sum_1^{\nu_1} \lambda \beta_i^1 z_{s_i^1, r_i^1, u_1, \xi_1} + \sum_1^{\nu_2} (1 - \lambda)\beta_i^2 z_{s_i^2, r_i^2, u_2, \xi_2}.$$

Let us introduce the notations

$$\begin{aligned} \xi &= \lambda \xi_1 + (1 - \lambda)\xi_2, \qquad u = \lambda u_1 + (1 - \lambda)u_2, \\ \lambda \beta_i^1 &= \beta_i, \quad s_i^1 = s_i, \quad r_i^1 = r_i \quad \text{for } i = 1, \cdots, \nu_1, \\ \lambda \beta_i^2 &= \beta_{\nu_1 + i}, \quad s_i^2 = s_{\nu_1 + i}, \quad r_i^2 = r_{\nu_1 + i} \quad \text{for } i = 1, \cdots, \nu_2. \end{aligned}$$

We have that

$$z = \lambda z_1 + (1 - \lambda)z_2 = \sum_1^{\nu_1 + \nu_2} \beta_i z_{s_i, r_i, u_i, \xi},$$

where

$$\beta \in P^{\nu_1 + \nu_2}, \quad \xi = \lambda \xi_1 + (1 - \lambda)\xi_2, \quad u = \lambda u_1 + (1 - \lambda)u_2.$$

(iii) Suppose that $t^2 < 1$. Let $I_3 \subset (t^2, 1]$ be the set of points of continuity of the function $\tilde{u}$. Let $s \in I_3$, $v \in U$, $\xi \in R^n$. We define the piecewise-continuous function $z_{s,v,\xi}\colon [t^2, 1] \to R^n$ by the formula

$$z_{s,v,\xi}(t) = \begin{cases} \Phi(t, t^2)\xi & \text{for } t^2 \leqq t \leqq s, \\ \Phi(t, t^2)\xi + \Phi(t, s)[f(s, \tilde{z}(s), v) - f(s, \tilde{z}(s), \tilde{u}(s))] & \text{for } s \leqq t \leqq 1. \end{cases}$$

Let P^v be as above. We define N_3 as the set of all functions $x = (z, 0)$ for which z has the form

$$z = \sum_1^v \beta_i z_{s_i, v_i, \xi_i},$$

where $s_i \in I_3$, $v_i \in U$, $\xi_i \in R^n$, $v \geqq 1$. Setting $v = \tilde{u}(s)$ and $\xi = 0$, we observe that the function which vanishes identically on $[t^2, 1]$ is in N_3.

It is easily seen that the set N_3 is convex.

Finally, we define the set M to be the set of all elements $x = (z, u) \in \mathscr{X}$ which satisfy the relations

$$\begin{aligned} &(z(t), u(t)) \in N_1 \quad \text{for } t \in [0, t^1), \\ &(z(t), u(t)) \in N_2 \quad \text{for } t \in [t^1, t^2], \\ &(z(t), u(t)) \in N_3 \quad \text{for } t \in (t^2, 1], \end{aligned}$$

and for which z is continuous at the points $t = t^1$ and $t = t^2$.

The set M is convex because the sets N_i, for $i = 1, 2, 3$, are. Also, $0 \in M$ because each N_i contains the function which vanishes identically on the corresponding interval. The functions $h_i(x) = \nabla\chi_i(\tilde{z}(1))z(1)$ for $i = 0, 1, \cdots, k-1$, $h_k(s) = \nabla g(\tilde{z}(t^1))z(t^1)$, and the operator T_0 are linear.

We shall show that there exists a set $M_1 \subset M$ such that $\Theta \in \operatorname{int} T_0 M_1$ and such that the h_i, $i = 0, 1, \cdots, k$, are bounded on M_1. It follows from Hypothesis B that there is a partition $t^1 < t_1 < \cdots < t_l < t^2$ of $[t^1, t^2]$ which contains all of the points of discontinuity of the function $\tilde{u}$ and which has the property that, on $[t_i, t_{i+1}]$ (for each i), the same minor of rank $k+1$ of the matrix $(\nabla_u q_1 \cdots \nabla_u q_r \nabla_u p) = K$ is different from zero (the derivatives are evaluated at the point $(t, \tilde{z}(t), \tilde{u}(t))$).

Let $y \in \mathscr{Y}$. We consider the system

$$\begin{aligned} &\nabla_z q_1(t, \tilde{z}(t), \tilde{u}(t))z + \nabla_u q_1(t, \tilde{z}(t), \tilde{u}(t))u = y_1(t), \\ &\quad\vdots \\ &\nabla_z q_r(t, \tilde{z}(t), \tilde{u}(t))z + \nabla_u q_r(t, \tilde{z}(t), \tilde{u}(t))u = y_r(t), \\ &\nabla_z p(t, \tilde{z}(t), \tilde{u}(t))z + \nabla_u p(t, \tilde{z}(t), \tilde{u}(t))u = y_{r+1}(t), \qquad t \in [t^1, t^2]. \end{aligned} \tag{11}$$

On each interval $[t_i, t_{i+1}]$, we can solve for $r+1$ components of the vector $u = (u^1, \cdots, u^p)$ as linear functions of $z, y(\cdot)$, and the other $p - r - 1$ components of u. We shall set these $p - r - 1$ components of the vector u equal to zero. The functions $u^i(t, z)$ obtained in this way are piecewise-continuous with respect to t and linear with respect to z and $y(\cdot)$. We substitute these functions into the system (10), which will then remain a nonhomogeneous linear system.

Let $z_2:[t^1,t^2]\to R^n$ be the absolutely continuous solution of the system obtained in this way, with $z(t^1)=0$. Then

$$|z_2(t^2)|\leqq \text{const.}\,\|y\|. \tag{12}$$

We substitute the function z_2 into $u^i(t,z)$ for each $i=1,\cdots,r+1$, and we obtain a piecewise-continuous function $u_2(t)=(u^1(t,z_2(t)),\cdots,u^{r+1}(t,z_2(t)),0,\cdots,0)$. Obviously, for each $y\in\mathscr{Y}$, we obtain in this manner that $(z_2,u_2)\in N_2$ ($z_2=z_{s,\tilde{u}(s),u_2,0}$). Let $y\in\mathscr{Y}$, and let $(z,u)_y$ be defined by

$$(z,u)_y(t)=\begin{cases}(0,0) & \text{for } t\in[0,t^1),\\ (z_2(t),u_2(t)) & \text{for } t\in[t^1,t^2],\\ (\Phi(t,t^2)z_2(t^2),0) & \text{for } t\in(t^2,1] \quad (\text{if } t^2<1).\end{cases}$$

Since $z_2(t^1)=0$, it follows that the function z in $(z,u)_y$ is continuous at the point $t=t^1$; it is also continuous at the point $t=t^2$ from its definition on $(t^2,1]$. Hence, $(z,u)_y\in M$ and $T_0((z,u)_y)=y$ (see (2)). Let $\varepsilon_0>0$, and let

$$V=\{y\in\mathscr{Y}:\|y\|\leqq\varepsilon_0\},\qquad M_1=\{(z,u)_y:y\in V\}.$$

Therefore, by relations (11), we have that $\Theta\in\operatorname{int}T_0M_1=\operatorname{int}V$, and by relation (12), we have that

$$|h_i(x)|=|\nabla\chi_i(\tilde{z}(1))z(1)|=|\nabla\chi_i(\tilde{z}(1))\Phi(1,t^2)z_2(t^2)|\leqq\text{const.}\,\varepsilon_0$$
$$\text{for each } i=0,1,\cdots,k-1 \quad \text{and all } x\in M_1.$$

Similarly, $h_k(x)=\nabla g(\tilde{z}(t^1))\cdot z_2(t^1)=0$ for all $(z,u)\in M_1$. Hence, the h_i, $i=0,1,\cdots,k$, are bounded on the set $M_1\subset M$.

Let $M_0=\{x:x\in M,T_0x=\Theta\}$, and let $x_1=(z_1,u_1),\cdots,x_{k+1}=(z_{k+1},u_{k+1})$ be linearly independent elements in M_0 such that

$$0\in\operatorname{int}\operatorname{co}\{h(x_1),\cdots,h(x_{k+1})\},$$

where $h=(h_1,\cdots,h_k)$. The points of discontinuity of the functions z_i are points of continuity of the function $\tilde{u}$.

Let $0<s_1<\cdots<s_m\leqq 1$ be the points of discontinuity of the functions $z_1,\cdots,z_{k+1}$. Let

$$\{s_1,\cdots,s_{m_1}\}=\{s_1,\cdots,s_m\}\cap[0,t^1).$$

On $[0,t^1]$, the set $\operatorname{co}\{z_1,\cdots,z_{k+1}\}$ is described by

$$z_\beta(t)=\begin{cases}0 & \text{for } 0\leqq t<s_1,\\ \Phi(t,s_1)\sum_{j=1}^{l_1}\left(\sum_1^{k+1}\beta_rC_1^{j,r}\right)[f(s_1,\tilde{z}(s_1),v_j^1)-f(s_1,\tilde{z}(s_1),\tilde{u}(s_1))] & \\ & \text{for } s_1\leqq t<s_2,\\ \quad\vdots & \\ z(s_{m_1}-0)+\Phi(t,s_{m_1})\sum_{j=1}^{l_{m_1}}\left(\sum_1^{k+1}\beta_rC_{m_1}^{j,t}\right)[f(s_{m_1},\tilde{z}(s_{m_1}),v_j^{m_1}) & \\ \qquad -f(s_{m_1},\tilde{z}(s_{m_1}),\tilde{u}(s_{m_1}))] & \text{for } s_{m_1}\leqq t\leqq t^1,\end{cases}$$

where $z_\beta=\sum_1^{k+1}\beta_rz_r$, $v_j^i\in U$, $C_i^{j,r}\geqq 0$ and $\beta\in P^{r+1}$.

It follows from the theorems on the continuous dependence with respect to initial conditions that there is an $\varepsilon_1 \in (0, 1)$ with the property that, for every $(\varepsilon, \beta) \in [0, \varepsilon_1] \times P^{k+1}$, there exist an absolutely continuous function $z^1_{\varepsilon,\beta}: [0, t^1] \to G_1$ and piecewise-continuous functions $u_{\varepsilon,\beta}: [0, t^1) \to U$ such that (see [3, Chap. I])

$$\frac{dz^1_{\varepsilon,\beta}}{dt} = f(t, z^1_{\varepsilon,\beta}(t), u^1_{\varepsilon,\beta}(t)), \qquad t \in [0, t^1], \quad z_{\varepsilon,\beta}(0) = z_0, \tag{13}$$

$$z^1_{\varepsilon,\beta}(t^1) = \tilde{z}(t^1) + \varepsilon z_\beta(t^1) + o(\varepsilon, \beta), \tag{14}$$

where $o(\varepsilon, \cdot): P^{k+1} \to R^n$ is continuous and

$$\lim_{\varepsilon \to 0^+} \frac{|o(\varepsilon, \beta)|}{\varepsilon} = 0 \quad \text{uniformly in } \beta \in P^{k+1}.$$

Let

$$\{s_{m_1+1}, \cdots, s_{m_2}\} = \{s_1, \cdots, s_m\} \cap [t^1, t^2],$$

We shall show that there is an $\varepsilon_2 \in (0, \varepsilon_1]$ with the property that, for every $(\varepsilon, \beta) \in [0, \varepsilon_2] \times P^{k+1}$, there exist an absolutely continuous function $z^2_{\varepsilon,\beta}: [t^1, t^2] \to G_1$ and a piecewise-continuous function $u^2_{\varepsilon,\beta}: [t^1, t^2] \to G_2$ such that

$$z^2_{\varepsilon,\beta}(t) = z_{\varepsilon,\beta}(t^1) + \int_{t^1}^{t} f(\sigma, z^2_{\varepsilon,\beta}(\sigma), u^2_{\varepsilon,\beta}(\sigma))\, d\sigma, \qquad t \in [t^1, t^2], \tag{15}$$

$$\begin{aligned} q_1(t, z^2_{\varepsilon,\beta}(t), u^2_{\varepsilon,\beta}(t)) &= \cdots = q_r(t, z^2_{\varepsilon,\beta}(t), u^2_{\varepsilon,\beta}(t)) \\ &= p(t, z^2_{\varepsilon,\beta}(t), u^2_{\varepsilon,\beta}(t)) = 0, \qquad t \in [t^1, t^2], \end{aligned} \tag{16}$$

$$z^2_{\varepsilon,\beta}(s_i) = \tilde{z}(s_i) + \varepsilon z_\beta(s_i) + o_i(\varepsilon, \beta), \tag{17}$$

where $o_i(\varepsilon, \cdot): P^{k+1} \to R^n$ is continuous and

$$\lim_{\varepsilon \to 0^+} \frac{|o_i(\varepsilon, \beta)|}{\varepsilon} = 0 \quad \text{uniformly in } \beta \in P^{k+1}$$

$$\text{for } i = 0, \cdots, m_1 + 1, \cdots, m_2 + 1,$$

where $s_0 = t^1$ and $s_1 = t^2$. We may suppose that $\tilde{u}(t^2 - 0) = \tilde{u}(t^2)$.

Our proof will proceed by induction. Let $i = 0$. If we choose $o_0(\varepsilon, \beta) = o(\varepsilon, \beta)$ (from relation (14)), then relations (15) and (17) are satisfied on $[t^1, s_0]$, i.e., at t^1. Therefore, relation (16) is satisfied. Since relations (16) are satisfied at $(t^1, \tilde{z}(t^1), \tilde{u}(t^1 + 0))$, and $\tilde{u}(t^1 + 0)$ satisfies Hypothesis B, it follows from the implicit function theorem that there is an $\varepsilon_2 \in (0, \varepsilon_1]$ with the property that, for every $(\varepsilon, \beta) \in [0, \varepsilon_2) \times P^{k+1}$, there exists a point $u^2_{\varepsilon,\beta} \in G_2$ such that relation (16) is satisfied at the point $(t^1, z^2_{\varepsilon,\beta}(t^1), u^2_{\varepsilon,\beta})$.

Now suppose that relations (15) and (16) hold on $[t^1, s_i]$, and that (17) holds for some i. The functions z_r, $r = 1, \cdots, k + 1$, are continuous on the interval $[s_i, s_{i+1})$, and, since $z_r \in M_0 \subset M$, it follows that

$$\frac{dz_r}{dt} = \frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t)) z_r(t) + \frac{\partial f}{\partial u}(t, \tilde{z}(t), \tilde{u}(t)) u_r(t), \tag{18}$$

$$
\begin{aligned}
&\nabla_z q_1(t, \tilde{z}(t), \tilde{u}(t))z_r(t) + \nabla_u q_1(t, \tilde{z}(t), \tilde{u}(t))u_r(t) \\
&\qquad = \\
&\qquad \vdots \\
&\qquad = \nabla_z q_r(t, \tilde{z}(t), \tilde{u}(t))z_r(t) + \nabla_u q_r(t, \tilde{z}(t), \tilde{u}(t))u_r(t) \\
&\qquad = \nabla_z p(t, \tilde{z}(t), \tilde{u}(t))z_r(t) + \nabla_u(t, \tilde{z}(t), \tilde{u}(t))u_r(t) \qquad \text{for all } t \in [s_i, s_{i+1}].
\end{aligned} \tag{19}
$$

By the induction hypothesis, $\varepsilon_i > 0$, $o_i(\varepsilon, \beta)$ and $z^i_{\varepsilon,\beta}$ have been determined. We choose $o(\varepsilon, \beta) = o_i(\varepsilon, \beta)$ and $\tau_0 = \varepsilon_i$. The optimal pair $(\tilde{z}, \tilde{u})$ satisfies Hypothesis B. Relations (18) and (19) coincide with relations (α) and (β) of Lemma 2 except that the interval $[t^1, t^2]$ has been replaced by the interval $[s_i, s_{i+1}]$. The function o_i in relation (17) satisfies conditions (γ) and (δ) of Lemma 2. Thus, the hypotheses of Lemma 2 are satisfied and it follows that there is an $\varepsilon^* \in (0, \varepsilon_i]$ such that, for every $(\varepsilon, \beta) \in [0, \varepsilon^*] \times P^{k+1}$, there exist an absolutely continuous function $y_{\varepsilon,\beta}$: $[s_i, s_{i+1}) \to G_1$ and a piecewise-continuous function $u_{\varepsilon,\beta}: [s_i, s_{i+1}) \to G_2$ with the following properties:

$$
y_{\varepsilon,\beta}(t) = z^i_{\varepsilon,\beta}(s_i) + \int_{s_i}^{t} f(\sigma, y_{\varepsilon,\beta}(\sigma), u_{\varepsilon,\delta}(\delta))\, d\sigma \quad \text{for all } t \in [s_i, s_{i+1}), \tag{20}
$$

$$
\begin{aligned}
q_1(t, y_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t)) &= \cdots = q_r(t, y_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t)) \\
&= p(t, y_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t)) = 0 \quad \text{for all } t \in [s_i, s_{i+1}),
\end{aligned} \tag{21}
$$

$$
\begin{aligned}
&y_{\varepsilon,\cdot}(t): P^{k+1} \to R^n \quad \text{is continuous for } (\varepsilon, t) \in [0, \varepsilon^*] \times [s_i, s_{i+1}), \\
&\lim_{\varepsilon \to 0^+} \frac{|y_{\varepsilon,\beta}(t) - \tilde{z}(t) - \varepsilon z_\beta(t)|}{\varepsilon} = 0 \quad \text{uniformly in } (t, \beta) \in [s_i, s_{i+1}) \times P^{k+1}.
\end{aligned} \tag{22}
$$

We define the functions $z^{i+1}_{\varepsilon,\beta}$ and $u^{i+1}_{\varepsilon,\beta}$ as follows:

$$
z^{i+1}_{\varepsilon,\beta}(t) = \begin{cases} z^i_{\varepsilon,\beta}(t) & \text{for } t \in [t^1, s_i], \\ y_{\varepsilon,\beta}(t) & \text{for } t \in [s_i, s_{i+1}), \end{cases} \qquad u^{i+1}_{\varepsilon,\beta}(t) = \begin{cases} u^i_{\varepsilon,\beta}(t) & \text{for } t \in [t^1, s_i], \\ u_{\varepsilon,\beta}(t) & \text{for } t \in [s_i, s_{i+1}). \end{cases}
$$

We shall extend the functions $z^{i+1}_{\varepsilon,\beta}$ and $u^{i+1}_{\varepsilon,\beta}$ from the interval $[t^1, s_{i+1})$ onto the closed interval $[t^1, s_{i+1}]$ in such a way that relations (15), (16) and (17) hold. For each $\beta \in P^{k+1}$, the function $z_\beta = \sum_1^{k+1} \beta_r z_r$ may have a jump discontinuity at $t = s_{i+1}$ of the following form:

$$
\begin{aligned}
z_\beta(s_{i+1}) = z_\beta(s_{i+1} - 0) + \sum_{j=1}^{l_{i+1}} \left(\sum_1^{k+1} \beta_r C^{j,r}_{i+1} \right) [&f(s_{i+1}, \tilde{z}(s_{i+1}), v^j_{i+1}) \\
&- f(s_{i+1}, \tilde{z}(s_{i+1}), \tilde{u}(s_{i+1}))],
\end{aligned} \tag{23}
$$

where $C^{j,r}_{i+1} \geqq 0$, $l_{i+1} \geqq 1$, and $v^j_{i+1} \in R(s_{i+1})$. Since s_{i+1} is a point of continuity of the function $\tilde{u}$, and $v^j_{i+1} \in R(s_{i+1})$ for each $j = 1, \cdots, l_{i+1}$, it follows from Lemma 3 that there exist a number $\varepsilon' \in (0, \varepsilon^*]$, a closed ball $S(\tilde{z}(s_{i+1}))$, and

functions

$$u_{i+1}^j : [s_{i+1} - \varepsilon', s_{i+1}] \times S(\tilde{z}(s_{i+1})) \to G_2 \quad \text{for } j = 1, \cdots, l_{i+1},$$

which are continuous and of class C^1 in $z \in S(\tilde{z}(s_{i+1}))$ such that:

$$\begin{aligned} u_{i+1}^j(s_{i+1}, \tilde{z}(s_{i+1})) &= v_{i+1}^j, \\ q_1(t, z, u_{i+1}^j(t, z)) = \cdots = q_r(t, z, u_{i+1}^j(t, z)) &= p(t, z, u_{i+1}^j(t, z)) = 0 \end{aligned} \tag{24}$$

for all $(t, z) \in [s_{i+1} - \varepsilon', s_{i+1}] \times S(\tilde{t}(s_{i+1}))$, for each $j = 1, \cdots, l_{i+1}$. We choose $s = s_{i+1}$, $\tau_0 = \varepsilon'$, $l = l_{i+1}$, $h(t, \beta) = z_\beta(t)$ (where $h(s_{i+1})$ is defined as $z_\beta(s_{i+1} - 0)$), $u_j = u_{i+1}^j$, $o_1(\varepsilon, t, \beta) = z_{\varepsilon,\beta}^{i+1}(t) - \tilde{z}(t) + \varepsilon z_\beta(t)$ (where $o_i(\varepsilon, s_{i+1}, \beta)$ is defined as $o_1(\varepsilon, s_{i+1} - 0, \beta)$), and $C^{j,r} = C_{i+1}^{j,r}$.

It follows from relation (22) and from the definition of the functions h and o_1 that condition (i) of Lemma 5 is satisfied. By Lemma 5, we now conclude that there exists an $\varepsilon_{i+1} \in (0, \varepsilon']$ such that, for each $(\varepsilon, \beta) \in [0, \varepsilon_{i+1}] \times P^{k+1}$, there exist functions

$$z_{\varepsilon,\beta} : \left[s_{i+1} - \varepsilon \sum_{j=1}^{l_{i+1}} \sum_{1}^{k+1} \beta_r C_{i+1}^{j,r}, s_{i+1} \right] \to S(\tilde{z}(s_{i+1}))$$

and

$$u_{\varepsilon,\beta} : \left[s_{i+1} - \varepsilon \sum_{j=1}^{l_{i+1}} \sum_{1}^{k+1} \beta_r C_{i+1}^{j,r}, s_{i+1} \right] \times S(\tilde{z}(s_{i+1})) \to G_2$$

such that

$$z_{\varepsilon,\beta}(t) = z_{\varepsilon,\beta}^{i+1}(s_{i+1}(\varepsilon)) + \int_{s_{i+1}(\varepsilon)}^{t} f(\sigma, z_{\varepsilon,\beta}(\sigma), u_{\varepsilon,\beta}(\sigma))\, d\sigma \quad \text{for all } t \in [s_{i+1}(\varepsilon), s_{i+1}], \tag{25}$$

where

$$s_{i+1}(\varepsilon) = s_{i+1} - \varepsilon \sum_{j=1}^{l_{i+1}} \sum_{1}^{k+1} \beta_r C_{i+1}^{j,r},$$

such that

$$\begin{aligned} z_{\varepsilon,\beta}(s_{i+1}) &= \tilde{z}(s_{i+1}) + \varepsilon z_\beta(s_{i+1} - 0) \\ &\quad + \varepsilon \sum_{j=1}^{l_{i+1}} \left(\sum_{1}^{k+1} \beta_r C_{i+1}^{j,r} \right) [f(s_{i+1}, \tilde{z}(s_{i+1}), v_{i+1}^j) \\ &\qquad - f(s_{i+1}, \tilde{z}(s_{i+1}), \tilde{u}(s_{i+1}))] + o_{i+1}(\varepsilon, \beta), \end{aligned} \tag{26}$$

where $o_{i+1}(\varepsilon, \cdot) : P^{k+1} \to R^n$ is continuous and

$$\lim_{\varepsilon \to 0^+} \frac{|o_{i+1}(\varepsilon, \beta)|}{\varepsilon} = 0 \quad \text{uniformly in } \beta \in P^{k+1},$$

and such that

$$
\begin{aligned}
q_1(t, z_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t, z_{\varepsilon,\beta}(t))) = \cdots = q_r(t, z_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t)) \\
= p(t, z_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t, z_{\varepsilon,\beta}(t))) = 0 \\
\text{for all } t \in \left[s_{i+1} - \varepsilon \sum_j \sum_r \beta_r C_{i+1}^{j,r}, s_{i+1} \right].
\end{aligned} \tag{27}
$$

Let

$$
u_{\varepsilon,\beta}(t) = \begin{cases} u_{\varepsilon,\beta}^{i+1}(t) & \text{for } t \in \left[t^1, s_{i+1} - \varepsilon \sum_j \sum_r \beta_r C_{i+1}^{j,r} \right), \\ u_{\varepsilon,\beta}(t, z_{\varepsilon,\beta}(t)) & \text{for } t \in \left[s_{i+1} - \varepsilon \sum_j \sum_r \beta_r C_{i+1}^{j,r}, s_{i+1} \right], \end{cases}
$$

$$
y_{\varepsilon,\beta}(t) = \begin{cases} z_{\varepsilon,\beta}^{i+1}(t) & \text{for } t \in \left[t^1, s_{i+1} - \varepsilon \sum_j \sum_r \beta_r C_{i+1}^{j,r} \right], \\ z_{\varepsilon,\beta}(t) & \text{for } t \in \left[s_{i+1} - \varepsilon \sum_j \sum_r \beta_r C_{i+1}^{j,r}, s_{i+1} \right]. \end{cases}
$$

Since the functions $z_{\varepsilon,\beta}^{i+1}$ and $u_{\varepsilon,\beta}^{i+1}$ satisfy (15), (16) and (17) for all $t \in [t^1, s_{i+1})$, the induction argument is complete (see (25), (26) and (27)). Hence, there is an $\varepsilon_2 \in (0, \varepsilon_1]$ such that, for each $(\varepsilon, \beta) \in [0, \varepsilon_2] \times P^{k+1}$, there exist an absolutely continuous function $z_{\varepsilon,\beta}:[0, t^2] \to G_1$ and a piecewise-continuous function $u_{\varepsilon,\beta}:[0, t^2] \to G_2$ with the following properties (see (13)–(17)):

$$
z_{\varepsilon,\beta}(t) = z_0 + \int_0^t f(\sigma, z_{\varepsilon,\beta}(\sigma), u_{\varepsilon,\beta}(\sigma))\, d\sigma \quad \text{for all } t \in [0, t^2], \tag{28}
$$

$$
u_{\varepsilon,\beta}(t) \in U \quad \text{for all } t \in [0, t^1), \tag{29}
$$

$$
\begin{aligned}
q_1(t, z_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t)) = \cdots = q_r(t, z_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t)) \\
= p(t, z_{\varepsilon,\beta}(t), u_{\varepsilon,\beta}(t)) = 0 \quad \text{for all } t \in [t^1, t^2],
\end{aligned} \tag{30}
$$

$$
z_{\varepsilon,\beta}(t^2) = \tilde{z}(t^2) + \varepsilon z_\beta(t^2) + \tilde{o}(\varepsilon, \beta), \tag{31}
$$

where $\tilde{o}(\varepsilon, \cdot):P^{k+1} \to R^n$ is continuous and

$$
\lim_{\varepsilon \to 0^+} \frac{|\tilde{o}(\varepsilon, \beta)|}{\varepsilon} = 0 \quad \text{uniformly in } \beta \in P^{k+1}.
$$

If $t^2 < 1$, let $\{s_{m_2+1}, \cdots, s_m\} = \{s_1, \cdots, s_m\} \cap (t^2, 1]$. On the interval $[t^2, 1]$, the set co $\{z_1, \cdots, z_{k+1}\}$ is given by

$$
z_\beta = \sum_1^{k+1} \beta_r z \quad \text{for all } \beta \in P^{k+1},
$$

where

$$z_\beta(t) = \begin{cases} z_\beta(t^2) & \text{for } t^2 \leqq t < s_{m_2+1}, \\ z_\beta(t^2) + \Phi(t, s_{m_2+1}) \sum_{j=1}^{l_{m_2+1}} \left(\sum_{1}^{k+1} \beta_r C_{m_2+1}^{j,r} \right) [f(s_{m_2+1}, \tilde{z}(s_{m_2+1}), v_{m_2+1}^j) \\ \qquad - f(s_{m_2+1}, \tilde{z}(s_{m_2+1}), \tilde{u}(s_{m_2+1}))] & \text{for } s_{m_2+1} \leqq t < s_{m_2+2}, \\ \vdots \\ z_\beta(s_m - 0) + \Phi(t, s_m) \sum_{j=1}^{l_m} \left(\sum_{1}^{k+1} \beta_r C_m^{j,r} \right) [f(s_m, \tilde{z}(s_m), v_m^j) - f(s_m, \tilde{z}(s_m), \tilde{u}(s_m))] \\ & \text{for } s_m \leqq t \leqq 1, \end{cases}$$

where $v_j^i \in U$ and $C_i^{j,r} \geqq 0$ for each $i = m_2 + 1, \cdots, m$ and each $r = 1, \cdots, k+1$.

It follows from the theorems on the continuous dependence with respect to initial conditions that there is an $\varepsilon_3 \in (0, \varepsilon_2]$ such that, for each $(\varepsilon, \beta) \in [0, \varepsilon_2] \times P^{k+1}$, there exist an absolutely continuous function $\bar{z}_{\varepsilon,\beta}:[t^2, 1] \to G_1$ and a piecewise-continuous function $\bar{u}_{\varepsilon,\beta}:[t^2, 1] \to U$ such that (see [3, Chap. I]):

$$z_{\varepsilon,\beta}(t) = z_{\varepsilon,\beta}(t^2) + \int_{t^2}^{t} f(\sigma, \bar{z}_{\varepsilon,\beta}(\sigma), \bar{u}_{\varepsilon,\beta}(\sigma))\, d\sigma \quad \text{for all } t \in [t^1, 1], \tag{32}$$

$$\bar{z}_{\varepsilon,\beta}(1) = \bar{z}(1) + \varepsilon z_\beta(1) + o(\varepsilon, \beta), \tag{33}$$

$$\bar{o}(\varepsilon, \cdot):P^{k+1} \to R^n \quad \text{is continuous,}$$

$$\lim_{\varepsilon \to 0^+} \frac{|\bar{o}(\varepsilon, \beta)|}{\varepsilon} = 0 \quad \text{uniformly in } \beta \in P^{k+1}, \tag{34}$$

where $z_{\varepsilon,\beta}:[0, t^2] \to G_1$ satisfies relations (17), (19) and (20).

Let

$$y_{\varepsilon,\beta}(t) = \begin{cases} z_{\varepsilon,\beta}(t) & \text{for } (t, \varepsilon) \in [0, t^2] \times [0, \varepsilon_3], \\ \bar{z}_{\varepsilon,\beta}(t) & \text{for } (t, \varepsilon) \in [t^2, 1] \times [0, \varepsilon_3], \end{cases}$$

$$v_{\varepsilon,\beta}(t) = \begin{cases} u_{\varepsilon,\beta}(t) & \text{for } (t, \varepsilon) \in [0, t^2] \times [0, \varepsilon_3], \\ \bar{u}_{\varepsilon,\beta}(t) & \text{for } (t, \varepsilon) \in (t^2, 1] \times [0, \varepsilon_3]. \end{cases}$$

It follows from (17), (18), (19) and (21) that

$$[(y_{\varepsilon,\beta}, v_{\varepsilon,\beta}) - (\tilde{z}, \tilde{u})] \in L,$$

and it follows from (19) that

$$T(y_{\varepsilon,\beta}, v_{\varepsilon,\beta}) = \Theta$$

for each $(\varepsilon, \beta) \in [0, \varepsilon_3] \times P^{k+1}$. Hence

$$[(y_{\varepsilon,\beta}, v_{\varepsilon,\beta}) - (\tilde{z}, \tilde{u})] \in L_0.$$

Let the function

$$\zeta:\text{co}\,\{0, x_1, \cdots, x_{k+1}\} \to L_0 = \{x \in L: Tx = \Theta\}$$

be defined as follows:

$$\zeta(\delta \cdot x_\beta) = \begin{cases} (y_{\delta,\beta}, v_{\delta,\beta}) - (\tilde{z}, \tilde{u}) & \text{for all } \delta \in [0, \varepsilon_3], \\ 0 & \text{for } \delta \in (\varepsilon_3, 1], \end{cases}$$

where

$$x_\beta = \sum_1^{k+1} \beta_r x_r \quad \text{for all } \beta \in P^{k+1}.$$

Let $\varphi = (\varphi_1, \cdots, \varphi_k)$. Then,

$$\varphi(\zeta(\delta \cdot x_\beta)) = (\chi(y_{\delta,\beta}(1)), g(y_{\delta,\beta}(t^1))),$$

where $\chi = (\chi_1, \cdots, \chi_{k-1})$. Since $y_{\delta,.}(1): P^{k+1} \to G_1$ and $y_{\delta,.}(t^1): P^{k+1} \to G_1$ are continuous (see (33), (34) and (17)), and $\chi: G_1 \to R^{k-1}$ and $g: G_1 \to R$ are continuous, it follows that

$$\varphi(\zeta(\delta(\cdot))): \text{co}\,\{x_1, \cdots, x_{k+1}\} \to R^k$$

is continuous ($x_\beta \to x_{\beta_0}$ is into $\text{co}\{x_1, \cdots, x_{k+1}\}$ if $\beta \to \beta_0$ is into P^{k+1}). Since χ and g are of class C^1 and

$$\chi_i(\tilde{z}(1)) = g(\tilde{z}(t^1)) = 0,$$

it follows that

$$\begin{aligned} \frac{\chi_i(y_{\delta,\beta}(1))}{\delta} &= \frac{\chi_i(y_{\delta,\beta}(1)) - \chi_i(\tilde{z}(1))}{\delta} \\ &= \nabla\chi_i(\tilde{z}(1)) \frac{y_{\delta,\beta}(1) - \tilde{z}(1)}{\delta} + \varepsilon_i(\delta), \\ \frac{g(y_{\delta,\beta}(t^1))}{\delta} &= \frac{g(y_{\delta,\beta}(t^1)) - g(\tilde{z}(t^1))}{\delta} \\ &= \nabla g(\tilde{z}(t^1)) \frac{y_{\delta,\beta}(t^1) - \tilde{z}(t^1)}{\delta} + \varepsilon_k(\delta), \end{aligned} \tag{35}$$

where

$$\lim_{\delta \to 0^+} \varepsilon_i(\delta) = 0 \quad \text{for } i = 0, 1, \cdots, k.$$

Further,

$$\begin{aligned} &\frac{\varphi(\zeta(\delta \cdot x_\beta)) - \delta h(x_\beta)}{\delta} \\ &= \left(\frac{\chi(y_{\delta,\beta}(1)) - \delta\nabla\chi(\tilde{z}(1)) z_\beta(1)}{\delta}, \frac{g(y_{\delta,\beta}(1)) - \delta\nabla g(\tilde{z}(t^1)) z_\beta(t^1)}{\delta} \right) \\ &= \left(\nabla\chi(\tilde{z}(1)) \frac{y_{\delta,\beta}(1) - \tilde{z}(1) - \delta z_\beta(1)}{\delta} + \varepsilon(\delta), \nabla g(\tilde{z}(1)) \frac{y_{\delta,\beta}(t^1) - \tilde{z}(t^1) - \delta z_\beta(t^1)}{\delta} + \varepsilon_k(\delta) \right), \end{aligned}$$

where

$$\varepsilon(\delta) = (\varepsilon_1(\delta), \cdots, \varepsilon_{k-1}(\delta)).$$

From relations (33) and (34), we have that

$$\lim_{\delta \to 0^+} \frac{|y_{\delta,\beta}(1) - \tilde{z}(1) - \delta z_\beta(1)|}{\delta} = 0 \quad \text{uniformly in } \beta \in P^{k+1},$$

and from relation (17) we have

$$\lim_{\delta \to 0^+} \frac{|y_{\delta,\beta}(t^1) - \tilde{z}(t^1) - \delta z_\beta(t^1)|}{\delta} = 0 \quad \text{uniformly in } \beta \in P^{k+1}.$$

Hence,

$$\lim_{\delta \to 0^+} \frac{|\varphi(\zeta(\delta x)) - \delta h(x)|}{\delta} = 0 \quad \text{uniformly in } x \in \operatorname{co}\{x_1, \cdots, x_{k+1}\}.$$

Let $(\bar{z}, \bar{u}) = \bar{x} \in \operatorname{co}\{x_1, \cdots, x_{k+1}\}$ be such that $h(\bar{x}) = \theta$. The hypotheses of Lemma 1 are satisfied, so that there exist functions $\delta:(0, 1) \to (0, 1)$ and $x:(0, 1) \to \operatorname{co}\{x_1, \cdots, x_{k+1}\}$ such that

$$\lim_{\varepsilon \to 0^+} x(\varepsilon) = \bar{x} \in \operatorname{co}\{x_1, \cdots, x_{k+1}\} \quad \text{and} \quad \lim_{\varepsilon \to 0^+} \delta(\varepsilon) = 0, \tag{36}$$

$$\varphi(\zeta(\delta)\varepsilon(x(\varepsilon))) \in \tilde{L}_0 \quad \text{for each } \varepsilon \in (0, 1). \tag{37}$$

Let us denote $\zeta(\delta(\varepsilon)x(\varepsilon))$ by $A(\delta(\varepsilon))$.

Since $x(\varepsilon) \in \operatorname{co}\{x_1, \cdots, x_{k+1}\}$, it follows that $z(\varepsilon) \in \operatorname{co}\{z_1, \cdots, z_{k+1}\}$, where $x(\varepsilon) = (z(\varepsilon), u(\varepsilon))$. Hence,

$$z(\varepsilon) = \sum_1^{k+1} \beta_r(\varepsilon) z_r = z_{\beta,\varepsilon},$$

and it follows from relation (36) that

$$\lim_{\varepsilon \to 0^+} |z_{\beta,\varepsilon}(1) - \bar{z}(1)| = 0. \tag{38}$$

From relation (35), we have that

$$\frac{\varphi_0(A(\delta(\varepsilon))) - h_0(\delta(\varepsilon)\bar{x})}{\delta(\varepsilon)} = \frac{\chi_0(y_{\delta(\varepsilon),\beta(\varepsilon)}(1)) - \chi_0(\tilde{z}(1)) - \delta(\varepsilon)\nabla\chi_0(\tilde{z}(1))\bar{z}(1)}{\delta(\varepsilon)}$$

$$= \nabla\chi_0(\tilde{z}(1))\frac{y_{\delta(\varepsilon),\beta(\varepsilon)}(1) - \tilde{z}(1) - \delta(\varepsilon)\bar{z}(1)}{\delta(\varepsilon)} + \varepsilon_0(\delta(\varepsilon)).$$

Since

$$\lim_{\varepsilon \to 0^+} \frac{|y_{\delta(\varepsilon),\beta(\varepsilon)}(1) - \tilde{z}(1) - \delta(\varepsilon)\bar{z}(1)|}{\delta(\varepsilon)}$$

$$\leqq \lim_{\varepsilon \to 0^+} \frac{|\bar{o}(\delta(\varepsilon), \beta(\varepsilon))|}{\delta(\varepsilon)} + \lim_{\varepsilon \to 0^+} |z_{\beta(\varepsilon)}(1) - \bar{z}(1)|,$$

it follows from (23), (25) and (27) that

$$\lim_{\varepsilon \to 0^+} \frac{|y_{\delta(\varepsilon),\beta(\varepsilon)}(1) - \tilde{z}(1) - \delta(\varepsilon)\bar{z}(1)|}{\delta(\varepsilon)} = 0.$$

Taking into account the relation

$$\lim_{\varepsilon \to 0^+} \varepsilon_0(\delta(\varepsilon)) = 0,$$

we see that condition II of Theorem 3 is satisfied.

Since all of the hypotheses of Theorem 3 are satisfied, it follows that there exist constants α_i, $i = 0, 1, \cdots, k$, and a functional $f \in \mathscr{Y}^*$ such that

$$\sum \alpha_i h_i(x) + f(T_0 x) \leqq 0 \quad \text{for all } x \in M, \tag{39}$$

$$\sum_0^k |\alpha_i| > 0, \qquad \alpha_0 \leqq 0. \tag{40}$$

Since $\mathscr{Y} \subset L^1([t^1, t^2], R) \times \cdots \times L^1([t^1, t^2], R)$ (algebraically and topologically), there exist functions $v_1, \cdots, v_r, \lambda \in L^\infty([t^1, t^2], R)$ such that

$$f(y) = -\sum_1^r \int_{t^1}^{t^2} y_i(t) v_i(t)\, dt - \int_{t^1}^{t^2} y_{r+1}(t)\lambda(t)\, dt \quad \text{for all } y \in \mathscr{Y}.$$

By definition of the functionals h_i and of the operator T_0, relations (29) and (30) yield that

$$\begin{aligned} &\sum_{i=0}^{k-1} \alpha_i \nabla \chi_i(\tilde{z}(1)) z(1) + \alpha_k \nabla g(\tilde{z}(t^1)) z(t^1) \\ &- \sum_1^r \int_{t^1}^{t^2} v_i(t)[\nabla_z q_i(t, \tilde{z}(t), \tilde{u}(t)) z(t) + \nabla_u q_i(t, \tilde{z}(t), \tilde{u}(t)) u(t)]\, dt \\ &\qquad - \int_{t^1}^{t^2} \lambda(t)[\nabla_z p(t, \tilde{z}(t), \tilde{u}(t)) z(t) + \nabla_u p(t, \tilde{z}(t), \tilde{u}(t)) u(t)]\, dt \leqq 0 \end{aligned} \tag{41}$$

$$\text{for all } x = (z, u) \in M,$$

$$\sum_0^k |\alpha_i| > 0, \qquad \alpha_0 \leqq 0. \tag{42}$$

Let $\psi_3 : [t^2, 1] \to R^n$ be the solution of the system

$$\frac{d\psi}{dt} = -\psi \frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t)), \qquad \psi(1) = \sum_0^{k-1} \alpha_i \nabla \chi_i(\tilde{z}(1)), \tag{43}$$

let $\psi_2 : [t^1, t^2] \to R^n$ be the solution of the system

$$\begin{aligned} &\frac{d\psi}{dt} = \psi \frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t)) + \lambda(t) \nabla_z p(t, \tilde{z}(t), \tilde{u}(t)) + \sum_1^r v_i(t) \nabla_z q_i(t, \tilde{z}(t), \tilde{u}(t)), \\ &\psi(t^2) = \psi_3(t^2), \end{aligned} \tag{44}$$

and let $\psi_1 : [0, t^1] \to R^n$ be the solution of the system

$$\frac{d\psi}{dt} = -\psi \frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t)), \qquad \psi(t^1) = \psi_2(t^1) + \alpha_k \nabla g(\tilde{z}(t^1)), \tag{45}$$

where the α_i and the functions v_i, λ satisfy (41) and (42).

For each $s \in (t^2, 1]$ which is a point of continuity for $\tilde{u}$ and each $v \in U$, let $x_{s,v} = (z_{s,v}, 0) \in M$, where

$$z_{s,v}(t) = \begin{cases} 0 & \text{for } 0 \leqq t < s, \\ \Phi(t, s)[f(s, \tilde{z}(s), v) - f(s, \tilde{z}(s), \tilde{u}(s))] & \text{for } s \leqq t \leqq 1. \end{cases}$$

By virtue of (43) and (41), we obtain that

$$\psi_3(1) z_{s,v}(1) \leqq 0 \quad \text{for all } v \in U \tag{46}$$

and for all $s \in (t^2, 1]$ which are points of continuity for $\tilde{u}$. Since

$$\psi_3(s) = \psi_3(1)\Phi(1, s) \quad \text{for all } s \in [t^2, 1],$$

we obtain that

$$\begin{aligned} &\psi_3(1)\Phi(1, s)[f(s, \tilde{z}(s), v) - f(s, \tilde{z}(s), \tilde{u}(s))] \\ &\qquad = \psi_3(s)[f(s, \tilde{z}(s), v) - f(s, \tilde{z}(s), \tilde{u}(s))] \leqq 0 \quad \text{for all } v \in U \end{aligned} \tag{47}$$

and for all $s \in (t^2, 1]$ which are points of continuity of $\tilde{u}$.

Therefore, from (47), we conclude that

$$\psi_3(t) f(t, \tilde{z}(t), \tilde{u}(t)) = \max_{v \in U} \psi_3(t) f(t, \tilde{z}(t), v) \tag{48}$$

for all $t \in (t^2, 1]$ which are points of continuity of $\tilde{u}$.

For each point of continuity $s \in (t^1, t^2]$ of $\tilde{u}$ and each $r \in R(s)$, let the function $z_{s,r}$ be defined as follows:

$$z_{s,r}(t) = \begin{cases} 0 & \text{for } 0 \leqq t < s, \\ \Phi(t, s)[f(s, \tilde{z}(s), r) - f(s, \tilde{z}(s), \tilde{u}(s))] & \text{for } s \leqq t \leqq 1. \end{cases}$$

Because $z_{s,r}$ is continuous at $t = t^1$ and at $t = t^2$, $x_{s,r} = (z_{s,r}, 0) \in M$. By virtue of (43) and (44), relation (41) yields

$$\begin{aligned} 0 &\geqq \psi_3(t^2) z_{s,r}(t^2) - \int_{t^1}^{t^2} [\lambda(t) \nabla_z p(t, \tilde{z}(t), \tilde{u}(t)) + \sum v_i(t) \nabla_z q_i(t, \tilde{z}(t), \tilde{u}(t))] z_{s,r}(t)\, dt \\ &= \psi_3(t^2) z_{s,r}(t^2) - \int_{t^1}^{t^2} \left[\frac{d\psi_2}{dt} + \psi_2 \frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t)) \right] z_{s,r}(t)\, dt \\ &= \psi_3(t^2) z_{s,r}(t^2) - \psi_2(t^2) z_{s,r}(t^2) + \psi_2(s)[f(s, \tilde{z}(s), r) - f(s, \tilde{z}(s), \tilde{u}(s))] \\ &= \psi_2(s)[f(s, \tilde{z}(s), r) - f(s, \tilde{z}(s), \tilde{u}(s))]. \end{aligned} \tag{49}$$

Hence,

$$\psi_2(t) f(t, \tilde{z}(t), \tilde{u}(t)) = \max_{r \in R(t)} \psi_2(t) f(t, \tilde{z}(t), r) \tag{50}$$

for all $t \in (t^1, t^2]$ which are points of continuity of $\tilde{u}$. If $\tilde{u}(t^1 + 0) = \tilde{u}(t^1)$, then relation (50) holds also for $t = t^1$. Hence,

$$\psi_2(t) f(t, \tilde{z}(t), \tilde{u}(t)) = \max_{r \in R(t)} \psi_2(t) f(t, \tilde{z}(t), r) \tag{51}$$

for all $t \in [t^1, t^2]$ which are points of continuity of $\tilde{u}$. For each point of continuity

$s \in (0, t^1)$ of $\tilde{u}$ and each $v \in U$, let us define the function $z_{s,v}$ as follows:

$$z_{s,v}(t) = \begin{cases} 0 & \text{for } 0 \leqq t < s, \\ \Phi(t, s)[f(s, \tilde{z}(s), v) - f(s, \tilde{z}(s), \tilde{u}(s))] & \text{for } s \leqq t \leqq 1. \end{cases}$$

Since $z_{s,v}$ is continuous at $t = t^1$ and at $t = t^2$, it follows that $x_{s,v} = (z_{s,v}, 0) \in M$. By virtue of (43) and (45), relation (41) yields

$$\begin{aligned} 0 \geqq \psi_2(t^2) z_{s,v}(t^2) - \int_{t^1}^{t^2} \left[\frac{d\psi_2}{dt} + \psi_2 \frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t))\right] z_{s,v}(t)\, dt \\ + \alpha_k \nabla g(\tilde{z}(t^1)) z_{s,v}(t^1) \\ = \psi_2(t^1) z_{s,v}(t^1) + \alpha_k \nabla g(\tilde{z}(t^1)) z_{s,v}(t^1) \\ = [\psi_2(t^1) + \alpha_k \nabla g(\tilde{z}(t^1))] \Phi(t^1, s)[f(s, \tilde{z}(s), v) - f(s, \tilde{z}(s), \tilde{u}(s))] \\ = \psi_1(s)[f(s, \tilde{z}(s), v) - f(s, \tilde{z}(s), \tilde{u}(s))]. \end{aligned} \tag{52}$$

Hence,

$$\psi_1(t) f(t, \tilde{z}(t), \tilde{u}(t)) = \max_{v \in U} \psi_1(t) f(t, \tilde{z}(t), v) \tag{53}$$

for all $t \in (0, t^1)$ which are points of continuity of $\tilde{u}$. It is obvious that if $\tilde{u}(0) = \tilde{u}(0 + 0)$, the relation (53) holds also for $t = 0$. Therefore,

$$\psi_1(t) f(t, \tilde{z}(t), \tilde{u}(t)) = \max_{v \in U} \psi_1(t) f(t, \tilde{z}(t), v) \tag{54}$$

for all t which are points of continuity of $\tilde{u}$.

LEMMA 6. *Let*

$$\psi(t) = \begin{cases} \psi_1(t) & \text{for } t \in [0, t^1], \\ \psi_2(t) & \text{for } t \in (t^1, t^2], \\ \psi_3(t) & \text{for } t \in [t^2, 1], \end{cases}$$

where ψ_1, ψ_2 and ψ_3 satisfy systems (45), (44) *and* (43), *respectively. Then the function $\psi : [0, 1] \to R^n$ is not identically equal to zero.*

Proof. Suppose that $\psi(t) = 0$ for all $t \in [0, 1]$. Then

$$\psi(1) = \sum_0^{k-1} \alpha_i \nabla \chi_i(\tilde{z}(1)) = 0,$$

and it follows from Hypothesis A that $\alpha_i = 0$ for $i = 0, 1, \cdots, k - 1$. Since $\psi(t^1) = \psi(t^1 + 0) = 0$, $\alpha_k \nabla g(\tilde{z}(t^1)) = 0$. Further, because $|\nabla g(\tilde{z}(t))| \neq 0$ for all $t \in [t^1, t^2]$ (see Hypothesis B), it follows that $\alpha_k = 0$. Hence $\alpha_i = 0$ for $i = 0, 1, \cdots, k$, contradicting relation (31).

LEMMA 7. *Let $\psi_2 : [t^1, t^2] \to R^n$ be the solution of system* (44). *Then the functions λ and v_i, $i = 1, \cdots, r$, are piecewise-continuous, and*

$$\psi_2(t) \nabla_u f(t, \tilde{z}(t), \tilde{u}(t)) = \lambda(t) \nabla_u p(t, \tilde{z}(t), \tilde{u}(t)) + \sum_1^r v_i(t) \nabla_u q_i(t, \tilde{z}(t), \tilde{u}(t))$$

for all $t \in [t^1, t^2]$.

Proof. Let $u:[t^1, t^2] \to R^p$ be piecewise-continuous, and let

$$z(t) = \begin{cases} 0 & \text{for } 0 \leqq t \leqq t^1, \\ \displaystyle\int_{t^1}^{t} \Phi(t, s)\nabla_u f(s, \tilde{z}(s), \tilde{u}(s))u(s)\, ds & \text{for } t^1 \leqq t \leqq t^2, \\ \Phi(t, t^2)z(t^2) & \text{for } t^2 \leqq t \leqq 1, \end{cases}$$

so that the function $z:[0, 1] \to R^n$ is absolutely continuous. Further, define x as follows:

$$x = (z, u) = \begin{cases} (0, 0) & \text{for } 0 \leqq t \leqq t^1, \\ (z(t), u(t)) & \text{for } t^1 < t \leqq t^2, \\ (z(t), 0) & \text{for } t^2 < t \leqq 1. \end{cases}$$

Obviously, $x \in M$, and relation (30) yields that

$$\begin{aligned} \psi_2(t^2)z(t^2) - \int_{t^1}^{t^2} \left[\lambda(t)\nabla_z p(t, \tilde{z}(t), \tilde{u}(t)) + \sum_1^r v_i(t)\nabla_z q_i(t, \tilde{z}(t), \tilde{u}(t)) \right] z(t)\, dt \\ - \int_{t^1}^{t^2} \left[\lambda(t)\nabla_u p(t, \tilde{z}(t), \tilde{u}(t)) + \sum_1^r v_i(t)\nabla_u q_i(t, \tilde{z}(t), \tilde{u}(t)) \right] u(t)\, dt \leqq 0 \end{aligned} \tag{55}$$

for all $x = (z, u)$. By virtue of (44), relation (55) implies that

$$\begin{aligned} \int_{t^1}^{t^2} \Big[\psi_2(t)\nabla_u f(t, \tilde{z}(t), \tilde{u}(t)) - \lambda(t)\nabla_u p(t, \tilde{z}(t), \tilde{u}(t)) \\ - \sum_1^r v_i(t)\nabla_u q_i(t, \tilde{z}(t), \tilde{u}(t)) \Big] u(t)\, dt \leqq 0 \end{aligned} \tag{56}$$

for all piecewise-continuous functions $u:[t^1, t^2] \to R$. Since (56) holds for both $\pm u(t)$, we have that

$$\int_{t^1}^{t^2} F(t)u(t)\, dt = 0 \tag{57}$$

for all piecewise-continuous functions $u:[t^1, t^2] \to R^p$, where

$$F(t) = \psi_2(t)\nabla_u f(t, \tilde{z}(t), \tilde{u}(t)) = \lambda(t)\nabla_u p(t, \tilde{z}(t), \tilde{u}(t)) - \sum_1^r v_i(t)\nabla_u q_i(t, \tilde{z}(t), \tilde{u}(t)).$$

Each component F^i, $i = 1, \cdots, p$, of the function F defines a continuous, linear functional φ_i on the space $C([t^1, t^2], R)$ through the relation

$$\varphi_i(c) = \int_{t^1}^{t^2} F^i(t)c(t)\, dt.$$

Now relation (57) implies that $\varphi_i = 0$ for $i = 1, \cdots, p$, so that

$$\int_{t^1}^{t^2} |F^i(t)|\, dt = 0 \quad \text{for } i = 1, \cdots, p,$$

and we deduce that $F(t) = 0$ a.e. in $[t^1, t^2]$.

On the other hand, by Hypothesis B, the vectors

$$\nabla_u p(t, \tilde{z}(t), \tilde{u}(t)), \nabla_u q_1(t, \tilde{z}(t), \tilde{u}(t)), \cdots, \nabla_u q_r(t, \tilde{z}(t), \tilde{u}(t))$$

are linearly independent for all $t \in [t^1, t^2]$ which are points of continuity of $\tilde{u}$. Hence, there exists a unique system of piecewise-continuous functions γ_i, $i = 0, 1, \cdots, r$, such that

$$\psi_2(t)\nabla_u f(t, \tilde{z}(t), \tilde{u}(t)) = \gamma_0(t)\nabla_u p(t, \tilde{z}(t), \tilde{u}(t)) + \sum_1^r \gamma_i(t)\nabla_u q_i(t, \tilde{z}(t), \tilde{u}(t))$$

for all $t \in [t^1, t^2]$.

Since $\lambda, \nu_1, \cdots, \nu_r \in L^\infty([t^1, t^2], R)$ and $F(t) = 0$ a.e. in $[t^1, t^2]$, it follows that λ, ν_i, $i = 1, \cdots, r$, may be identified with γ_0, γ_i, $i = 1, \cdots, r$, respectively, and the proof of Lemma 7 is complete.

THEOREM 4. *Let $0 < t^1 < t^2 \leqq 1$, let $(\tilde{z}, \tilde{u})$ be an optimal pair, and suppose that Hypotheses* A *and* B *hold. Then there exist absolutely continuous functions $\psi_1 : [0, t^1] \to R^n$, $\psi_2 : [t^1, t^2] \to R^n$, and $\psi_3 : [t^2, 1] \to R^n$, piecewise-continuous functions $\lambda, \nu_1, \cdots, \nu_r : [t^1, t^2] \to R$, and constants α_i, $i = 0, 1, \cdots, k$, such that:*

(a)
$$\frac{d\psi_3(t)}{dt} = -\psi_3(t)\frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t)) \quad \text{for } t \in [t^2, 1],$$

$$\psi_3(1) = \sum_1^{k-1} \alpha_i \nabla \chi_i(\tilde{z}(1)),$$

$$\psi_3(t) f(t, \tilde{z}(t), \tilde{u}(t)) = \max_{v \in U} \psi_3(t) f(t, \tilde{z}(t), v)$$

for all $t \in (t^2, 1]$ which are points of continuity of $\tilde{u}$;

(b)
$$\frac{d\psi_2(t)}{dt} = -\psi_2(t)\frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t)) + \lambda(t)\nabla_z p(t, \tilde{z}(t), \tilde{u}(t)) + \sum_1^r \nu_i(t)\nabla_z q_i(t, \tilde{z}(t), \tilde{u}(t))$$

$$\text{for } t \in [t^1, t^2], \quad \psi_2(t^2) = \psi_3(t^2),$$

$$\psi_2(t) f(t, \tilde{z}(t), \tilde{u}(t)) = \max_{r \in R(t)} \psi_2(t) f(t, \tilde{z}(t), r)$$

for all $t \in [t^1, t^2]$, which are points of continuity of $\tilde{u}$, where $R(t)$ is as defined in Hypothesis B,

$$\psi_2(t)\nabla_u f(t, \tilde{z}(t), \tilde{u}(t)) = \lambda(t)\nabla_u p(t, \tilde{z}(t), \tilde{u}(t)) + \sum_1^r \nu_i(t)\nabla_u q_i(t, \tilde{z}(t), \tilde{u}(t))$$

for $t \in [t^1, t^2]$;

(c)
$$\frac{d\psi_1(t)}{dt} = -\psi_1(t)\frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t)) \quad \text{for } t \in [0, t^1],$$

$$\psi_1(t^1) = \psi_2(t^1) + \alpha_k \nabla g(\tilde{z}(t^1)),$$

$$\psi_1(t) f(t, \tilde{z}(t), \tilde{u}(t)) = \max_{v \in U} \psi_1(t) f(t, \tilde{z}(t), v)$$

for all $t \in [0, t^1)$ *which are points of continuity of* $\tilde{u}$;

(d) *the function* ψ *defined by*

$$\psi(t) = \begin{cases} \psi_1(t) & \text{for } 0 \leqq t \leqq t^1, \\ \psi_2(t) & \text{for } t^1 < t \leqq t^2, \\ \psi_3(t) & \text{for } t^2 \leqq t \leqq 1 \end{cases}$$

is nonzero, and

$$\sum_0^k |\alpha_i| > 0, \qquad \alpha_0 \leqq 0.$$

Proof. Relations (a) follow from (43) and (48). System (44), Lemma 7 and relation (51) give rise to the relations (b). System (45) and relation (54) yield the relations (c). Relation (d) follows from Lemma 6 and from relation (42).

THEOREM 5. *Let* $t^2 = 1$, *and let* $(\tilde{z}, \tilde{u})$ *be an optimal pair which satisfies Hypothesis* B. *If the vectors* $\nabla\chi_0(\tilde{z}(1)), \cdots, \nabla\chi_{k-1}(\tilde{z}(1)), \nabla g(\tilde{z}(1))$ *are linearly independent, then there exist an absolutely continuous function* $\psi : [0, 1] \to R_n$, *piecewise-continuous functions* λ, v_i $(i = 1, \cdots, r) : [t^1, 1] \to R$, *and constants* α_i, $i = 0, 1, \cdots, k$, *with* $\alpha_0 \leqq 0$ *such that*:

$$\text{(b)} \quad \frac{d\psi}{dt} = -\psi \frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t)) + \lambda(t)\nabla_z p(t, \tilde{z}(t), \tilde{u}(t)) + \sum_1^r v_i(t)\nabla_z q_i(t, \tilde{z}(t), \tilde{u}(t))$$

for $t \in [t^1, 1]$,

$$\psi(1) = \sum_0^{k-1} \alpha_i \nabla\chi_i(\tilde{z}(1)) + \alpha_k \nabla g(\tilde{z}(1)),$$

$$\psi(t) f(t, \tilde{z}(t), \tilde{u}(t)) = \max_{r \in R(t)} \psi(t) f(t, \tilde{z}(t), r)$$

for all $t \in [t^1, t^2]$ *which are points of continuity of* $\tilde{u}$,

$$\psi(t)\nabla_u f(t, \tilde{z}(t), \tilde{u}(t)) = \lambda(t)\nabla_u p(t, \tilde{z}(t), \tilde{u}(t)) + \sum_1^r v_i(t)\nabla_u q_i(t, \tilde{z}(t), \tilde{u}(t));$$

$$\text{(c)} \quad \frac{d\psi}{dt} = -\psi \frac{\partial f}{\partial z}(t, \tilde{z}(t), \tilde{u}(t)) \quad \text{for } t \in [0, t^1],$$

$$\psi(t) f(t, \tilde{z}(t), \tilde{u}(t)) = \max_{v \in U} \psi(t) f(t, \tilde{z}(t), v)$$

for all $t \in [0, t^1)$ *which are points of continuity of* $\tilde{u}$;

$$\text{(d)} \quad |\psi(t)| \neq 0 \quad \text{for } t \in [0, 1].$$

The theorem follows in a straightforward way from Theorem 4.

Remark 2. In [3] Gamkrelidze considered an optimal control problem with inequality-type phase coordinate restrictions $(g(z) \leqq 0)$. The method used by Gamkrelidze in order to obtain his jump condition at $t = t^1$ (the same as relations (b) and (c) in Theorem 4) was based on the possibility of considering variations of the form $\tilde{z}(t^1) + \varepsilon\delta\xi_0$, where $\nabla g(\tilde{z}(t^1)) \cdot \delta\xi_0 < 0$. In our case, where the phase restriction is of the form $g(z) = 0$, the method used in [3] cannot be applied to obtain the relations (b) and (c) in Theorem 4.

Remark 3. In order to consider restrictions of the type $q_i(t, z, u) = 0$, $i = 0, 1, \cdots, r$, Hypothesis B must be strengthened by requiring that for every triple (t, z, u) which satisfies $q_i(t, z, u) = 0$ for $i = 0, 1, \cdots, r$, the vectors $\nabla_u q_0(t, z, u)$, $\cdots, \nabla_u q_r(t, z, u)$ must be linearly independent. This assumption was made in [4] and [5].

Acknowledgment. The author is indebted to Professor L. W. Neustadt for pointing out the problem and for his many valuable comments and corrections to early versions of this paper.

REFERENCES

[1] H. HALKIN, *Nonlinear nonconvex programming in an infinite dimensional space*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 10–25.

[2] H. HALKIN AND L. W. NEUSTADT, *General necessary conditions for optimization problems*, Proc. Nat. Acad. Sci. U.S.A., 56 (1966), pp. 1066–1072.

[3] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962, pp. 269–271 and pp. 303–314.

[4] T. GUINN, *The problem of bounded space coordinates as a problem of Hestenes*, this Journal, 3 (1965), pp. 181–190.

[5] L. D. BERKOVITZ, *On control problems with bounded state variables*, J. Math. Anal. Appl., 5 (1962), pp. 488–498.

[6] R. PALLU DE LA BARRIÈRE, *Optimal Control Theory*, W. B. Saunders, Philadelphia, 1967.

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

NECESSARY AND SUFFICIENT DYNAMIC PROGRAMMING CONDITIONS FOR CONTINUOUS TIME STOCHASTIC OPTIMAL CONTROL*

RAYMOND RISHEL†

1. Introduction. The purpose of this paper is to extend methods of dynamic programming to very general types of continuous time stochastic control systems. The controlled processes involved are not required to be Markovian and the control laws are allowed to be functionals on incomplete measurements of the system. The processes are assumed to stop at a random time which may depend on the control.

For each control u a value function $V_u(t)$ is constructed. This function at time t is the essential infimum, over the class of controls which agree with u up to t, of the conditional expectation of the remaining contribution to the performance, given the past measurements of the process. For each control u two spaces of functionals on the past measurements of the processes are constructed and semigroups of operators on each are defined.

A concept called relative completeness of the class of controls is defined. Under this condition an analogue of Bellman's principle of optimality is established. This principle and further weak conditions imply that the value function V_u is in the domain of the weak infinitesimal generator A_u of the appropriate semigroup. Then necessary conditions for optimality are phrased in terms of inequalities involving $A_u V_u$ and conditional expectations of the performance rates. It is shown that the existence of a function $W_u(t)$ with properties similar to the necessary conditions for $V_u(t)$ is a sufficient condition for optimality.

Dynamic programming type conditions for optimality of stochastic control systems have been given by a number of authors. Only a few of these which are closely related to this paper will be mentioned. Necessary and sufficient conditions for optimality of discrete time finite state systems with incomplete measurements have been given by Dynkin in [4]. Fleming in a series of papers [6], [7], [8] has given necessary and sufficient conditions for optimality of diffusion type systems. Kushner in [11] gave a sufficient condition for optimality phrased in terms of the infinitesimal generator of a semigroup. Krasovskii and Lidskii [10] gave a similar type of sufficient condition in a different setting. For deterministic optimal control problems, Boltyanskii [1] showed that dynamic programming conditions gave a sufficient condition for optimality. The stochastic optimality conditions of this paper are expressed in a form analogous to those of [1]. A recent survey by Fleming [9] gives a very complete exposition of the stochastic optimal control field with a very extensive list of references.

2. Controlled systems and the optimal control problem. Let (R, M, Λ) denote the real half-line $0 \leqq t < \infty$ with the usual Borel field M, and Lebesgue mea-

* Received by the editors July 14, 1969.

† Department of Mathematics, Washington State University, Pullman, Washington, and The Boeing Company, Seattle, Washington. Now at Bell Telephone Laboratories, Inc., Whippany, New Jersey 07981.

sure Λ. Let (Ω, β, P) denote a triple of a probability space Ω, Borel field β and probability measure P. Let (X, B) be a measure space and $x(t)$ a measurable stochastic process with values in X. For a stochastic process $y(t)$, let the notation $F\{y(t):0 \leqq t \leqq T\}$ denote the Borel subfield of β generated by the random variables $y(t)$ for $0 \leqq t \leqq T$. Let $N_t = F\{x(s):0 \leqq s \leqq t\}$. A finite stopping time η for $x(t)$ is a nonnegative real-valued random variable on Ω such that

$$\{\omega:\eta(\omega) \leqq t\} \in N_t \quad \text{and} \quad \|\eta\| = P\text{-ess sup}_{\omega} |\eta(\omega)| < \infty. \tag{1}$$

For a stopping time η, let $\chi(t)$ be the stochastic process defined by

$$\chi(t) = \begin{cases} 1 & \text{if } \eta \geqq t, \\ 0 & \text{if } \eta < t. \end{cases} \tag{2}$$

Recall that if F_t is an increasing collection of Borel fields and $x(t)$ a stochastic process, $x(t)$ is said to be adapted to F_t, if $x(t)$ is F_t-measurable for each fixed t. Notice that (1) and (2) imply $\chi(t)$ is a measurable stochastic process adapted to N_t.

A nonnegative real-valued measurable stochastic process $k(t)$ will be called a performance rate of the stochastic process $x(t)$, if $k(t)$ is adapted to N_t. Notice, by [12, p. 502], assuming that $k(t)$ is adapted to N_t implies that $k(t)$ is a functional on the past of $x(t)$. This remark justifies the terminology "$k(t)$ is a performance rate of $x(t)$." A performance rate will be called dominated if there is a $q > 1$ and a random variable b such that $k(t\omega) \leq b(\omega)$ and $E\{b^q\} < \infty$.

Let the measure space (X, B) be the Cartesian product of measure spaces (Y, B_1) and (Z, B_2). Then the process $x(t) = (y(t), z(t))$ is a vector of two-component processes. Consider $y(t)$ as a component that is observable, and $z(t)$ as a component that is unobservable. For each t let A_t denote a set of times less than t such that $A_{t_1} \subset A_{t_2}$ if $t_1 < t_2$. Let

$$H_t = F\{y(s), s \in A_t, \text{ and } \chi(s), 0 \leqq s < t\}. \tag{3}$$

The fields H_t will be called the information fields of the process $x(t)$ corresponding to observations made at times of A_t. Recall that $\chi(s)$ is one until the stopping time η is reached and zero afterwards. Thus, knowledge of $y(s)$, $s \in A_t$, and $\chi(s)$, $0 \leqq s < t$, is equivalent to knowledge of measurements of $y(s)$ for times s in A_t and whether the process has stopped or not. This is assumed to be the information available at time t about the process $x(t)$.

Let (U, B_3) be a measure space. A measurable stochastic process $u(t)$ with values in U will be called a "feedback control on observations made at times of A_t on the components $y(t)$ of the process $x(t)$" if $u(t)$ is adapted to H_t. Recall again by [12, p. 502] that if $u(t)$ is adapted to H_t, $u(t)$ is a functional on $\{y(s), s \in A_t, \text{ and } \chi(s), 0 \leqq s < t\}$.

A controlled system will be defined to consist of a class U of controls $u(t)$ such that corresponding to each control of the class there are:

(A) a stochastic process $x_u(t)$ representing the trajectory of the system,

(B) a finite stopping time η_u,

(C) a dominated performance rate $k_u(t)$;

and each control $u(t)$ is a feedback control on observations made at times A_t on the components $y_u(t)$ of the process $x_u(t)$. In addition it will be assumed that the

processes of the system have the following property which can be summarized intuitively by saying that the future does not influence the past. If $u(t)$ and $u^*(t)$ are controls such that $u(s) = u^*(s)$ for $0 \leqq s < t$, then

$$x_u(s) = x_{u^*}(s) \quad \text{and} \quad k_u(s) = k_{u^*}(s) \quad \text{for } 0 \leqq s < t,$$

$\{\omega : \eta_u(\omega) > t\} = \{\omega : \eta_{u^*}(\omega) > t\}$, and $\eta_u(\omega) = \eta_{u^*}(\omega)$ on this set.

Consider the optimal control problem of finding the control u in the class U of the controlled system for which

$$E\left\{\int_0^{\eta_u} k_u(s)\, ds\right\} \tag{4}$$

is a minimum.

3. Spaces of functionals on the information fields of the system. Optimality conditions for the controlled system will be defined in terms of spaces of stochastic processes adapted to the information fields of the system. For a fixed control u in U consider real-valued stochastic processes $\phi_u(t)$ and the equivalence relation among the processes $\phi_u(t)$ given by

$$\phi_u^1 \equiv \phi_u^2 \quad \text{if} \quad \chi_u(t, \omega)\phi_u^1(t, \omega) = \chi_u(t, \omega)\phi_u^2(t, \omega),$$

$\Lambda \times P$ almost everywhere. Let H_{ut}, $0 \leqq t < \infty$, denote the information fields corresponding to the control u. For a real number $p > 1$ define $L_p(u)$ to be the space of equivalence classes of measurable real-valued stochastic processes $\phi_u(t)$, each equivalence class containing an element $\phi_u(t)$ such that

(A) $\phi_u(t)$ is adapted to H_{ut}, and

(B) $\|\phi_u\| = E\left\{\int_0^{\eta_u} |\phi_u(t)|^p\, dt\right\}^{1/p} < \infty.$

For $q = p(p-1)^{-1}$ define $L_q(u)$ to be a similar space of equivalence classes $\psi_u(t)$ satisfying the above conditions (A) and (B) with p replaced by q in condition (B).

The standard abuse of using equivalence classes and representative elements of equivalence classes interchangeably will be followed. This will cause no confusion if every statement that an element of $L_p(u)$ or $L_q(u)$ has a property P is interpreted to mean there is at least one member in the equivalence class that has the property P.

In the remainder of the paper, in certain instances to save repeticious writing of the symbol u, the dependence of spaces, functions, stopping times, trajectory process, and fields on the control u will not be indicated. The reader should keep in mind that each of these are defined with respect to some control u and depend on which control is being used.

Remark 1. *The spaces $L_p(u)$ and $L_q(u)$ are Banach spaces under their respective norms.*

Proof. To show they are complete spaces, notice that Cauchy convergence of a sequence ψ_n in either $L_p(u)$ or $L_q(u)$ is equivalent to Cauchy convergence of $\chi\psi_n$ in either $L_p(\Lambda \times P)$ or $L_q(\Lambda \times P)$. Since the latter are complete spaces, completeness of $L_p(u)$ or $L_q(u)$ will follow if the limit lies in these spaces. For a sequence

$\chi\psi_n$ converging in either $L_p(\Lambda \times P)$ or $L_q(\Lambda \times P)$ there is a subsequence $\chi\psi_{n_j}$ converging $\Lambda \times P$ almost everywhere. Define

$$\psi(t,\omega) = \begin{cases} \lim_{j\to\infty} \chi(t,\omega)\psi_{n_j}(t,\omega) & \text{if the limit exists,} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Then $\psi(t, \omega) = \chi(t, \omega)\psi(t, \omega)$, and either the $L_p(u)$ and $L_p(\Lambda \times P)$ or $L_q(u)$ and $L_q(\Lambda \times P)$ norms of $\psi(t, \omega)$ agree. From (5) it is seen that $\psi(t)$ is measurable and adapted to H_t since the $\psi_{n_j}(t)$ are.

Remark 2. *If $h(t)$ is a measurable real-valued stochastic process such that* $E\left\{\int_0^\eta |h(t)|^q\, dt\right\} < \infty$, *then* $E\{\chi(t)h(t)|H_t\} \in L_q(u)$.

Proof. It follows from [2, p. 439] that there is a sequence of essentially bounded stochastic step functions $h_n(t)$ converging in $L_q(\Lambda \times P)$ to $\chi(t)h(t)$. Since $H_{t_1} \subset H_{t_2}$ if $t_1 < t_2$, it follows by combining Theorem 4.3 of [2, p. 355] and Theorem 2.5 of [2, p. 60] that the remark is true for stochastic step functions. Since

$$|E\{\chi(t)h_n(t)|H_t\} - E\{\chi(t)h(t)|H_t\}|^q \leqq E\{\chi(t)|h_n(t) - h(t)|^q H_t\},$$

$E\{\chi(t)h_n(t)|H_t\}$ converges to $E\{\chi(t)h(t)|H_t\}$ in $L_q(u)$. Since $L_q(u)$ is complete this establishes the remark.

Remark 3. *The space $L_q(u)$ is the dual space of $L_p(u)$. Each linear functional on $L_p(u)$ has the form*

$$L[\phi] = E\int_0^\eta \phi(s)\psi(s)\, ds \tag{6}$$

for some element ψ of $L_q(u)$.

Proof. Since the mapping taking $\phi(s)$ into $\chi(s)\phi(s)$ maps $L_p(u)$ isomorphically as a Banach space into $L_p(\Lambda \times P)$, every linear functional on $L_p(u)$ has the form

$$E\int_0^\infty \chi(s)\psi(s)\phi(s)\, ds \tag{7}$$

for some element ψ of $L_q(\Lambda \times P)$. Now interchanging the order of integration, using the fact that $\chi(t)\phi(t)$ is adapted to H_t, and the formula for iterated expectations gives

$$L(\phi) = \int_0^\infty E\left\{\chi(s)\phi(s)E\{\chi(s)\psi(s)|H_s\}\right\} ds = E\int_0^\eta \phi(s)E\{\chi(s)\psi(s)|H_s\}\, ds. \tag{8}$$

By Remark 2, $E\{\chi(s)\psi(s)|H_s\}$ is an element of $L_q(u)$.

In the remainder of the paper the abbreviated notation (ϕ, ψ) will be used for $E\left\{\int_0^\eta \phi(s)\psi(s)\, ds\right\}$.

4. The operators R_h and T_h. Define operators R_h on $L_p\,(u)$ by

$$R_h[\phi](t) = \begin{cases} \phi(t-h) & \text{if } t \geqq h, \\ 0 & \text{if } t < h. \end{cases} \tag{9}$$

It is easily seen that R_h forms a norm decreasing semigroup of operators on $L_p(u)$.

THEOREM 1. *The operators R_h satisfy*

$$\lim_{h \downarrow 0} \|R_h[\phi] - \phi\| = 0. \tag{10}$$

Proof. This proof is an argument which is similar to [2, pp. 440–441]. Since the theorem is very important for later work this argument is carried out below.

It is not a loss of generality to assume ϕ is bounded. To see this define

$$\phi_n(t, \omega) = \begin{cases} \phi(t, \omega) & \text{if } |\phi(t, \omega)| \leqq n, \\ 0 & \text{if } |\phi(t, \omega)| > n. \end{cases} \tag{11}$$

Since ϕ is in $L_p(u)$, $\|\phi_n - \phi\|$ approaches zero. Since R_n is norm decreasing, $\|R_h\phi - \phi\| \leqq 2\|\phi_n - \phi\| + \|R_h\phi_n - \phi_n\|$ so the theorem will follow if it has been established for bounded elements of $L_p(u)$. For bounded $\phi(t, \omega)$, since $\|\eta\| < \infty$, by Lusin's theorem [13, p. 72] there is a continuous function $f_\varepsilon(t)$ so that for P—almost every fixed ω,

$$\left[\int_0^{\eta(\omega)} |\phi(s, \omega) - f_\varepsilon(s)|^p \, ds\right]^{1/p} < \varepsilon. \tag{12}$$

For such a fixed ω, applying Minkowski's inequality gives

$$\begin{aligned} &\limsup_{h \to 0} \left[\int_0^{\eta(\omega)} |R_h(\phi)(s, \omega) - \phi(s, \omega)|^p \, ds\right]^{1/p} \\ &\leqq \limsup_{h \to 0} \left\{ \left[\int_h^{\eta(\omega)} |\phi(s - h, \omega) - f_\varepsilon(s - h)|^p \, ds\right]^{1/p} \right. \\ &\quad + \left[\int_h^{\eta(\omega)} |f_\varepsilon(s - h) - f_\varepsilon(s)|^p \, ds\right]^{1/p} \\ &\quad \left. + \left[\int_h^{\eta(\omega)} |f_\varepsilon(s) - \phi(s, \omega)|^p \, ds\right]^{1/p} + \left[\int_0^h |\phi(s, \omega)|^p \, ds\right]^{1/p} \right\} \leqq 2\varepsilon. \end{aligned} \tag{13}$$

Since ε was arbitrary, the pth power of the quantity inside the lim sup on the left-hand side of (13) must approach zero as h approaches zero. Applying Lebesgue's dominated convergence theorem to the expected value of this quantity then implies (10).

THEOREM 2. *Let T_h be the adjoint operator of R_h. Then T_h has the representation*

$$T_h[\psi](t) = E\{\chi(t + h)\psi(t + h) | H_t\}. \tag{14}$$

Proof. Since T_h is the adjoint of R_h,

$$(\phi, T_h\psi) = E\left\{\int_0^\eta R_h[\phi](s)\psi(s) \, ds\right\} = E\left\{\int_h^\infty \chi(s)\phi(s - h)\psi(s) \, ds\right\}. \tag{15}$$

Now $\chi(t + h)\chi(t) = \chi(t + h)$; hence by interchanging the order of integration and changing the variable of integration, we have that (15) equals

$$\int_0^\infty E\{\chi(t)\phi(t)\chi(t + h)\psi(t + h)\} \, dt. \tag{16}$$

Since $\chi(t)$ and $\phi(t)$ are H_t-measurable, by the formula for iterated conditional

expectations the integrand in (16) equals

$$E\left\{\chi(t)\phi(t)E\{\chi(t+h)\psi(t+h)|H_t\}\right\}. \tag{17}$$

Using this and interchanging the order of integration once more gives

$$(\phi, T_h\psi) = E\int_0^\eta \phi(t)E\{\chi(t+h)\psi(t+h)|H_t\}\,dt, \tag{18}$$

which establishes the theorem.

Let X be a Banach space and R_h a norm decreasing semigroup of operators on X. Let Y be the dual space of X and T_h the adjoint semigroup of operators on Y. Recall [5, p. 37] that an element y in Y is in the domain D_A of the weak infinitesimal generator A of T_h if there is an element Ay in Y such that

$$\lim_{h\downarrow 0} h^{-1}[(x, T_hy) - (x, y)] = (x, Ay) \tag{19}$$

for all x in X. Recall [5, p. 40] that if y is in D_A, the formula

$$(x, T_hy) - (x, y) = \int_0^h (x, T_sAy)\,ds \tag{20}$$

holds.

THEOREM 3. *Let X be a Banach space, Y its dual, R_h a semigroup of norm decreasing operators on X, T_h the adjoint semigroup of operators on Y. For each x in X, let*

$$\lim_{h\downarrow 0} \|R_hx - x\| = 0. \tag{21}$$

Then a necessary and sufficient condition that an element y of Y be in D_A, the domain of the weak infinitesimal generator A of T_h, is that there exist a real number K such that

$$\|T_hy - y\| \leqq Kh. \tag{22}$$

Proof. Necessity. If $y \in D_A$ for each x in X,

$$\begin{aligned}(x, h^{-1}(T_hy - y)) &\leqq h^{-1}\int_0^h |(x, T_sAy)|\,ds \\ &= h^{-1}\int_0^h |(R_sx, Ay)|\,ds \leqq \|Ay\|\,\|x\|.\end{aligned} \tag{23}$$

Sufficiency.

$$|(x, T_{t+h}y - T_ty)| \leqq \|R_tx\|\,\|T_hy - y\| \leqq Kh\|x\|. \tag{24}$$

Hence (x, T_ty) is a Lipschitzian real-valued function of t for fixed x and y. Therefore, it has a derivative everywhere on R except at a set τ_x of Lebesgue measure zero. For each s in the complement of τ_x,

$$\lim_{h\downarrow 0}(R_sx, h^{-1}(T_hy - y)) = \lim_{h\downarrow 0}(x, h^{-1}(T_{s+h}y - T_sy)) \tag{25}$$

exists. Now (21) implies that the elements $R_s x$ for s in $R - \tau_x$ and x in X are dense in X. Thus (25), by the characterization of weak convergence, implies that $h^{-1}(T_h y - y)$ converges weakly to an element Ay of Y.

5. The value function. For a given control u let the notation $u^* \in u(t)$ mean that $u^*(s)$ and $u(s)$ agree on $0 \leqq s < t$. For $u^* \in u(t)$, if the control u^* is used after time t, consider the conditional expectation of the remaining performance given the field H_{ut} of measurements; that is,

$$E\left\{\int_{t \wedge \eta_{u^*}}^{\eta_{u^*}} k_{u^*}(s)\, ds | H_{ut}\right\}. \tag{26}$$

In subsequent formulas to shorten the notation the symbol $\int_t^\eta$ will be used to mean $\int_{t \wedge \eta}^\eta$. It will be always understood that if the upper limit of an integral is a stopping time the lower limit is the minimum of this stopping time and the time indicated.

Since performance rates are nonnegative, the random variables (26) are bounded below by zero. For fixed t define $V_u(t)$ to be the essential infimum of these random variables; that is,

$$V_u(t) = P\text{-ess}\inf_{u^* \in u(t)} E\left\{\int_t^{\eta_{u^*}} k_{u^*}(s)\, ds | H_{ut}\right\}. \tag{27}$$

Since $L_\infty(P)$ is a complete lattice [3, p. 302], $V_u(t)$ is a random variable for each fixed t, in fact, an H_{ut}-measurable random variable.

Remark 4. *$V_u(0)$ is independent of u and gives the infimum of the values of the performance indices* (4) *of the optimal control problem.*

Proof. The set $u^* \in u(0)$ includes the entire class of controls U because the set $0 \leqq s < 0$ is empty. Similarly from (3), H_{u0} is the field consisting of the empty space and the whole space Ω. Therefore, (27) implies that $V_u(0)$ is the infimum of the quantities (4).

The next objective will be to establish a stochastic analogue of Bellman's principle of optimality. The intuitive idea behind the principle of optimality is that it is always better to use immediately an optimal control than to use some other control for a short while and then use an optimal control.

The validity of inequality (28) with probability one for each t will be called the principle of optimality.

$$V_u(t) \leqq E\left\{\int_t^{t+h} \chi_u(s) k_u(s)\, ds | H_{ut}\right\} + E\{\chi(t+h) V_u(t+h) | H_{ut}\}. \tag{28}$$

Before establishing (28) under certain conditions, it will be shown that (28) does not always hold by giving a counterexample.

The following is a rather naive feedback control system for which (28) is false. Let Ω be the space of the two points $\Omega = \{0, 4\}$ with respective probabilities p and $1 - p$, where $0 < p < \frac{1}{2}$. Let

$$x_u(t, \omega) = \begin{cases} 4^{-1}\omega & \text{if } t \leqq 2 + u(t), \\ 4 & \text{if } 2 + u(t) < t \leqq 4. \end{cases} \tag{29}$$

Suppose the entire process $X_u(t, \omega)$ is observable but that A_t is empty except for $A_2 = \{1\}$. Let the class U of controls consist of two controls $u_1(t, \omega)$ and $u_2(t, \omega)$ given by

$$\begin{aligned} &u_1(t, \omega) = u_2(t, \omega) = 0 \quad \text{if } t \leqq 2, \\ &u_1(t, \omega) = x_{u_1}(1, \omega), \quad u_2(t, \omega) = 1 - x_{u_2}(1, \omega) \quad \text{if } 2 < t \leqq 4. \end{aligned} \tag{30}$$

Let the stopping time $\eta_u(\omega)$ equal the first time $x_u(t, \omega)$ equals four. Consider then the problem of minimizing $E\{\eta_u\}$.

It is easily seen that $\eta_{u_1}(0) = 2$, $\eta_{u_1}(4) = 3$, $\eta_{u_2}(0) = 3$, and $\eta_{u_2}(4) = 2$. Hence, $E\{V_{u_1}(2)\} = 0$ which implies that

$$2p + 3(1 - p) = V_{u_1}(0) \nleqq \int_0^2 dt + V_{u_1}(2) = 2. \tag{31}$$

In the example above it happened at a certain time that there was no control whose remaining performance approximated the value function at that time. To avoid this type of situation the following concept will be introduced.

The class of controls will be called relatively complete if for each control u, time t, and $\varepsilon > 0$, there is a control $u^* \in u(t)$ such that

$$V_u(t) \geqq E\left\{ \int_t^{\eta_{u^*}} k_{u^*}(s)\, ds | H_{ut} \right\} - \varepsilon \tag{32}$$

with probability one. Following [4, pp. 9–10], such a control will be called $(u(t), \varepsilon)$-optimal.

THEOREM 4. *If the class U of controls is relatively complete, the principle of optimality is valid.*

Proof. The controls $u' \in u(t + h)$ all belong to $u(t)$; so, with probability one,

$$V_u(t) \leqq E\left\{ \int_t^{t+h} \chi_u(s) k_u(s)\, ds | H_{ut} \right\} + P\text{-ess inf}_{u' \in u(t+h)} E\left\{ \int_{t+h}^{\eta_{u'}} k_{u'}(s)\, ds | H_{ut} \right\}. \tag{33}$$

Let $u^* \in u(t + h)$ be a $(u(t + h), \varepsilon)$-optimal control. Then

$$V_u(t + h) \geqq E\left\{ \int_{t+h}^{\eta_{u^*}} k_{u^*}(s)\, ds | H_{u(t+h)} \right\} - \varepsilon. \tag{34}$$

The definition (27) of $V_u(t)$ implies that $\chi_u(t + h)V_u(t + h) = V_u(t + h)$. Therefore, (34) implies

$$E\{\chi_u(t + h)V_u(t + h) | H_{ut}\} \geqq P\text{-ess inf}_{u' \in u(t+h)} E\left\{ \int_{t+h}^{\eta_{u'}} k_{u'}(s)\, ds | H_{ut} \right\} - \varepsilon. \tag{35}$$

Since ε is arbitrary, and a reverse inequality follows directly from the properties of conditional expectations and (27), it follows that

$$E\{\chi_u(t + h)V_u(t + h) | H_{ut}\} = P\text{-ess inf}_{u' \in u(t+h)} E\left\{ \int_{t+h}^{\eta_{u'}} k_{u'}(s)\, ds | H_{ut} \right\}. \tag{36}$$

Substituting (36) in (33) gives (28) which is the desired conclusion.

A control u will be called value decreasing if for each t,

$$V_u(t) \geqq E\{\chi_u(t + h)V_u(t + h) | H_{ut}\} \tag{37}$$

with probability one.

Not all controls are value decreasing. The reader will have little difficulty in constructing situations where the control "is in the wrong direction for certain times" in which the control is not value decreasing. Theorem 6 to be proved below and the nonnegativity of performance rates show that if the class of controls is relatively complete then any optimal control is value decreasing.

Notice that the right side of inequality (37) has the same formula as T_u operating on V_u would have, but we do not know that V_u is in $L_q(u)$. The notation

$$\|k_u\| = \sup_t E\{|k_u(t)|^q\}^{1/q} \tag{38}$$

will be used for the norm of a performance rate $k_u(t)$. Since $k_u(t)$ is dominated by some b_u, (38) is finite.

THEOREM 5. *If the class U of controls is relatively complete and if u is a value decreasing control, $V_u(t)$ belongs to $L_q(u)$.*

Proof of Theorem 5. The inequality

$$E|V_u(t)|^q \leqq E\left|E\left\{\int_t^{\eta_u} k_u(s)\,ds|H_{ut}\right\}\right|^q \leqq \int_0^{\|\eta_u\|} E\{k_u(s)^q\}\,ds \leqq \|\eta_u\|\,\|k_u\|^q \tag{39}$$

holds. Let

$$0 = t_0 < t_1 < t_2 < \cdots < t_n = \|\eta_u\| \tag{40}$$

be points of R and let

$$V_u^n(t) = V_u(t_{i+1}) \quad \text{if } t_i \leqq t < t_{i+1}. \tag{41}$$

Since $V_u^n(t)$ is a stochastic step function, it is jointly measurable on $R \times \Omega$. From (39), (41), the finiteness of $\|\eta_u\|$, and Remark 2, $E\{\chi_u(t)V_u^n(t)|H_{ut}\}$ is in $L_q(u)$. Since u is value decreasing,

$$\begin{aligned} 0 &\leqq V_u(t) - E\{\chi_u(t)V_u^n(t)|H_{ut}\} \\ &= V_u(t) - E\{\chi_u(t_{i+1})V_u(t_{i+1})|H_{ut}\} \quad \text{if } t_i \leqq t < t_{i+1}. \end{aligned} \tag{42}$$

By Theorem 4, the last term of (42) is smaller than

$$E\left\{\int_t^{t_{i+1}} \chi_u(s)k_u(s)\,ds|H_{ut}\right\} \quad \text{if } t_i \leqq t \leqq t_{i+1}. \tag{43}$$

The expected value of the qth power of expression (43) is bounded by $\|k_u\|^q(t_{i+1} - t)$. This can be made uniformly small by the appropriate choice of $\{t_0, t_1, \cdots, t_n\}$. Therefore, we conclude that choices can be made for $t_0, t_1, \cdots, t_n$ so that the corresponding functions $E\{\chi_u(t)V_u^n(t)|H_{ut}\}$ converge to $V_u(t)$ in the $L_q(u)$-norm. Since $L_q(u)$ is complete this completes the proof.

THEOREM 6. *If the class U of controls is relatively complete and u is an optimal control,*

$$V_u(t) = E\left\{\int_t^{\eta_u} k_u(s)\,ds|H_{ut}\right\} \tag{44}$$

with probability one for each t.

Proof. From Remark 4, for an optimal control u,

$$V_u(0) = E\left\{\int_0^{\eta_u} k_u(s)\,ds\right\} = E\left\{\int_0^t \chi_u(s)k_u(s)\,ds\right\} + E\left\{E\left\{\int_t^{\eta_u} k_u(s)\,ds|H_{ut}\right\}\right\}. \tag{45}$$

Subtracting (45) from (28) evaluated at $t = 0$, $h = t$, and using that $\chi_u(t)V_u(t) = V_u(t)$ gives by Theorem 4 that

$$0 \leqq E\left\{V_u(t) - E\left\{\int_t^{\eta_u} k_u(s)\,ds|H_{ut}\right\}\right\} \tag{46}$$

for each t. By definition (27) the quantity within the first expected value sign of (46) is nonpositive with probability one. Therefore, (46) implies it must equal zero with probability one, which establishes the theorem.

THEOREM 7. *If* $\psi(t) = E\left\{\int_t^{\eta} k(s)\,ds|H_t\right\}$, *then* $\psi \in D_A$ *and* $-A[\psi](t) = E\{\chi(t)k(t)|H_t\}$.

Proof. Let $\theta(t)$ denote $E\{\chi(t)k(t)|H_t\}$. By the formula for iterated expectations and interchanging orders of integration,

$$\begin{aligned} &E\left\{\int_t^{\eta} k(s)\,ds|H_t\right\} - E\left\{\chi(t+h)E\left\{\int_{t+h}^{\eta} k(s)\,ds|H_{t+h}\right\}|H_t\right\} \\ &\qquad = E\left\{\int_t^{t+h} \chi(s)k(s)\,ds|H_t\right\} = \int_t^{t+h} E\{\chi(s)k(s)|H_t\}\,ds. \end{aligned} \tag{47}$$

Hence, Theorem 2 and (47) imply for any $\phi \in L_1(u)$ that

$$\begin{aligned} (\phi, \psi) - (\phi, T_h\psi) &= E\left\{\int_0^{\infty} \chi(s)\phi(s)\int_0^h E\{\chi(s+v)k(s+v)|H_s\}\,dv\,ds\right\} \\ &= \int_0^h \int_v^{\infty} E\{\phi(t-v)\chi(t)k(t)\}\,dt\,dv \\ &= \int_0^h E\left\{\int_0^{\eta} R_v[\phi](t)E\{\chi(t)k(t)|H_t\}\,dt\right\} dv \\ &= \int_0^h (R_v\phi, \theta)\,dv. \end{aligned} \tag{48}$$

Since $(R_v\phi, \theta)$ is continuous at $v = 0$, the difference quotient defining $A\psi$ converges to $-(R_0\phi, \theta) = -(\phi, \theta)$. Since this is true for each ϕ in $L_p(u)$ the theorem follows.

Remark 5. *If* ψ_1 *and* ψ_2 *are in* $L_q(u)$ *and*

$$(\phi, \psi_1) \leqq (\phi, \psi_2) \tag{49}$$

for every nonnegative ϕ *of* $L_p(u)$, *then*

$$\chi(t, \omega)\psi_1(t, \omega) \leqq \chi(t, \omega)\psi_2(t, \omega), \qquad \Lambda \times P \text{ almost everywhere.} \tag{50}$$

Proof. Let λ be the characteristic function of the set on which $\psi_2 - \psi_1 < 0$. Then since $E\left\{\int_0^{\eta} |\lambda(s)|^p\,ds\right\} < \|\eta\|$ and $\lambda(t)$ is a measurable stochastic process

adapted to H_t, λ is in $L_p(u)$. Substituting λ in (49) gives

$$E\int_0^\infty \chi(t)\lambda(t)(\psi_2(t) - \psi_1(t))\,dt \geqq 0, \tag{51}$$

which can happen only if $\chi(t)\lambda(t)$ vanishes $\Lambda \times P$ almost everywhere since the integrand is negative. The definition of λ implies (50).

THEOREM 8 (Necessary conditions for optimality). *If the class U of controls is relatively complete, for any value decreasing control u, the following conditions are satisfied*:

There is a constant $V(0)$ independent of u such that

$$\lim_{t\downarrow 0} E\{V_u(t)\} = V(0), \tag{52}$$

$$V_u(t) \in L_q(u) \cap D_{A_u} \tag{53}$$

and

$$-A_u[V_u](t) \leqq E\{\chi_u(t)k_u(t)|H_{ut}\} \tag{54}$$

for $\Lambda \times P$ almost every (t, ω). If u' is an optimal control, (54) *holds for u' with equality replacing the inequality.*

Proof. Since $\chi_u(t + h)V_u(t + h) = V_u(t + h)$, taking expectations of (28) for $t = 0$ implies

$$0 \leqq E\{V_u(0)\} - E\{V_u(h)\} \leqq E\left\{\int_0^h \chi_u(s)k_u(s)\,ds\right\} \leqq (1 + \|k_u\|^q)h. \tag{55}$$

Remark 4 implies $V_u(0)$ is a constant independent of u. Set $V(0)$ equal to this constant. Inequality (55) then implies (52).

From (28), for a value decreasing control u, since $k_u(t)$ is dominated by some b_u,

$$\|V_u - T_hV_u\| \leqq E\left\{\int_0^{\eta_u} E\left\{\int_t^{t+h} \chi_u(s)k_u(s)\,ds|H_{ut}\right\}^q dt\right\}^{1/q} \leqq \|\eta_u\|^{1/q}E\{b_u^q\}^{1/q}h. \tag{56}$$

Theorem 5 asserts that $V_u \in L_q(u)$; equation (56) and Theorem 3 imply $V_u \in D_{A_u}$. If

$$\psi_u(t) = E\left\{\int_t^{\eta_u} \chi_u(s)k_u(s)\,ds|H_{ut}\right\}, \tag{57}$$

Theorem 4 implies for any nonnegative $\phi_u \in L_p(u)$ that

$$0 \leqq (\phi_u, V_u - T_hV_u) \leqq (\phi_u, \psi_u - T_h\psi_u). \tag{58}$$

Therefore, by dividing by h and passing to the limit using Theorem 7,

$$0 \leqq (\phi_u, -A_uV_u) \leqq (\phi_u, E\{\chi_u(t)k_u(t)|H_{ut}\}). \tag{59}$$

Inequality (54) now follows from (59) and Remark 5.

Theorems 6 and 7 show (54) holds with equality for an optimal control u'.

THEOREM 9 (Sufficient conditions for optimality). *A sufficient condition for a control u' to be optimal is that there exist a constant $W(0)$ and for each u in U a stochastic process $W_u(t)$ such that*

$$\lim_{t\downarrow 0} E\{W_u(t)\} = W(0), \tag{60}$$

(61) $W_u(t) \in L_q(u) \cap D_{A_u}$,

(62) $-A_u W_u \leqq E\{\chi_u(t) k_u(t) | H_{ut}\}$, $\Lambda \times P$ *almost everywhere,*

(63) $-A_{u'} W_{u'} = E\{\chi_{u'}(t) k_{u'}(t) | H_{u't}\}$, $\Lambda \times P$ *almost everywhere.*

Proof. Again let $\psi_u(t) = E\left\{ \int_t^{\eta_u} k_u(s)\, ds | H_{ut} \right\}$. From Theorem 7, (61) and (62), $\psi_u(t) - W_u(t) \in D_{A_u}$ and

(64) $A_u(\psi_u(t) - W_u(t)) \leqq 0$, $\Lambda \times P$ almost everywhere.

Then for any nonnegative element ϕ_u of $L_p(u)$,

(65) $(\phi_u, \psi_u - W_u) - (\phi_u, T_h(\psi_u - W_u)) \geqq -\int_0^h (R_s \phi_u, A_u(\psi_u - W_u))\, ds \geqq 0.$

Note that for any elements ϕ of $L_p(u)$ and ψ of $L_q(u)$ that

(66) $(\phi, T_h \psi) = (R_h \phi, \psi) = E\left\{ \int_h^{\eta} \phi(t - h)\psi(t)\, dt \right\}.$

Recalling the meaning of $\int_h^{\eta}$ and that $\|\eta\| < \infty$, it follows that (66) is zero for large h. This and (65) imply

(67) $(\phi_u, \psi_u) \geqq (\phi_u, W_u)$

for every nonnegative element of $L_p(u)$. By Remark 5,

(68) $\psi_u \geqq W_u$, $\Lambda \times P$ almost everywhere.

A similar argument using (63) implies that

(69) $\psi_{u'} = W_{u'}$, $\Lambda \times P$ almost everywhere.

Now

(70) $0 \leqq E\{\psi_u(0)\} - E\{\psi_u(h)\} = E\left\{ \int_0^h \chi_u(s) k_u(s)\, ds \right\} \leqq (1 + \|k_u\|)h,$

so

(71) $\lim_{h \downarrow 0} E\{\psi_u(h)\} = E\{\psi_u(0)\}.$

From (68), (69), (60) and (71) it follows that

(72) $E\{\psi_u(0)\} \geqq W(0) = E\{\psi_{u'}(0)\}.$

Rewriting (72) as

(73) $E\left\{ \int_0^{\eta_u} k_u(s)\, ds \right\} \geqq E\left\{ \int_0^{\eta_{u'}} k_{u'}(s)\, ds \right\}$

gives the desired statement of optimality of u'.

REFERENCES

[1] V. G. BOLTYANSKI, *Sufficient conditions for optimality and the justification of the dynamic programming method*, Izv. Akad. Nauk SSSR Ser. Mat., 28 (1964), pp. 481–514; English transl., this Journal, 44 (1966), pp. 326–361.

[2] J. L. DOOB, *Stochastic Process*, John Wiley, New York, 1953.

[3] N. DUNFORD AND J. T. SCHWARTZ, Linear Operators, Interscience, New York, 1958.

[4] E. B. DYNKIN, *Controlled stochastic processes—discrete parameter*, Theor. Probability Appl., 10 (1966), pp.1–14.

[5] ———, *Markov Processes. I*, Springer–Verlag, Berlin, 1965.

[6] W. H. FLEMING, *Some Markovian optimization problems*, J. Math. Mech., 12 (1963), pp. 131–140.

[7] ———, *Duality and a'priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., 16 (1966), pp. 131–140.

[8] ———, *Optimal control of partially observable diffusions*, this Journal, 6 (1968), pp. 194–214.

[9] ———, *Optimal continuous parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–509.

[10] N. N. KRASOVSKII AND E. A. LIDSKII, *Analytic design of controls in systems with random characteristics, Parts I–III*, Avtomat. i Telemekh., 22 (1961), pp. 1145–1150, 1273–1278, 1425–1431.

[11] H. J. KUSHNER, *Sufficient conditions for the optimality of a stochastic control*, this Journal, 3 (1965), pp.499–508.

[12] M. LOEVE, *Probability Theory*, 3rd ed., Van Nostrand, New York, 1963.

[13] S. SAKS, *Theory of the Integral*, Hafner, New York, 1937.

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

COMMENTS ON THE PAPER "OPTIMAL CONTROL OF PROCESSES DESCRIBED BY INTEGRAL EQUATIONS. I" BY V. R. VINOKUROV*

L. W. NEUSTADT† AND J. WARGA‡

A number of statements and arguments in [1] are inaccurate. Theorem 2.1 in [1, p. 327] is not valid. This theorem states that, subject to conditions specified in [1, § 1, pp. 324–325], a couple (x, u) is an optimal solution if and only if it satisfies the maximum principle [1, pp. 326–327]. It is well known, however, that, in the special case where equation (1.1) in [1, p. 324] is equivalent to an ordinary differential equation, a couple (x, u) may satisfy the maximum condition without being optimal. One such counterexample is due to Bolza and is described by Goursat [2, p. 599]. A simple case of this general counterexample is provided by the problem of minimizing $x_1(1)$ subject to the relations

$$(x_1(t), x_2(t)) = x(t) = \int_0^t K(t, x(s), u(s), s)\, ds, \qquad t \in [0, 1],$$

$$x_2(1) = 0,$$

where

$$K = (K_1, K_2), \quad K_1(t, x, u, s) = tu^4 - 2u^3x_2, \quad K_2(t, x, u, s) = u,$$

and the set U is the real line (or, alternately, U may be a closed interval $[-N, N]$ for an arbitrary positive N). It can then be verified that the couple $(x, u) = (0, 0)$ satisfies the maximum condition but that it is not optimal; the control $u(t) = \alpha$ for $t \in [0, h]$, $u(t) = \alpha \cdot (1 - t)h/(1 - h)$ for $t \in (h, 1]$ (with $0 \leqq \alpha \leqq N$ and $0 < h < \frac{1}{3}$) yielding $x_1(1) = -\frac{1}{2}\alpha^4h^2(1 - 3h)/(1 - h)^3 < 0$, $x_2(1) = 0$.

The proof of the sufficiency condition [1, p. 335] is in error because sufficiency follows from the relation $I_0(x_0, u_0) \leqq I_0(x, u)$ and not from $I_0(x, u_0) < I_0(x, u)$.

The proof of "necessity" [1, p. 327] is purely formal and no justification is provided for introducing the Lagrange coefficients μ_j[1, § 2, p. 3251] or assuming that the Jacobian determinant (2.13) in [1, p. 328] does not vanish.

REFERENCES

[1] V. R. VINOKUROV, *Optimal control of processes described by integral equations. I*, this Journal, 7 (1969), pp. 324–336.

[2] E. GOURSAT, *Cours d'analyse mathematique*, Time III, 5th ed., Gauthier-Villars, Paris, 1942.

* Received May 15, 1970.

† Department of Electrical Engineering, University of Southern California, Los Angeles, California 90007.

‡ Mathematics Department, Northeastern University, Boston, Massachusetts 02115.

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

CONTROLABILITE DES SYSTEMES NON LINEAIRES*

CLAUDE LOBRY†

Introduction. Dans cet article on étudie le problème de l'accessibilité pour des systèmes non linéaires du type:

$$\frac{dx}{dt} = f(x, t, u), \qquad x \in R^n, \qquad u \in \Omega \subset R^p, \tag{1}$$

et plus particulièrement dans le cas où Ω n'est pas un ensemble convexe.

Cette étude repose pour l'essentiel sur un théorème dû à Chow [2]. R. Hermann a, le premier, montré dans [7] et [8] comment ce théorème pouvait être appliqué avec fruit en théorie du contrôle. Depuis H. Hermes [10]–[12], utilisant le théorème de Chow également, a abordé ce problème en termes de systèmes de Pfaff. Parallèlement dans [14] et [15] Kučera fait une étude très fine concernant les propriétés géométriques de l'ensemble des états accessibles, pour un système linéaire particulier. Toutes ces études reposent fondamentalement sur l'utilisation de dérivées de Lie de champs de vecteurs. Nous ne proposons pas dans cet article des résultats nouveaux importants mais plutôt une approche géométrique systématique du problème. Pour cela nous avons partagé l'exposé en deux parties. Dans la première nous exposons en termes purement mathématiques des résultats dus essentiellement à Hermann et Chow; dans la seconde, nous interprétons ces résultats en termes de contrôlabilité. Lorsqu'un résultat est proche d'un résultat classique des références sont données, cependant la bibliographie proposée est loin d'être exhaustive, en particulier tous les travaux concernant les équations du type:

$$\frac{dx}{dt} \in \Gamma(x, t), \tag{2}$$

tels que ceux de Wajeski, Filippov, Castaing, etc. ont été délibérément omis. En fait, les méthodes et les résultats proposés ici sont de nature très différente.

Le paragraphe 1.1 est uniquement consacré à l'introduction de définitions classiques en géométrie différentielle. Il ne contient aucun résultats.

Le paragraphe 1.2 est consacré à l'étude des "variétés intégrales" d'une famille de champs de vecteurs. C'est en un certain sens une généralisation de l'étude de R. Hermann [7]. La proposition 1.2.1 est le résultat central de ce paragraphe. Les exemples qui l'accompagnent montrent que c'est le résultat le plus précis que l'on puisse obtenir dans le contexte choisi. Au point de vue géométrique il serait plus logique de s'intéresser à l'intégration des "distributions cohérentes" [22], [23], i.e., se donner, de manière suffisamment régulière, en chaque point de la variété un sous-espace de l'espace tangent en ce point. Beaucoup des résultats énoncés ici peuvent se traduire immédiatement sauf, précisément, la proposition 1.2.1. Le point de vue adopté (famille de champs de vecteurs) permet l'utilisation du language géométrique et s'interprète immédiatement en termes de contrôle.

* Received by the editors November 18, 1969, and in final revised form February 24, 1970.

† Mathématiques Appliquées, Université de Grenoble, Cedex 53, 38 Grenoble-Gare, France.

Le paragraphe 1.3 propose une démonstration du théorème de Chow dont l'interpretation en termes de la théorie du contrôle est immédiate.

Le paragraphe 2.1 établit les liens entre le formalisme précédemment développé et le formalisme classique de la théorie du contrôle. Les propositions qui y sont énoncées sont des conséquences immédiates des définitions.

Le paragraphe 2.2 est consacré à l'étude locale de l'ensemble des états accessibles d'un système "contrôlé." Il s'agit de corollaires des résultats de la première partie. Ces résultats ne sont pas classiques, et il est possible qu'une étude plus précise menée dans la même direction apporte d'autres renseignements.

Le paragraphe 2.3 est consacré à l'étude des systèmes du type:

$$\frac{dx}{dt} = H(x)\cdot u, \qquad x \in R^n, \qquad u \in R^p, \qquad H(x) \in \mathscr{L}(R^p, R^n).$$

Ces systèmes ont été introduits dans [10]. Une conjecture raisonnable concernant la "bang-bang" contrôlabilité est proposée. Cette conjecture fait apparaître la possibilité de décrire l'ensemble des états accessibles de certains systèmes comme l'ensemble:

$$\{x \in R^n ; g_i(x) \leqq 0, i = 1, 2, \cdots, p\},$$

où les applications g_i sont des applications différentiables de R^n dans R.

Les problèmes variationnels issus de problèmes de contrôle, (contrôle optimal) n'ont pas étés abordés. Il est clair que l'étude locale du paragraphe 2.2 peut être utile dans l'étude de problèmes d'optimisation.

1. Intégrabilité des familles de champs de vecteurs.

1.1. Notations. Nous introduisons les notations utilisées dans la suite. Pour la définition du vocabulaire de géométrie différentielle utilisée on pourra se reporter aux ouvrages classiques suivants: [1], [9], [18], [19].

Nous supposons systématiquement que les variétés, champs de vecteurs, fonctions que nous utilisons sont de classe C^∞. Cette hypothèse ne sera plus mentionnée par la suite. On supposera de plus que toutes les variétés sont séparées.

Soit M une variété, on note:

TM_x l'espace tangent à M en x;

$C^\infty(M)$ l'anneau des fonctions (C^∞) définies sur M;

$V(M)$ le $C(M)$-module des champs de vecteurs (C^∞) sur M.

Soient M et N deux variétés, $\phi : M \to N$ une application de M dans N. On note:

ϕ^* la différentielle de ϕ;

$\phi^*_{(x)}$ la valeur de ϕ^* au point x. $\phi^*_{(x)}$ est alors une application

$$\phi^*(x) : TM_x \to TN_{\phi(x)}$$

dont on note

$$\phi^*(x)\cdot h,$$

la valeur en un point h de TM_x.

Soit X un champ de vecteur défini sur M. On note :

$X_t(\cdot)$ le groupe local à un paramètre engendré par X. On sait qu'en général $X_t(\cdot)$ n'est défini que pour des valeurs de t suffisamment petites. Pour simplifier les notations nous omettrons systématiquement le "pour t assez petit." Il est facile de voir qu'aucune difficulté supplémentaire n'est liée à cette question dans ce qui suit. Sous les réserves exprimés ci-dessus on peut dire alors que $X_t(\cdot)$ est une application de $R \times M$ dans M :

$$(x, t) \to X_t(x).$$

On a de plus les relations :

$$X_0(x) = x,$$

$$X_{t+t'}(x) = X_t(X_{t'}(x)).$$

Pour t fixé on note : X_t^* la différentielle de l'application

$$x \to X_t(x).$$

Dans ces conditions $X_t^*(x)$ est une application linéaire inversible de TM_x dans $TM_{X_t(x)}$ satisfaisant aux relations :

$$X_0^*(x) = \text{identité},$$

$$(X_t^*(x))^{-1} = X_{-t}^*(X_t(x)).$$

Soient X et Y deux champs de vecteurs définis sur M, on note :

$[XY]$ le crochet de Jacobi des champs X et Y.

On sait que si on note $V_x(t)$ le vecteur de TM_x défini par

$$V_x(t) = (Y_t^*(x)^{-1}(X(Y_t(x))),$$

on a par définition de $[XY]$,

$$\left(\frac{d}{dt} V_x(t)\right)_{t=t_0} = (Y_{t_0}^*(x))^{-1} \cdot [XY](Y_{t_0}(x)).$$

On pourra trouver dans [24] une interprètation géométrique de cette notion. Pour ce qui nous intéresse la meilleure interprètation que l'on puisse donner est le théorème de Chow lui méme tel qu'il est démontré en § 1.3.

Si d'autre part $(x_1, x_2, \cdots, x_n)$ est un système de coordonnées locales de M, $(\partial x_1, \cdots, \partial x_n)$ la base de TM_x associée, si on note

$$X(x) = \sum_{i=1}^{n} X_i(x_1, \cdots, x_n) \frac{\partial}{\partial x_i},$$

on a

$$[XY](x_1 \cdots x_n) = X^*(x_1 \cdots x_n) \cdot Y(x_1 \cdots x_n) - Y^*(x_1 \cdots x_n) X(x_1 \cdots x_n),$$

où $X^*(x_1 \cdots x_n)$ est la matrice :

$$X^*(x_1 \cdots x_n) = \left(\frac{\partial X_i}{\partial x_j}(x_1 \cdots x_n)\right) \qquad i = 1, \cdots, n, \quad j = 1, \cdots, n.$$

Rappelons pour terminer une conséquence classique du théorème des fonctions implicites. Soit $\phi : M \to N$ une application de la variété M dans la variété N. On dit que ϕ est une immersion si quel que soit x dans M la valeur en x de la différentielle

$$\phi^*(x) : TM_x \to TN_{\phi(x)}$$

est une application linéaire injective.

PROPOSITION 1.1.1. *Soit ϕ une immersion de M dans N. Quel que soit x dans M il existe un voisinage $\mathcal{U}$ de x et un voisinage $\mathcal{V}$ de $\phi(x)$ tels que :*

(i) *ϕ restreinte à $\mathcal{U}$ est injective ;*

(ii) *il existe un système de coordonnées locales sur $\mathcal{V}$, $(y_1 \cdots y_n)$ tel que $\phi(\mathcal{U})$ soit défini par*

$$\phi(\mathcal{U}) = \{(y_1 \cdots y_n) : y_1 = y_2 = \cdots = y_p = 0\},$$

où p est égal à $m - n$, m et n désignant respectivement la dimension de M et de N.

1.2. Intégrabilité des familles de champs de vecteurs. Soit M une variété. Introduisons la définition suivante.

DÉFINITION 1.2.1. Soit D une famille de champs de vecteurs définis sur M. On dit qu'une *sous-variété N de M est une sous-variété intégrale de D* si N est connexe et si pour tout x de N on a l'égalité

$$TN_x = \mathcal{L}(D(x)),$$

où $\mathcal{L}(D(x))$ est l'espace vectoriel engendré par l'ensemble

$$D(x) = \{X(x) ; X \in D\}.$$

Le résultat essentiel que nous démontrons dans ce paragraphe est la proposition suivante.

PROPOSITION 1.2.1. *Soit D une famille de champs de vecteurs analytiques définis sur la variété analytique M, stable pour l'opération de crochet, c'est à dire telle que*

$$(X \in D,\ Y \in D) \Rightarrow [XY] \in D.$$

Par tout point x de M il passe une unique sous-variété intégrale de D, maximale pour l'inclusion.

Nous nous proposerons de plus une description précise de la structure de variété de la sous-variété de D qui sera interprétée par la suite (§ 2.1) en termes de

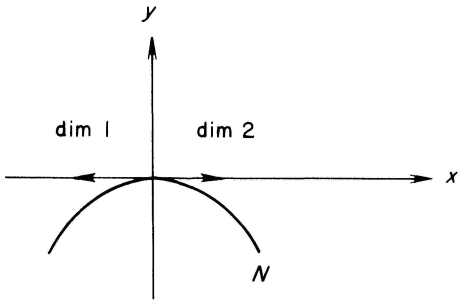


FIG. 1.

contrôle. Les idées contenues dans ce paragraphe sont très directement inspirées de celles de R. Hermann [8] et plus précisément le point essentiel, le lemme 1.2.1, correspond au lemme 2.1 de Hermann.

La difficulté de ce théorème est due à ce que la dimension de l'espace vectoriel $D(x)$ n'est pas supposée être constante. Lorsqu'elle est constante le classique théorème de Frobenius (cf. prop. 1.2.2) s'applique. L'exemple qui suit montre ce qui peut se produire quand la dimension n'est pas constante.

Exemple 1. Considèrons dans R^2 les deux champs suivants:

$$X(x, y) = \begin{pmatrix} 1 \\ 0 \end{pmatrix};$$

$$Y(x, y) = \begin{cases} \begin{pmatrix} 1 \\ \exp\left(-\dfrac{1}{x^2}\right) \end{pmatrix} & \text{si} \quad x > 0, \\ \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \text{si} \quad x \leqq 0. \end{cases}$$

Ces deux champs sont de classe C^∞. On vérifie immédiatement que la famille stable par crochet engendrée par les champs $\pm X$ et $\pm Y$ est la famille D définie par

$$D = \{\pm X;\ \pm Y;\ \pm Y^{(n)},\ n \in N\},$$

où le champ $Y^{(n)}$ est le champ

$$Y^{(n)}_{(x,y)} = \begin{cases} \begin{pmatrix} 1 \\ \left(\exp\left(-\dfrac{1}{x^2}\right)\right)^{(n)} \end{pmatrix} & \text{si} \quad x > 0, \\ \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \text{si} \quad x \leqq 0, \end{cases}$$

où $(\exp(-1/x^2))^{(n)}$ désigne la dérivée n-ième de $\exp(-1/x^2)$. On a alors

$$\dim \mathscr{L}(D_{(xy)}) = \begin{cases} 2 & \text{si} \quad x > 0, \\ 1 & \text{si} \quad x \leqq 0 \end{cases} \qquad \text{(cf. Fig. 1.)}$$

Supposons que par $(0, 0)$ il passe une sous-variété intégrale de la famille D. L'espace tangent en $(0, 0)$ à cette sous-variété est donc la droite $\{(x, y); y = 0\}$. Il s'ensuit que cette sous-variéte "pénètre" nécessairement dans le demi-espace $\{x, y; x \geqq 0\}$, où elle devrait avoir une dimension égale à 2 ce qui est évidemment impossible.

Notons "F.C.V." une famille de champs de vecteurs sur une variété M.

DÉFINITION 1.2.2. Soit D une F.C.V. On dit qu'un arc continu:

$$\alpha : [ab] \to M$$

est un *chemin intégral* de D si α est indéfiniment différentiable par morceaux et si pour tout intervalle I inclus dans $[ab]$ sur lequel α est différentiable il existe un

champ X de D tel que

$$\frac{d\alpha}{dt}(t) = X(\alpha(t)), \qquad t \in I.$$

Par indéfiniment différentiable par morceaux il faut entendre plus précisément que $[ab]$ est union finie d'intervalles $I_1 \cdots I_q$ tels que α restreint à I_j; soit la restriction à I_j d'une application indéfiniment différentiable de R dans M. On obtient donc un chemin intégral de D en "recollant continuement" un nombre fini de courbes intégrales de champs X de D.

DÉFINITION 1.2.3. Soit D une F.C.V. sur M. On appelle *feuille intégrale* de D passant par X et on note: L_x l'ensemble des points de M qui peuvent être joints à x par un chemin intégral de D.

On verra qu'à quelques réserves près la feuille intégrale L_x est précisément la sous-variété intégrale de D passant par x. On déduit immédiatement du classique théorème de Frobenius sur la complète intégrabilité des systèmes de Pfaff le résultat suivant concernant L_x.

PROPOSITION 1.2.2. *Soit D une F.C.V. définie sur M telle que*:

(i) *D est un sous-module de $V(M)$;*

(ii) $(X \in D,\ Y \in D) \Rightarrow [XY] \in D$;

(iii) *la dimension de l'espace vectoriel*

$$D(x) = \{X(x);\ X \in D\}$$

est indépendante de x et égale à p.

Alors l'ensemble L_x peut être muni d'une structure de variété différentiable S telle que la sous-variété (L_x, S) soit l'unique sous-variété intégrale maximale de D passant par x.

Démonstration. On sait d'après le théorème de Frobenius qu'il existe une unique sous-variété intégrale maximale de D passant par X. Il suffit donc de constater, ce qui est très clair d'après les définitions, que L_x coïncide avec cette sous-variété.

Dans [8] Hermann propose un théorème dans lequel l'hypothèse selon laquelle la dimension de $D(x)$ est constante est supprimée au profit d'autres conditions de régularité. C'est ce théorème que nous allons montrer après les lemmes suivants.

DÉFINITION 1.2.4. On dit qu'une famille D de champs de vecteurs définis sur M est *localement de type fini* si quelque soit x dans M il existe un nombre fini de champs de D:

$$X^1 \cdots X^q$$

tels que pour tout X de D il existe un voisinage $\mathscr{V}_{x,X}$ de x sur lequel on ait

$$[XX^j](y) = \sum_{i=1}^{q} f_i^j(y) X^i(y), \qquad y \in \mathscr{V}_{x,X}.$$

Cette condition est légèrement plus faible que la condition "locally finitely generated" de Hermann. On verra qu'elle est réalisée pour des familles de champs de vecteurs *analytiques*. Le lemme 2.1 de Hermann reste vrai sous cette hypothèse et on a alors le lemme suivant.

LEMME 1.2.1. (Hermann). *Soit D une F.C.V. sur M localement de type fini. Quels que soient x dans M, X dans D et t dans R tels que $X_t(x)$ soit défini, l'application*

$$X_t^*(x): TM_x \to TM_{X_t(x)}$$

définit un isomorphisme entre les espaces vectoriels

$$\mathscr{L}(D(x)) \quad et \quad \mathscr{L}(D(X_t(x))).$$

Démonstration. Il suffit de prouver que si le vecteur V appartient à l'espace $\mathscr{L}(D(x))$, alors le vecteur $X_t^*(x) \cdot V$ appartient à l'espace $\mathscr{L}(D(X_t(x)))$.

D'autre part l'arc:

$$\theta \to X_\theta(x), \qquad \theta \in [0, t],$$

est compact, il suffit donc de montrer la proposition plus faible suivante:

"*Pour tout x dans M il existe un réel $\varepsilon(x)$ strictement positif tel que pour tout t en valeur absolue inférieur à $\varepsilon(x)$ on ait*

$$V \in \mathscr{L}(D(x)) \Rightarrow X_t^*(x) \cdot V \in \mathscr{L}(D(X_t(x))).$$"

La famille D est localement de type fini, il existe donc q champs de vecteurs:

$$X^1 \cdots X^q$$

et un voisinage de x sur lequel

$$[XX^j](y) = \sum_{i=1}^{q} f_i^j(y) X^i(y), \qquad j = 1, \cdots, q.$$

Notons $V^j(t)$ le vecteur de TM_x défini par

$$V^j(t) = (X_t^*(x))^{-1} \cdot (X^j(X_t(x)).$$

Par définition du crochet de deux champs on a

$$\frac{d}{dt} V^j(t) = (X_t^*(x))^{-1} \cdot [XX^j](X_t(x)).$$

Soit $\varepsilon(x)$ un réel tel que

$$|t| \leqq \varepsilon(x) \Rightarrow X_t(x) \in \mathscr{V}_{x,X}.$$

On a alors pour $|t| \leqq \varepsilon(x)$,

$$\frac{d}{dt} V^j(t) = \sum_{i=1}^{q} f_i^j(X_t(x))(X_t^*(x))^{-1} \cdot X^i(X_t(x)), \qquad j = 1, \cdots, q.$$

Soit encore

$$\frac{d}{dt} V^j(t) = \sum_{i=1}^{q} f_i^j(X_t(x)) V^i(t), \qquad j = 1, \cdots, q.$$

Les q vecteurs $V^j(t)$ sont solution d'un système de q équations différentielles linéaires. Il existe donc des fonctions

$$\alpha_i^j(t), \qquad i = 1, \cdots, q, \quad j = 1, \cdots, q,$$

telles que

$$V^j(0) = \sum_{i=1}^{q} \alpha_i^j(t) V^i(t).$$

Soit encore

$$X_t^*(x) \cdot X^j(x) = \sum_{i=1}^{q} \alpha_i^j(t) X^i(X_t(x)).$$

Le vecteur $X_t^*(x) \cdot X^j(x)$ appartient donc à $\mathscr{L}(D(X_t(x)))$. Les vecteurs $X^j(x)$ constituant un système de générateurs de $\mathscr{L}(D(x))$; le lemme est démontré.

DÉFINITION 1.2.5. Soit D une F.C.V. sur M. On dit que D est *symétrique* si quel que soit X dans D le champ—X appartient à D.

LEMME 1.2.2. *Soit D une F.C.V. sur M, symétrique, localement de type fini. Soient $X^1, X^2, \cdots, X^p$ p champs définissant une base de $\mathscr{L}(D(x))$. Soit ϕ_x l'application définie par*

$$(t_1, t_2, \cdots, t_p) \in R^p \to \phi_x(t_1, \cdots, t_p) = X_{t_p}^p \circ X_{t_{p-1}}^{p-1} \circ \cdots \circ X_{t_1}^1(x) \in M.$$

Il existe un voisinage $\mathscr{U}_x$ de x tel que ϕ_x restreinte à $\mathscr{U}_x$ ait les propriétés suivantes:

(i) $\phi_x(\mathscr{U}_x) \subset L_x$;

(ii) *ϕ_x est une immersion injective;*

(iii) $\operatorname{Im}(\phi_x^*(t_1, \cdots, t_p)) = \mathscr{L}(D(\phi_x(t_1, \cdots, t_p)))$, *où* $\operatorname{Im}(\phi_x^*(t_1, \cdots, t_p))$ *désigne l'image de l'application*

$$\phi_x^*(t_1, \cdots, t_p)) : R^p \to TM_{\phi_x(t_1, \cdots, t_p)}.$$

En d'autres termes, ϕ_x définit une sous-variété intégrale de D passant par x.

Démonstration. La définition d'un chemin intégral impose que l'on parcoure les courbes intégrales des champs de D dans le sens des t croissants. Comme la famille D est symétrique on aura

$$X_{-t}(x) = -X_t(x),$$

et par suite: le point (i) est démontré. Démontrons le point (ii).

L'application ϕ_x est évidemment de classe C^∞, d'autre part la dérivée à l'origine de ϕ_x est déterminée par les vecteurs:

$$\left(\frac{\partial \phi_x}{\partial t_i}\right)_{(0,\cdots,0)} = X^i(x).$$

ϕ_x est donc de rang p à l'origine ce qui démontre le point (ii). Le point (iii) est une conséquence du lemme 1.2.1. En effet on a vu que

$$\mathscr{L}\left[\left(\frac{\partial \phi_x}{\partial t_i}\right)_{(0,\cdots,0)}; i = 1, \cdots, p\right] = \mathscr{L}(D(x)).$$

On a d'autre part:

$$\left(\frac{\partial \phi_x}{\partial t_i}\right)_{(t_1,0,0,0,)} = X_{t_1}^*(x) \cdot \left(\frac{\partial \phi_x}{\partial t_i}\right)_{(0,\cdots,0)}$$

par conséquence d'après le lemme 1.2.1 :

$$\left(\frac{\partial \phi_x}{\partial t_i}\right)_{(t_1,0,0,0)} \in \mathscr{L}(D(X_{t_1}(x))) = \mathscr{L}(D(\phi_x(t_1, 0, \cdots, 0))).$$

En opérant de proche en proche on démontre ainsi le point (iii).

Démontrons pour terminer un troisième lemme. On considère la situation suivante. Soit D une F.C.V. sur M, symétrique, localement de type fini. Soit

$$\psi : \mathscr{U} \to M$$

une immersion définie sur un ouvert de R^p telle que l'on ait

$$\operatorname{Im}(\psi^*(t)) = \mathscr{L}(D(\psi(t))), \qquad t \in \mathscr{U}.$$

Soit x un point de $\psi(\mathscr{U})$. Soit ϕ_x l'application définie par

$$\phi_x(t_1, \cdots, t_p) = X^p_{t_p} \circ \cdots \circ X^1_{t_1}(x),$$

où les champs $X^1, \cdots, X^p$ définissent une base de $\mathscr{L}(D(x))$.

On peut alors énoncer le lemme suivant.

LEMME 1.2.3. *L'ensemble* $(\phi_x)^{-1}(\psi(\mathscr{U}))$ *est un voisinage de* 0.

Démonstration. Soit $(t_1, \cdots, t_p)$ un point de $\mathscr{V}_x$ tel que

$$\phi_x(t_1, \cdots, t_p) \in \psi(\mathscr{U}).$$

L'application ψ étant une immersion on pourra toujours par application de la proposition 1.1.1 se ramener à la situation suivante :

M est un ouvert de R^n contenant l'origine.

$\psi(\mathscr{U})$ est l'intersection d'un ouvert $\mathscr{W}$ de R^n contenant l'origine avec la variété linéaire :

$$L_p = \{x_1 \cdots x_n ; x_{p+1} = x_{p+2} = x_n = 0\}.$$

La famille D est telle que

$$x \in L_p \cap \mathscr{W} \Rightarrow \mathscr{L}(D(x)) = L_p,$$

$$\phi_x(t_1, t_2, \cdots, t_p) = 0.$$

Notons θ l'application définie par

$$(\theta_1, \cdots, \theta_p) \in R^p \to \theta(\theta_1, \cdots, \theta_p) = \phi_x(t_1 + \theta_1, \cdots, t_p + \theta_p).$$

Cette application est définie sur un ouvert de R^p. On vérifie d'autre part que

$$\theta(\theta_1, \cdots, \theta_p) = Y^p_{\theta_p} \circ \cdots \circ Y^1_{\theta_1}(0),$$

où les champs Y^i sont définis par

$$Y^p(m) = X^p(m),$$

$$Y^{p-1}(m) = X^{p*}_{t_p}(X^p_{-t_p}(m) \cdot X^{p-1}(X^p_{-t_p}(m)),$$

$$Y^i(m) = X^{p*}_{t_p} \circ X^{p-1*}_{t_{p-1}} \circ \cdots \circ X^{i+1*}_{t_{i+1}}(X^{i+1}_{-t_{i+1}} \circ \cdots \circ X^p_{-t_p}(m)) \cdot X^i(X^{i+1}_{-t_{i+1}} \circ \cdots \circ X^p_{-t_p}(m)).$$

D'après le lemme 1.2.1 les champs Y^i sont tels que

$$Y^i(m) \in \mathscr{L}(D(m)).$$

Par conséquence sur $\mathscr{L}_p \cap \mathscr{W}$ on a

$$Y^i(m) \in L_p, \qquad i = 1, \cdots, p, \quad m \in L_p \cap \mathscr{W},$$

et par suite d'après les théorèmes classiques sur l'intégration des équations différentielles on voit que pour $(\theta_1, \cdots, \theta_p)$ suffisamment voisins de 0 le point $\theta(\theta_1, \cdots, \theta_p) = \phi(t_1 + \theta_1, \cdots, t_p + \theta_p)$ appartient à $L_p \cap \mathscr{W}$, donc à $\psi(\mathscr{U})$ ce qui démontre le lemme.

Nous pouvons maintenant démontrer la proposition suivante qui est une petite généralisation de la proposition 2.1 de [8].

PROPOSITION 1.2.3 (Hermann). *Soit D une F.C.V. sur M, symétrique, localement de type fini. Soit x un point de M. La feuille intégrale L_x de D passant par x peut être munie d'une structure de variété différentiable S telle que le couple (L_x, S) soit une sous-variété intégrale de D.*

Démonstration. Soit pour chaque y de L_x l'application :

$$\phi_y : \mathscr{V}_y \to L_x$$

définie au lemme 1.2.2. Montrons que la famille de "cartes" définie par les ϕ_y munit L_x d'une structure de variété différentiable.

(i) Soit $p(y)$ la dimension de l'espace vectoriel sur lequel est défini $\mathscr{V}_y$. On vérifie immédiatement par application du lemme 1.2.1 que $p(y)$ est constante sur tout chemin intégral de D et, par suite, constante sur L_x.

(ii) Les applications $\phi_y : \mathscr{U}_y \to M$ sont des applications injectives de $\mathscr{U}_y$ dans L_x. Montrons que les changements de cartes sont bien C^∞. Soit

$$\phi_z : \mathscr{U}_z \to M$$

une autre carte. Si on applique le lemme 1.2.3 en prenant pour ψ l'application ϕ_z on voit que l'ensemble

$$\phi_y^{-1}(\phi_z(\mathscr{U}_z))$$

est un ouvert de R^p. Le théorème des fonctions implicites montre ensuite que l'application

$$\phi_z \circ \phi_z : \phi_y^{-1}(\phi_z(\mathscr{U}_z)) \to \mathscr{U}_z$$

est de classe C^∞.

(iii) Soit $S = \{\phi_y : \mathscr{U}_y \to L_x \,;\, y \in L_x\}$. L'atlas S définit sur L_x une structure de variété différentiable. Par construction même de S on a

$$T(L_x, S)_y = \phi_y^*(0) \cdot (R^p) = \mathscr{L}(D(y)).$$

Il ne reste donc qu'à montrer de (L_x, S) est une variété connexe ((L_x, S) est trivialement séparée). Pour cela il suffit de montrer qu'un chemin intégral joignant x à y est un arc de (L_x, S), donc vérifier que pour tout X dans D et tout y dans L_x l'application

$$t \to X_t(y) \in L_x$$

est continue ce qui se vérifie aisément en procédant comme au lemme 1.2.3.

Remarque 1. La construction d'une structure de variété différentiable sur L_x qui vient d'être faite diffère sensiblement de celle que propose Hermann dans [8].

Les cartes que nous proposons constituent un système local de "coordonnées curvilignes" de L_x reposant sur la connaissance d'une base de D_x.

Hermann suppose que D est un sous-espace vectoriel et définit une carte au point y en choisissant un sous-espace D_y de D tel que l'application

$$Y \to Y(y)$$

soit un isomorphisme de D_y sur $D(y)$. Ensuite à chaque Y de D_y il associe le point $Y_1(y)$ de L_x.

A condition de choisir Y suffisamment petit on peut ainsi définir une carte au voisinage de y.

D'un point de vue pratique (numérique) dans la construction que nous avons proposé une carte peut être complètement décrite par l'intégration de p équations différentielles.

Remarque 2. L'hypothèse selon laquelle D est localement de type fini est essentielle comme le montre l'exemple suivant. C'est le même exemple que l'exemple 1 mais interprèté de façon différente.

Exemple 2. Soit D la F.C.V. définie sur R^2 dans l'exemple 1. L'allure des courbes intégrales du champ X est donnée par Fig. 2. L'allure des courbes intégrales du champ Y est donnée par Fig. 3. Il est clair que la feuille intégrale de D passant par l'origine est égale à R^2 tout entier alors que la dimension de D à l'origine est seulement égale à 1. Ce contre exemple repose sur le fait qu'il existe des fonctions C^∞ nulles sur un ouvert, non identiquement nulles ce qui n'est pas le cas pour des fonctions analytiques.

On déduit immédiatement de la proposition 1.2.3 et du lemme 1.2.3 la proposition suivante.

PROPOSITION 1.2.4. *Soit D une F.C.V. sur M symétrique, localement de type fini. Soit x un point de M. Toute sous-variété intégrale de D passant par x est une sous-variété de (L_x, S), où S est la structure définie en* 1.2.3.

Supposons maintenant que M soit une variété analytique et $V(M)$ le module des champs de vecteurs analytiques définis sur M. Rappelons les résultats classiques d'algèbre suivants.

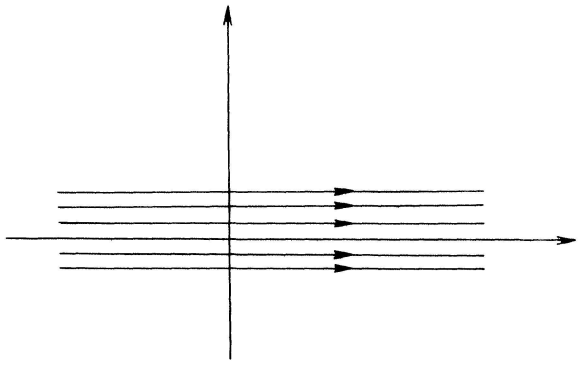


FIG. 2.

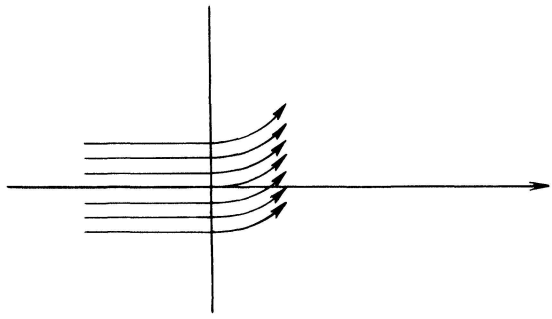


FIG. 3.

DÉFINITION 1.2.6. On dit qu'un *A-module E est Noetherien* si toute famille non vide de sous-modules de E possède un élément maximal. Un *anneau A est Noetherien* si, considéré comme A-module, il est Noetherien.

PROPOSITION 1.2.5. *Si A est un anneau Noetherien, l'A-module* $\prod_{i=1}^{n} A_i (A_i = A)$ *est Noetherien* [20, p. 55].

PROPOSITION 1.2.6. *L'anneau des séries entières convergentes à n indéterminées réelles,* $K[[x_1, \cdots, x_n]]$ *est un anneau Noetherien* [21, p. 149].

PROPOSITION 1.2.7. *Soit D une famille de champs de vecteurs analytiques définis sur la variété M, close pour l'opération de crochet, c'est à dire telle que*

$$(X \in D,\ Y \in D) \Rightarrow [X\,Y] \in D.$$

La famille D est alors "localement de type fini."

Démonstration. La condition étant locale il suffit de démontrer la proposition 1.2.7 pour une famille D d'applications analytiques de R^n dans R^n.

Soit $\mathscr{A}(R^n, R^n)$ l'ensemble des applications analytiques de R^n dans R^n. Soit $\mathscr{GA}(R^n, R^n)$ l'ensemble des germes en 0 d'application analytiques de R^n dans R^n.

On sait que $\mathscr{GA}(R^n, R^n)$ est le quotient de $\mathscr{A}(R^n, R^n)$ par la relation d'équivalence:

"$f \sim g \Leftrightarrow f$ et g coïncident sur un voisinage de 0".

Notons π la projection canonique:

$$\pi : \mathscr{A}(R^n, R^n) \to \mathscr{GA}(R^n, R^n).$$

Le module $\mathscr{GA}(R^n, R^n)$ est isomorphe au module: $(K[[x_1, \cdots, x_n]])^n$ des séries entières convergentes à n indéterminées réelles qui est Noetherien (prop. 1.2.5 et 1.2.6).

Soit $\mathscr{T}_f$ l'ensemble des parties finies de $\pi(D)$. La famille

$$\{M_\alpha ; \alpha \in \mathscr{T}_f\}$$

(M_α = module engendré par α) possède un élément maximal: $M_{(a_1 \cdots a_p)}$; $a_i \in \pi(D)$.

Tout élément a de $\pi(D)$ peut donc s'écrire:

$$a = \sum_{i=1}^{p} \phi_i a_i, \qquad \phi_i \in K[[x_1, \cdots, x_n]].$$

Soient $X^1, \cdots, X^p$ p éléments de D tels que $\pi(X_i) = a_i$. Etant donné un élément quelconque X de D (en particulier de la forme $X = [YX^i]$, $Y \in D$) on a

$$\pi_X = \sum_{i=1}^{p} \phi_i \pi_{Xi}.$$

Par conséquence il existe un voisinage de l'origine sur lequel on a

$$X(x_1, \cdots, x_n) = \sum_{i=1}^{p} \phi_i(x_1, \cdots, x_n) X^i(x_1, \cdots, x_n)$$

ce qui démontre la proposition 1.2.7.

On déduit des propositions 1.2.3 et 1.2.7 la proposition suivante.

PROPOSITION 1.2.8. *Soit D une famille de champs de vecteurs analytiques, symétrique, définie sur la variété analytique M, stable par crochet (i.e., $(X \in D, Y \in D) \Rightarrow [XY] \in D$). Quel que soit x dans M, l'ensemble L_x peut être muni d'une structure de variété S telle que (L_x, S) soit une sous-variété intégrale de D.*

Enfin on obtient la proposition 1.2.1 en appliquant la proposition 1.2.4 à la famille symétrique engendrée par D.

Remarque. La proposition 1.2.3 s'applique à la situation décrite par Kučera dans [14]. En effet ce dernier énonce : Soit $\mathcal{U}$ un sous-espace vectoriel de $\mathcal{L}(R^n, R^n)$ tel que

$$(A \in \mathcal{U}, B \in \mathcal{U}) \Rightarrow AB - BA \in \mathcal{U}.$$

La famille D de champs de vecteurs :

$$D = \{x \to Ax \,;\, A \in \mathcal{U}\}$$

est stable par crochet et localement de type fini ($\mathcal{L}(R^n, R^n)$ est un espace vectoriel de dimension finie).

1.3. Le théorème de Chow. Nous énonçons dans ce paragraphe une forme particulière du théorème de Chow.

Nous donnons sur un exemple une idée de la démonstration. La demonstration générale n'est pas plus difficile mais la multiplication des indices la rend fastidieuse. Elle est rédigée dans [17].

Soit dans R^3 la famille de champs de vecteurs :

$$D = \{\pm X^1 \,;\, \pm X^2\}.$$

PROPOSITION 1.3.1. *Si les vecteurs*

$$X^1(x_0), X^2(x_0), [X^1 X^2](x_0)$$

sont indépendants, alors la feuille intégrale de D passant par x_0 est un voisinage de x_0.

Démonstration. Soit Z_t^λ la famille de groupes à un paramètre définis par

$$Z_t^\lambda(x) = X_{-\lambda}^1 \circ X_t^2 \circ X_\lambda^1(x).$$

Si on note $V_x(\lambda)$ le vecteur

$$V_x(\lambda) = \left(\frac{d}{dt} Z_t^\lambda(x)\right)_{t=0},$$

on a

$$V_x(\lambda) = X^{1*}_{-\lambda}(X^1_\lambda(x)) \cdot X^2(X^1_\lambda(x)) ;$$

donc par définition du crochet de deux champs on a

$$\left(\frac{d}{d\lambda} V_x(\lambda)\right)_{\lambda=0} = [X^1 X^2](x).$$

Soit ψ_λ la famille d'applications de R^3 dans R^3 définie par

$$\psi_\lambda(t_1, t_2, t_3) = Z^\lambda_{t_3} \circ X^2_{t_2} \circ X^1_{t_1}(x_0).$$

Par définition de l'ensemble L_x on a l'inclusion

$$\psi_\lambda(\mathcal{U}) \subset L_x,$$

où $\mathcal{U}$ est un voisinage de 0 dans R^3 sur lequel ψ_λ est bien définie. Montrons que ψ_λ est pour un λ convenable un difféomorphisme local au voisinage de 0 ce qui montrera la proposition. Pour cela calculons $\psi^*_\lambda(0,0,0)$. On a

$$\frac{\partial \psi_\lambda}{\partial t_1}(0,0,0) = X^1(x_0),$$

$$\frac{\partial \psi_\lambda}{t_2}(0,0,0) = X^2(x_0),$$

$$\frac{\partial \psi_\lambda}{t_3}(0,0,0) = V_{x_0}(\lambda).$$

Il est clair que

$$V_{x_0}(0) = X^2(x_0).$$

Soit $\Delta(\lambda)$ l'application de R dans R défini par

$$\Delta(\lambda) = \det(X^1(x_0), X^2(x_0), V_{x_0}(\lambda)).$$

On a

$$\Delta(0) = 0, \qquad \frac{d\Delta}{d\lambda}(\lambda)_{\lambda=0} = \det(X^1(x_0), X^2(x_0), [X^2 X^1](x_0)).$$

Par hypothèse les vecteurs

$$X^1(x_0), X^2(x_0), [X^1 X^2]x_0$$

sont indépendants.

On a donc

$$\Delta(0) = 0, \qquad \Delta'(0) \neq 0.$$

Par conséquence pour λ suffisamment petit on a $\Delta(\lambda) \neq 0$.

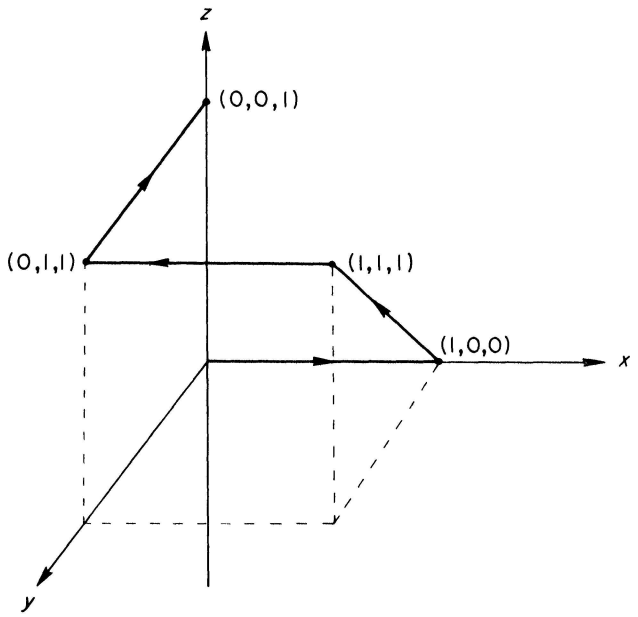


FIG. 4.

Exemple 3. Sur Fig. 4 on voit comment en utilisant les courbes intégrales des champs

$$X = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad Y = \begin{pmatrix} 0 \\ 1 \\ x \end{pmatrix}$$

on peut atteindre le point (0, 0, 1) en partant de l'origine.

Afin de généraliser la proposition 1.3.1 à un système symétrique quelconque de vecteurs de R^n introduisons la notion de saturée d'une famille de champs de vecteurs.

Soit D une famille de champs de vecteurs sur une variété M. On note ΔD la famille:

$$\Delta D = \{[XY]; X; Y : X \in D, Y \in D\}$$

et $\Delta^n(D)$ la famille définie par

$$\Delta^{n+1}(D) = \Delta(\Delta^n(D)).$$

DÉFINITION 1.3.1. Soit D une F.C.V. sur la variété M. On appelle *saturée de* D et on note $\tilde{D}$ la famille définie par

$$\tilde{D} = \bigcup_{n \in N} \Delta^n(D).$$

On remarque que si D est symétrique, $\tilde{D}$ l'est également. D'autre part même si la famille D est finie on peut avoir pour tout n,

$$\Delta^n(D) \not\supseteq \Delta^{n-1}(D).$$

D'autre part, par construction même, $\tilde{D}$ est stable par crochet. On a alors la proposition suivante.

PROPOSITION 1.3.2. THÉORÈME DE CHOW. *Soit D une F.C.V. sur R^n symétrique telle que* $\dim(\mathscr{L}(\tilde{D}(x_0))) = n$. *Alors la feuille intégrale de D passant par x_0 est un voisinage de x_0.*

Démonstration. On procède comme dans la démonstration de la proposition 1.3.1 en utilisant des familles de groupes à un paramètre un peu plus élaborées que la famille Z_λ (cf. [17]).

Du théorème de Chow on déduit immédiatement les propositions:

PROPOSITION 1.3.3. *Soit D une F.C.V. symétrique définie sur une variété M de dimension n. Si la dimension de l'espace vectoriel $\mathscr{L}(\tilde{D}(x_0))$ est égale à n, alors la feuille intégrale L_{x_0} de D passant par x_0 est un ouvert de M.*

Démonstration. C'est une conséquence immédiate de la proposition 1.3.2.

PROPOSITION 1.3.4. *Soit D une F.C.V. symétrique définie sur une variété connexe M de dimension n. Si pour tout x de M on a*

$$\dim \mathscr{L}(\tilde{D}(x)) = n,$$

alors la feuille intégrale passant par x est égale à M.

Démonstration. C'est une conséquence immédiate de la connexité et de la proposition 1.3.1.

PROPOSITION 1.3.5. *Soit D une famille de champs de vecteurs analytique, symétrique sur la variété M. Notons L_x et $\tilde{L}_x$ les feuilles intégrales de D et $\tilde{D}$ passant par x.*

Quel que soit x dans M on a l'égalité

$$L_x = \tilde{L}_x.$$

Démonstration. D'après la proposition 1.2.8 appliquée à la famille $\tilde{D}$ on peut munir $\tilde{L}_x$ d'une structure de variété différentiable. Il suffit ensuite d'appliquer la proposition précédente à la restriction de D à $\tilde{L}_x$.

En l'absence d'hypothèse d'analyticité on obtiendrait une proposition analogue en supposant que $\tilde{D}$ est localement de type fini.

En théorie du contrôle il est souvent important de savoir si l'ensemble des états accessibles est d'intérieur vide. La réponse à cette question sera donnée par la proposition suivante.

PROPOSITION 1.3.6. *Soit D une famille de champs de vecteurs analytiques sur R^n. Une condition nécessaire pour que la feuille intégrale L_x de D passant par x soit d'intérieur non vide est que la dimension de l'espace vectoriel*

$$\mathscr{L}(\tilde{D}(x))$$

soit égale à n. Si de plus la famille D est symétrique la condition est suffisante.

Démonstration. Supposons que

$$\dim(\mathscr{L}(\tilde{D}(x))) = p < n.$$

Soit D' la famille symétrique engendrée par D. On a évidemment

$$\mathscr{L}(\tilde{D}(x)) = \mathscr{L}(\tilde{D}'(x)).$$

La feuille intégrale L'_x de D' passant par x contient L_x. D'autre part on peut munir L'_x d'une structure de variété différentiable S telle que (L'_x, S) soit une sous-variété connexe de R^n de dimension P.

On sait que toute sous-variété connexe et séparée d'une variété σ compacte est σ-compacte [1, p. 97]. La variété R^n est σ-compacte, (L'_x, S) est donc σ-compacte, comme elle est de dimension strictement plus petite que n l'ensemble L'_x est d'intérieur vide. La condition est donc nécessaire. Si D est symétrique la condition est suffisante d'après la proposition 1.3.3.

2. Critères intrinsèques en théorie du contrôle. Nous allons utiliser les résultats précédents pour obtenir des critères intrinsèques en théorie du contrôle. Nous entendons par critère intrinsèque tout critère qui peut se déduire par des opérations *algébriques* sur la fonction $f(x, t, u)$. Par exemple, le critère de contrôlabilité de Kalman [13]: "Le système

$$\frac{dx}{dt} = Ax + Bu, \quad x \in R^n, \qquad u \in R^p,$$

est contrôlable si

$$\operatorname{rang}[A, AB, \cdots, A^{n-1}B] = n,"$$

est un critère intrinsèque. Par contre le critère: "Le système

$$\frac{dx}{dt} = A(t)x + B(t)u, \qquad x \in R^n, \qquad u \in R^p,$$

est contrôlable si

$$\operatorname{rang} \int_0^T \phi^{-1}(t_0 t)B(t)^t B(t)^t \phi^{-1}(t_0 t)\, dt = n,$$

où $\phi(t_0 t)$ est la matrice fondamentale du système," n'est pas un critère intrinsèque car il s'exprime à l'aide de $\phi(t_0 t)$ qui s'obtient à partir des données en *intégrant une équation différentielle*, ce qui n'est pas toujours possible pratiquement.

Le "principe du maximum" n'est pas non plus un critère intrinsèque puisqu'il fait intervenir également la résolvante du système le long d'une trajectoire présumée optimale.

2.1. Théorie du contrôle et familles de champs de vecteurs. Les problèmes de contrôle se formulent classiquement de la manière suivante.

Soit $f: R^n \times R \times R^p \to R^n$ une application suffisamment régulière. Soit $\mathcal{M}$ un sous-ensemble de l'ensemble des applications intégrables d'un intervalle de R dans une partie U de R^p. A tout couple:

$$(x_0, (\mathcal{U} : [t_0 t_1] \to U))$$

constitué d'une condition initiale x_0 de R^n et d'un "contrôle" admissible $(\mathcal{U} : [t_0 t_1] \to U)$ on associe l'état final:

$$(x_0, (\mathcal{U} : [t_0 t_1] \to U)) \to x(x_0, t_0, t_1, \mathcal{U}) \in R^n$$

défini par

$$\frac{d}{dt}x(x_0, t_0, t, \mathcal{U}) = f(x, t, u(t)),$$

$$x(t_0) = x_0.$$

Cette formulation, très générale, rend compte de nombreuses situations pratiques. Nous allons diminuer la généralité de cette situation en supposant que f est une fonction C^∞ ou analytique de tous ses arguments et que $\mathcal{M}$ est l'ensemble des fonctions C^∞ ou analytique par morceaux d'un intervalle de R dans U. La perte de généralité entraînée par ces restrictions n'est pas très grave dans la pratique.

Donnons maintenant quelques définitions précises.

DÉFINITION 2.1.1. L'expression:

"Soit le système

$$\frac{dx}{dt} = f(x, t, u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)."}$$

Sous-entend:

(i) Soit $f: R^n \times R \times R^p \to R^n$ de classe C^∞ (resp. analytique).

(ii) Soit U une partie quelconque de R^p, $\mathcal{M}(U)$ l'ensemble des applications C^∞ par morceaux (resp. analytique par morceaux) d'un intervalle de R dans U.

(iii) Au couple $x_0 \in R^n$, $(\mathcal{U}: [t_0 t_1] \to U)$ de $\mathcal{M}(U)$ on associe l'élément

$$x(x_0, t_0, t_1, \mathcal{U}) \in R^n,$$

défini par

$$(*) \qquad \begin{cases} t \in [t_0 t_1] \dfrac{d}{dt}x(x_0, t_0, t, \mathcal{U}) = f(x(x_0, t_0, t, \mathcal{U}), t, \mathcal{U}(t)), \\ x(x_0, t_0, t_0, \mathcal{U}) = x_0. \end{cases}$$

L'application f étant supposée telle que les solutions de l'équation différentielle (*) soient toujours définies pour $t = t_1$.

DÉFINITION 2.1.2. Soit le système

$$\frac{dx}{dt} = f(x, t, u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)}.$$

On appelle ensemble des états (x_0, t_0)-*accessibles à l'instant* t_1 $(t_1 > t_0)$, et on note $A(x_0, t_0, t_1, U)$ l'ensemble:

$$A(x_0, t_0, t_1, U) = \{x(x_0, t_0, t_1, \mathcal{U}) : \mathcal{U} \in C^\infty_{\mathrm{M}}([t_0 t_1] \to U)\}$$

$$\text{(resp. } \mathrm{anal}_{\mathrm{M}}([t_0 t_1] \to U)),$$

où $C^\infty_{\mathrm{M}}([t_0 t_1] \to U)$ (resp. $\mathrm{anal}_{\mathrm{M}}([t_0 t_1] \to U)$) désigne l'ensemble des applications C^∞ (resp. analytiques) par morceaux de $[t_0 t_1]$ dans U.

On appelle ensemble des états (x_0, t_0)-accessibles et on note $A(x_0, t_0, U)$ l'ensemble:

$$A(x_0, t_0, U) = \bigcup_{t_1 > t_0} A(x_0, t_0, t_1, U).$$

Pour finir définissons la contrôlabilité "bang-bang" de la manière suivante.

DÉFINITION 2.1.3. Soient les systèmes:

$$\text{(i)}\quad \frac{dx}{dt} = f(x,t,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)},$$

$$\text{(ii)}\quad \frac{dx}{dt} = f(x,t,u), \qquad x \in R^n, \qquad u \in \overline{U} \subset R^p, \qquad C^\infty \text{ (resp. anal.)}.$$

Supposons que $\overline{U}$ soit inclus dans U. On dit que le système (i) est *strictement* "$\overline{U}$-*bang-bang*" *contrôlable* si quel que soient x_0, t_0, t_1 on a l'égalité

$$A(x_0, t_0, t_1, \overline{U}) = A(x_0, t_0, t_1, U).$$

On dit que le système (i) est "$\overline{U}$-*bang-bang*" *contrôlable* (*au sens large*) si

$$A(x_0, t_0, \overline{U}) = A(x_0, t_0, U).$$

Il est clair que lorsqu'on s'intéresse à des problèmes de contrôle en temps minimum seule la notion de stricte "bang-bang" contrôlabilité est utilisable. Par contre si dans un problème d'optimisation on étudie un critère sur l'état final on pourra utiliser la "bang-bang" contrôlabilité large. Dans la suite nous n'examinerons que la contrôlabilité "bang-bang" au sens large.

Nous relions maintenant la notion de système à celle de famille de champs de vecteurs.

DÉFINITION 2.1.4. Soit le système:

$$\frac{dx}{dt} = f(x,t,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal)}.$$

Notons I l'ensemble des applications C^∞ (resp. analytiques) de R dans U. On appelle *famille de champs de vecteurs associée à* (f, U) et on note:

$$(X^i)_{i \in I}; \qquad I = C^\infty(R, U) \text{ (resp. anal. } (R \to U)),$$

la famille de champs de vecteurs sur R^{n+1} définie par

$$X^i(x,t) = \begin{pmatrix} f(x,t,i(t)) \\ 1 \end{pmatrix}, \qquad i \in I.$$

On utilise également la notation:

$$(X^i)_{i \in I} = D(f, U).$$

La proposition suivante est une conséquence immédiate des définitions 1.2.2, 2.1.2, 2.1.4.

PROPOSITION 2.1.1. *Soit le systéme*:

$$\frac{dx}{dt} = f(x,t,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)}$$

et $D(f, U)$ *sa famille de champs de vecteurs associée.*

Soit x_0 un point de R^n, t_0 un réel. On a les égalités suivantes:

$$A(x_0, t_0, t_1, U) = \pi(L_{(x_0,t_0)} \cap \{(x,t); t = t_1\}),$$
$$A(x_0, t_0, U) = \pi(L_{(x_0,t_0)}),$$

où π est la projection de R^{n+1} sur R^n:

$$(x,t) \to x$$

et $L_{(x_0,t_0)}$ est la feuille intégrale (déf. 1.2.3) *de $D(f, U)$ passant par (x_0, t_0).*

Les systèmes autonomes sont susceptibles d'un traitement différent que nous définissons maintenant.

DÉFINITION 2.1.5. On appelle système *autonome* un système de la forme:

$$\frac{dx}{dt} = f(x,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)}.$$

On dit de plus que le système est dénombrable si l'ensemble U est dénombrable.

DÉFINITION 2.1.6. Soit

$$\frac{dx}{dt} = f(x,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)}$$

un système autonome *dénombrable*. On appelle famille de champs de vecteurs associée à un tel système et on note:

$$(X^u)u \in U$$

la famille de champs de vecteurs sur R^n définie par

$$X^u(x) = f(x,u).$$

On a alors la proposition suivante.

PROPOSITION 2.1.2. *Soit*

$$(*) \qquad \frac{dx}{dt} = f(x,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)}$$

un système autonome dénombrable. On a l'égalité:

$$A(x_0, 0, U) = L_{x_0},$$

où L_{x_0} est la feuille intégrale de la famille de champs de vecteur $(X^u)u \in U$ associée à (*).

Démonstration. Il suffit de remarquer que si U est dénombrable les applications C^∞ (resp. analytiques) par morceaux de R dans U sont constantes par morceaux.

2.2. Ensemble des états accessibles étude locale. Comme exemples d'application des résultats de §§ 1.2 et 1.3 nous donnons les propositions suivantes. Elles ne sont pas très fines et il est certain qu'un travail du type de celui de Kučera [14], [15] permettrait d'obtenir des résultats plus intéressants.

On considère le système:

$$(1) \qquad \frac{dx}{dt} = f(x, t, u), \qquad x \in R^n, \qquad u \in U \subset R^p \text{ (anal.)},$$

et sa famille de champs de vecteurs associée $D(f, U)$ (cf. déf. 2.1.4).

On note enfin $\tilde{D}(f, U)$ la saturée (déf. 1.3.1) de $D(f, U)$.

PROPOSITION 2.2.1. *Soit dans* R^{n+1} *un couple*:

$$(\bar{x}, t_1); \quad \bar{x} \in A(x_0, t_0, t_1, U).$$

L'ensemble des points de R^{n+1} *de la forme*

$$(x, t); \qquad x \in A(x_0, t_0, t, U)$$

est inclus dans une sous-variété de R^{n+1} *dont l'espace tangent au point*

$$(\bar{x}, t_1)$$

est défini par

$$\mathscr{L}(\tilde{D}(f, U)(\bar{x}, t_1)).$$

Démonstration. D'après la proposition 2.1.1 le point $(\bar{x}, t_1)$ appartient, ainsi que tous les points (x, t) tels que $x \in A(x_0, t_0, t, U)$, à la feuille intégrale $L_{(x_0 t_0)}$ de $D(f, U)$ qui d'après (prop. 1.2.1) est incluse dans une sous-variété de R^{n+1} dont l'espace tangent en $(\bar{x}, t_1)$ est $\mathscr{L}(\tilde{D}(f, U)(\bar{x}, t_1))$.

Remarque 1. Ce résultat concernant le couple

$$(\bar{x} = x(x_0, t_0, t_1, \bar{u}), t_1)$$

dans l'espace des (x, t) ne fait intervenir que le couple $(\bar{x}, t_1)$ et pas le contrôle $\bar{u}:[t_0 t_1] \to U$ conduisant à $\bar{x}$. En principe l'espace $\mathscr{L}(D(f, U)(\bar{x}, t_1))$ peut être déterminé par des opérations de crochet (pas nécessairement en nombre fini!).

Remarque 2. Le résultat est très faible et ne renseigne que très peu sur le comportement de l'ensemble des états accessibles au voisinage de $\bar{x}$.

Remarque 3. On ne peut rien obtenir par ce procédé concernant les états accessibles $A(x_0, t_0, U)$ car la projection selon π de la sous-variété $L_{(x_0, t_0)}$ n'est pas une sous-variété de R^n. Par contre concernant l'ensemble des états $A(x_0 t_0, t_1, U)$ on peut obtenir la proposition suivante.

PROPOSITION 2.2.2. *Soit dans* R^n *le point* $\bar{x}$ *appartenant à* $A(x_0, t_0, t_1, U)$. *L'ensemble* $A(x_0, t_0, t_1, U)$ *est inclus dans une sous-variété dont l'espace tangent au point* $\bar{x}$ *est l'espace vectoriel*:

$$\mathscr{L}(D(f, U)(\bar{x}, t_1)) \cap \{(x, t); t = 0\}.$$

Démonstration. L 'ensemble de R^{n+1}

$$\{(x, t_1); x \in A(x_0, t_0, t_1, U)\}$$

est inclus dans la sous-variété de R^n:

$$L_{(x_0, t_0)} \cap \{(x, t); t = t_1\}.$$

Remarque. Ce résultat est un peu plus précis que le précédent. Il reste cependant très insuffisant comme le montre l'exemple classique suivant, du à Filippov.

Exemple 4.

$$\frac{dx}{dt} = -y^2 + u^2, \qquad x \in R,$$

$$u \in R, \ |u| = 1.$$

$$\frac{dy}{dt} = u, \qquad y \in R,$$

Dans R^3 la famille de champs de vecteurs associée est :

$$X^1(x, y, t) = \begin{pmatrix} -y^2 + 1 \\ 1 \\ 1 \end{pmatrix},$$

$$X^2(x, y, t) = \begin{pmatrix} -y^2 + 1 \\ -1 \\ 1 \end{pmatrix}.$$

Calculons les crochets. On a

$$[X^1X^2](xy) = \begin{pmatrix} 0 & -2y & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} -y^2 + 1 \\ -1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 & -2y & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} -y^2 + 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4y \\ 0 \\ 0 \end{pmatrix},$$

$$[[X^1X^2]X^1] = \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}.$$

Les vecteurs X^1, X^2, $[[X^1X^2]X^1]$ sont toujours indépendants. L'ensemble $A(x_0, 0, 1, U)$ est inclus dans un ouvert de R^2, ce qui n'apporte évidemment aucun renseignement supplémentaire!

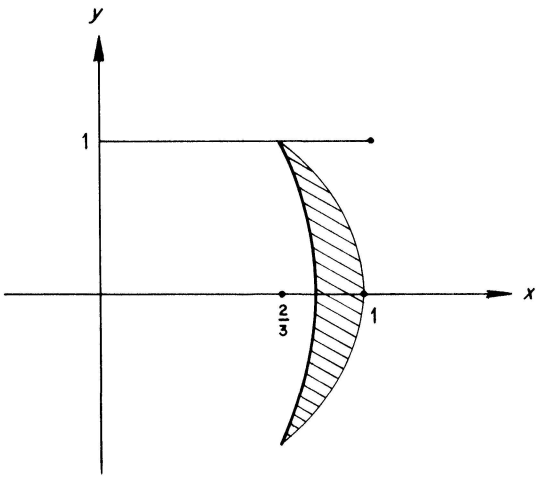


FIG. 5.

Quelques calculs élémentaires montrent que l'allure de $A(x_0, 0, 1, U)$ est donnée par Fig. 5. L'ensemble des états accessibles à l'instant 1 est un ensemble limité par deux arcs, l'arc passant par le point (1, 0) n'appartenant pas à l'ensemble. Filippov a montré par cet exemple qu'en l'absence d'hypothèse de convexité sur l'ensemble :

$$\{f(x, u); u \in U\},$$

il ne peut pas exister de contrôle en temps minimum. Ceci est très clair lorsqu'on représente dans R^3 la feuille intégrale du système défini par les deux champs X^1 et X^2. La troisième composante des champs étant constamment égale à 1 la troisième coordonnée s'interprète comme le temps écoulé depuis le départ. On voit facilement que cette feuille intégrale est une partie de R^3 limitée par deux surfaces, la surface inférieure n'appartenant pas à la partie, ce qui entraine la non-existence d'un contrôle en temps minimum.

On peut enfin obtenir une condition nécessaire pour que l'ensemble $A(x_0, t_0, t_1, U)$ soit d'intérieur non vide.

PROPOSITION 2.2.3. *Une condition nécessaire pour que l'ensemble $A(x_0, t_0, t_1, U)$ des états (x_0, t_0)-accessibles à l'instant t_1 du système* (1) *(analytique) soit d'intérieur non vide est que*

$$\mathscr{L}(\tilde{D}(f, U) \cdot (x_0, t_0))$$

soit de dimension $n + 1$.

Démonstration. Supposons que $\dim(\mathscr{L}(\tilde{D}(f, U) \cdot (x_0, t_0))) \leqq n$; alors pour tout $\bar{x} \in A(x_0, t_0, t_1, U)$ on a

$$\dim\{\mathscr{L}(\tilde{D}(f, U) \cdot (\bar{x}, t_1)) \cap \{(x, t); t = 0\}\} < n$$

car l'espace $\mathscr{L}(\tilde{D}(f, U) \cdot (\bar{x}t_1))$ est toujours transverse à l'hyperplan $\{(x, t); t = 0\}$.

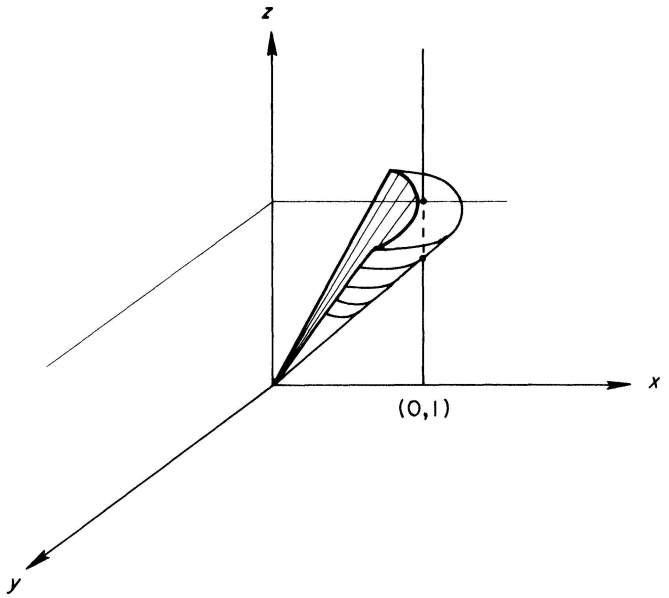


FIG. 6.

La proposition découle alors de la proposition 1.3.5.

Exemple 5. Appliquons ce résultat au système:

$$\frac{dx}{dt} = Ax + Bu, \qquad x \in R^n, \qquad u \in R^p, \qquad U = \{(u_1, \cdots, u_p); |u_i| = 1\},$$

$$A \in \mathscr{L}(R^n, R^n), \qquad B \in \mathscr{L}(R^p, R^n),$$

$$x(0) = 0.$$

On sait qu'une condition nécessaire et suffisante (Kalman et Halkin) pour que l'ensemble des états (0, 0)-accessibles du système ci-dessus soit d'intérieur non vide est que

$$\text{rang}\,[B, AB, \cdots, A^{n-1}B] = n;$$

en appliquant la proposition 2.2.2 on obtient la condition *nécessaire*. En effet, la famille de champs de vecteurs associée est:

$$D(f, U) = \{Ax \pm V_i; i = 1, \cdots, p\},$$

où V_1, V_p représentent les colonnes de B.

On a alors

$$[Ax + V_i, Ax + V_j] = AV_j - AV_i = A[V_j - V_i].$$

On obtient donc tous les champs:

$$AV_1, AV_2, \cdots, AV_p$$

(à un facteur multiplicatif près). Si on recommence on a

$$[Ax + V_i, AV_j] = A \cdot AV_j = A^2V_j.$$

On voit donc que la saturée de la famille $D\,(f, U)$ contient les champs:

$$Ax \pm V_i, \pm AV_i, \pm A^2V_i, \cdots, \pm A^{n-1}V_i, \qquad i = 1, \cdots, p.$$

Le théorème de Hamilton-Caeley montre qu'il est inutile de continuer (A^n s'exprime à partir des A^i, $i \leqq n - 1$) et que par conséquence on a obtenu la saturée de $\tilde{D}\,(f, U)$. D'après la proposition 2.2.3 une condition nécessaire pour que $A\,(0, 0, t_1, U)$ soit d'intérieur non vide est que l'espace vectoriel engendré par les vecteurs

$$\pm V_i, \pm AV_i, \cdots, \pm A^{n-1}V_i, \qquad i = 1, \cdots, p,$$

soit de dimension n, ce qui peut également s'écrire

$$\text{rang}\,[B, AB, \cdots, A^{n-1}B] = n.$$

2.3. Etats accessibles d'un système autonome symétrique. Nous étudions ici les systèmes autonomes particuliers du type suivant.

DÉFINITION 2.3.1. On appelle système *symétrique* un système de la forme

$$\frac{dx}{dt} = H(x) \cdot u, \qquad x \in R^n, \qquad u \in U \subset R^p \text{ (anal.)},$$

SIAM J. CONTROL
Vol. 8, No. 4, November 1970

CONTROLABILITE DES SYSTEMES NON LINEAIRES*

CLAUDE LOBRY†

Introduction. Dans cet article on étudie le problème de l'accessibilité pour des systèmes non linéaires du type:

$$\frac{dx}{dt} = f(x, t, u), \qquad x \in R^n, \qquad u \in \Omega \subset R^p, \tag{1}$$

et plus particulièrement dans le cas où Ω n'est pas un ensemble convexe.

Cette étude repose pour l'essentiel sur un théorème dû à Chow [2]. R. Hermann a, le premier, montré dans [7] et [8] comment ce théorème pouvait être appliqué avec fruit en théorie du contrôle. Depuis H. Hermes [10]–[12], utilisant le théorème de Chow également, a abordé ce problème en termes de systèmes de Pfaff. Parallèlement dans [14] et [15] Kučera fait une étude très fine concernant les propriétés géométriques de l'ensemble des états accessibles, pour un système linéaire particulier. Toutes ces études reposent fondamentalement sur l'utilisation de dérivées de Lie de champs de vecteurs. Nous ne proposons pas dans cet article des résultats nouveaux importants mais plutôt une approche géométrique systématique du problème. Pour cela nous avons partagé l'exposé en deux parties. Dans la première nous exposons en termes purement mathématiques des résultats dus essentiellement à Hermann et Chow; dans la seconde, nous interprétons ces résultats en termes de contrôlabilité. Lorsqu'un résultat est proche d'un résultat classique des références sont données, cependant la bibliographie proposée est loin d'être exhaustive, en particulier tous les travaux concernant les équations du type:

$$\frac{dx}{dt} \in \Gamma(x, t), \tag{2}$$

tels que ceux de Wajeski, Filippov, Castaing, etc. ont été délibérément omis. En fait, les méthodes et les résultats proposés ici sont de nature très différente.

Le paragraphe 1.1 est uniquement consacré à l'introduction de définitions classiques en géométrie différentielle. Il ne contient aucun résultats.

Le paragraphe 1.2 est consacré à l'étude des "variétés intégrales" d'une famille de champs de vecteurs. C'est en un certain sens une généralisation de l'étude de R. Hermann [7]. La proposition 1.2.1 est le résultat central de ce paragraphe. Les exemples qui l'accompagnent montrent que c'est le résultat le plus précis que l'on puisse obtenir dans le contexte choisi. Au point de vue géométrique il serait plus logique de s'intéresser à l'intégration des "distributions cohérentes" [22], [23], i.e., se donner, de manière suffisamment régulière, en chaque point de la variété un sous-espace de l'espace tangent en ce point. Beaucoup des résultats énoncés ici peuvent se traduire immédiatement sauf, précisément, la proposition 1.2.1. Le point de vue adopté (famille de champs de vecteurs) permet l'utilisation du language géométrique et s'interprète immédiatement en termes de contrôle.

* Received by the editors November 18, 1969, and in final revised form February 24, 1970.

† Mathématiques Appliquées, Université de Grenoble, Cedex 53, 38 Grenoble-Gare, France.

Le paragraphe 1.3 propose une démonstration du théorème de Chow dont l'interpretation en termes de la théorie du contrôle est immédiate.

Le paragraphe 2.1 établit les liens entre le formalisme précédemment développé et le formalisme classique de la théorie du contrôle. Les propositions qui y sont énoncées sont des conséquences immédiates des définitions.

Le paragraphe 2.2 est consacré à l'étude locale de l'ensemble des états accessibles d'un système "contrôlé." Il s'agit de corollaires des résultats de la première partie. Ces résultats ne sont pas classiques, et il est possible qu'une étude plus précise menée dans la même direction apporte d'autres renseignements.

Le paragraphe 2.3 est consacré à l'étude des systèmes du type:

$$\frac{dx}{dt} = H(x)\cdot u, \qquad x \in R^n, \qquad u \in R^p, \qquad H(x) \in \mathscr{L}(R^p, R^n).$$

Ces systèmes ont été introduits dans [10]. Une conjecture raisonnable concernant la "bang-bang" contrôlabilité est proposée. Cette conjecture fait apparaître la possibilité de décrire l'ensemble des états accessibles de certains systèmes comme l'ensemble:

$$\{x \in R^n ; g_i(x) \leqq 0, i = 1, 2, \cdots, p\},$$

où les applications g_i sont des applications différentiables de R^n dans R.

Les problèmes variationnels issus de problèmes de contrôle, (contrôle optimal) n'ont pas étés abordés. Il est clair que l'étude locale du paragraphe 2.2 peut être utile dans l'étude de problèmes d'optimisation.

1. Intégrabilité des familles de champs de vecteurs.

1.1. Notations. Nous introduisons les notations utilisées dans la suite. Pour la définition du vocabulaire de géométrie différentielle utilisée on pourra se reporter aux ouvrages classiques suivants: [1], [9], [18], [19].

Nous supposons systématiquement que les variétés, champs de vecteurs, fonctions que nous utilisons sont de classe C^∞. Cette hypothèse ne sera plus mentionnée par la suite. On supposera de plus que toutes les variétés sont séparées.

Soit M une variété, on note:

TM_x l'espace tangent à M en x;

$C^\infty(M)$ l'anneau des fonctions (C^∞) définies sur M;

$V(M)$ le $C(M)$-module des champs de vecteurs (C^∞) sur M.

Soient M et N deux variétés, $\phi : M \to N$ une application de M dans N. On note:

ϕ^* la différentielle de ϕ;

$\phi^*_{(x)}$ la valeur de ϕ^* au point x. $\phi^*_{(x)}$ est alors une application

$$\phi^*(x) : TM_x \to TN_{\phi(x)}$$

dont on note

$$\phi^*(x)\cdot h,$$

la valeur en un point h de TM_x.

Soit X un champ de vecteur défini sur M. On note :

$X_t(\cdot)$ le groupe local à un paramètre engendré par X. On sait qu'en général $X_t(\cdot)$ n'est défini que pour des valeurs de t suffisamment petites. Pour simplifier les notations nous omettrons systématiquement le "pour t assez petit." Il est facile de voir qu'aucune difficulté supplémentaire n'est liée à cette question dans ce qui suit. Sous les réserves exprimés ci-dessus on peut dire alors que $X_t(\cdot)$ est une application de $R \times M$ dans M :

$$(x, t) \to X_t(x).$$

On a de plus les relations :

$$X_0(x) = x,$$

$$X_{t+t'}(x) = X_t(X_{t'}(x)).$$

Pour t fixé on note : X_t^* la différentielle de l'application

$$x \to X_t(x).$$

Dans ces conditions $X_t^*(x)$ est une application linéaire inversible de TM_x dans $TM_{X_t(x)}$ satisfaisant aux relations :

$$X_0^*(x) = \text{identité},$$

$$(X_t^*(x))^{-1} = X_{-t}^*(X_t(x)).$$

Soient X et Y deux champs de vecteurs définis sur M, on note :

$[XY]$ le crochet de Jacobi des champs X et Y.

On sait que si on note $V_x(t)$ le vecteur de TM_x défini par

$$V_x(t) = (Y_t^*(x)^{-1}(X(Y_t(x))),$$

on a par définition de $[XY]$,

$$\left(\frac{d}{dt} V_x(t)\right)_{t=t_0} = (Y_{t_0}^*(x))^{-1} \cdot [XY](Y_{t_0}(x)).$$

On pourra trouver dans [24] une interprètation géométrique de cette notion. Pour ce qui nous intéresse la meilleure interprètation que l'on puisse donner est le théorème de Chow lui méme tel qu'il est démontré en § 1.3.

Si d'autre part $(x_1, x_2, \cdots, x_n)$ est un système de coordonnées locales de M, $(\partial x_1, \cdots, \partial x_n)$ la base de TM_x associée, si on note

$$X(x) = \sum_{i=1}^{n} X_i(x_1, \cdots, x_n) \frac{\partial}{\partial x_i},$$

on a

$$[XY](x_1 \cdots x_n) = X^*(x_1 \cdots x_n) \cdot Y(x_1 \cdots x_n) - Y^*(x_1 \cdots x_n) X(x_1 \cdots x_n),$$

où $X^*(x_1 \cdots x_n)$ est la matrice :

$$X^*(x_1 \cdots x_n) = \left(\frac{\partial X_i}{\partial x_j}(x_1 \cdots x_n)\right) \qquad i = 1, \cdots, n, \quad j = 1, \cdots, n.$$

Rappelons pour terminer une conséquence classique du théorème des fonctions implicites. Soit $\phi : M \to N$ une application de la variété M dans la variété N. On dit que ϕ est une immersion si quel que soit x dans M la valeur en x de la différentielle

$$\phi^*(x) : TM_x \to TN_{\phi(x)}$$

est une application linéaire injective.

PROPOSITION 1.1.1. *Soit ϕ une immersion de M dans N. Quel que soit x dans M il existe un voisinage $\mathcal{U}$ de x et un voisinage $\mathcal{V}$ de $\phi(x)$ tels que :*

(i) *ϕ restreinte à $\mathcal{U}$ est injective ;*

(ii) *il existe un système de coordonnées locales sur $\mathcal{V}$, $(y_1 \cdots y_n)$ tel que $\phi(\mathcal{U})$ soit défini par*

$$\phi(\mathcal{U}) = \{(y_1 \cdots y_n) : y_1 = y_2 = \cdots = y_p = 0\},$$

où p est égal à $m - n$, m et n désignant respectivement la dimension de M et de N.

1.2. Intégrabilité des familles de champs de vecteurs. Soit M une variété. Introduisons la définition suivante.

DÉFINITION 1.2.1. Soit D une famille de champs de vecteurs définis sur M. On dit qu'une *sous-variété N de M est une sous-variété intégrale de D* si N est connexe et si pour tout x de N on a l'égalité

$$TN_x = \mathcal{L}(D(x)),$$

où $\mathcal{L}(D(x))$ est l'espace vectoriel engendré par l'ensemble

$$D(x) = \{X(x) ; X \in D\}.$$

Le résultat essentiel que nous démontrons dans ce paragraphe est la proposition suivante.

PROPOSITION 1.2.1. *Soit D une famille de champs de vecteurs analytiques définis sur la variété analytique M, stable pour l'opération de crochet, c'est à dire telle que*

$$(X \in D,\ Y \in D) \Rightarrow [XY] \in D.$$

Par tout point x de M il passe une unique sous-variété intégrale de D, maximale pour l'inclusion.

Nous nous proposerons de plus une description précise de la structure de variété de la sous-variété de D qui sera interprétée par la suite (§ 2.1) en termes de

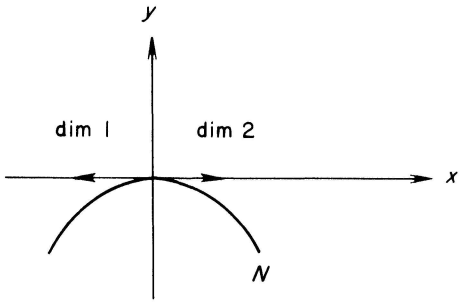


FIG. 1.

contrôle. Les idées contenues dans ce paragraphe sont très directement inspirées de celles de R. Hermann [8] et plus précisément le point essentiel, le lemme 1.2.1, correspond au lemme 2.1 de Hermann.

La difficulté de ce théorème est due à ce que la dimension de l'espace vectoriel $D(x)$ n'est pas supposée être constante. Lorsqu'elle est constante le classique théorème de Frobenius (cf. prop. 1.2.2) s'applique. L'exemple qui suit montre ce qui peut se produire quand la dimension n'est pas constante.

Exemple 1. Considèrons dans R^2 les deux champs suivants:

$$X(x, y) = \begin{pmatrix} 1 \\ 0 \end{pmatrix};$$

$$Y(x, y) = \begin{cases} \begin{pmatrix} 1 \\ \exp\left(-\dfrac{1}{x^2}\right) \end{pmatrix} & \text{si} \quad x > 0, \\ \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \text{si} \quad x \leqq 0. \end{cases}$$

Ces deux champs sont de classe C^∞. On vérifie immédiatement que la famille stable par crochet engendrée par les champs $\pm X$ et $\pm Y$ est la famille D définie par

$$D = \{\pm X;\ \pm Y;\ \pm Y^{(n)},\ n \in N\},$$

où le champ $Y^{(n)}$ est le champ

$$Y^{(n)}_{(x,y)} = \begin{cases} \begin{pmatrix} 1 \\ \left(\exp\left(-\dfrac{1}{x^2}\right)\right)^{(n)} \end{pmatrix} & \text{si} \quad x > 0, \\ \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \text{si} \quad x \leqq 0, \end{cases}$$

où $(\exp(-1/x^2))^{(n)}$ désigne la dérivée n-ième de $\exp(-1/x^2)$. On a alors

$$\dim \mathscr{L}(D_{(xy)}) = \begin{cases} 2 & \text{si} \quad x > 0, \\ 1 & \text{si} \quad x \leqq 0 \end{cases} \qquad \text{(cf. Fig. 1.)}$$

Supposons que par $(0, 0)$ il passe une sous-variété intégrale de la famille D. L'espace tangent en $(0, 0)$ à cette sous-variété est donc la droite $\{(x, y); y = 0\}$. Il s'ensuit que cette sous-variéte "pénètre" nécessairement dans le demi-espace $\{x, y; x \geqq 0\}$, où elle devrait avoir une dimension égale à 2 ce qui est évidemment impossible.

Notons "F.C.V." une famille de champs de vecteurs sur une variété M.

DÉFINITION 1.2.2. Soit D une F.C.V. On dit qu'un arc continu:

$$\alpha : [ab] \to M$$

est un *chemin intégral* de D si α est indéfiniment différentiable par morceaux et si pour tout intervalle I inclus dans $[ab]$ sur lequel α est différentiable il existe un

champ X de D tel que

$$\frac{d\alpha}{dt}(t) = X(\alpha(t)), \qquad t \in I.$$

Par indéfiniment différentiable par morceaux il faut entendre plus précisément que $[ab]$ est union finie d'intervalles $I_1 \cdots I_q$ tels que α restreint à I_j; soit la restriction à I_j d'une application indéfiniment différentiable de R dans M. On obtient donc un chemin intégral de D en "recollant continuement" un nombre fini de courbes intégrales de champs X de D.

DÉFINITION 1.2.3. Soit D une F.C.V. sur M. On appelle *feuille intégrale* de D passant par X et on note: L_x l'ensemble des points de M qui peuvent être joints à x par un chemin intégral de D.

On verra qu'à quelques réserves près la feuille intégrale L_x est précisément la sous-variété intégrale de D passant par x. On déduit immédiatement du classique théorème de Frobenius sur la complète intégrabilité des systèmes de Pfaff le résultat suivant concernant L_x.

PROPOSITION 1.2.2. *Soit D une F.C.V. définie sur M telle que*:

(i) *D est un sous-module de $V(M)$;*

(ii) $(X \in D,\ Y \in D) \Rightarrow [XY] \in D$;

(iii) *la dimension de l'espace vectoriel*

$$D(x) = \{X(x);\ X \in D\}$$

est indépendante de x et égale à p.

Alors l'ensemble L_x peut être muni d'une structure de variété différentiable S telle que la sous-variété (L_x, S) soit l'unique sous-variété intégrale maximale de D passant par x.

Démonstration. On sait d'après le théorème de Frobenius qu'il existe une unique sous-variété intégrale maximale de D passant par X. Il suffit donc de constater, ce qui est très clair d'après les définitions, que L_x coïncide avec cette sous-variété.

Dans [8] Hermann propose un théorème dans lequel l'hypothèse selon laquelle la dimension de $D(x)$ est constante est supprimée au profit d'autres conditions de régularité. C'est ce théorème que nous allons montrer après les lemmes suivants.

DÉFINITION 1.2.4. On dit qu'une famille D de champs de vecteurs définis sur M est *localement de type fini* si quelque soit x dans M il existe un nombre fini de champs de D:

$$X^1 \cdots X^q$$

tels que pour tout X de D il existe un voisinage $\mathscr{V}_{x,X}$ de x sur lequel on ait

$$[XX^j](y) = \sum_{i=1}^{q} f_i^j(y) X^i(y), \qquad y \in \mathscr{V}_{x,X}.$$

Cette condition est légèrement plus faible que la condition "locally finitely generated" de Hermann. On verra qu'elle est réalisée pour des familles de champs de vecteurs *analytiques*. Le lemme 2.1 de Hermann reste vrai sous cette hypothèse et on a alors le lemme suivant.

LEMME 1.2.1. (Hermann). *Soit D une F.C.V. sur M localement de type fini. Quels que soient x dans M, X dans D et t dans R tels que $X_t(x)$ soit défini, l'application*

$$X_t^*(x): TM_x \to TM_{X_t(x)}$$

définit un isomorphisme entre les espaces vectoriels

$$\mathscr{L}(D(x)) \quad et \quad \mathscr{L}(D(X_t(x))).$$

Démonstration. Il suffit de prouver que si le vecteur V appartient à l'espace $\mathscr{L}(D(x))$, alors le vecteur $X_t^*(x) \cdot V$ appartient à l'espace $\mathscr{L}(D(X_t(x)))$.

D'autre part l'arc:

$$\theta \to X_\theta(x), \qquad \theta \in [0, t],$$

est compact, il suffit donc de montrer la proposition plus faible suivante:

"*Pour tout x dans M il existe un réel $\varepsilon(x)$ strictement positif tel que pour tout t en valeur absolue inférieur à $\varepsilon(x)$ on ait*

$$V \in \mathscr{L}(D(x)) \Rightarrow X_t^*(x) \cdot V \in \mathscr{L}(D(X_t(x))).$$"

La famille D est localement de type fini, il existe donc q champs de vecteurs:

$$X^1 \cdots X^q$$

et un voisinage de x sur lequel

$$[XX^j](y) = \sum_{i=1}^{q} f_i^j(y) X^i(y), \qquad j = 1, \cdots, q.$$

Notons $V^j(t)$ le vecteur de TM_x défini par

$$V^j(t) = (X_t^*(x))^{-1} \cdot (X^j(X_t(x)).$$

Par définition du crochet de deux champs on a

$$\frac{d}{dt} V^j(t) = (X_t^*(x))^{-1} \cdot [XX^j](X_t(x)).$$

Soit $\varepsilon(x)$ un réel tel que

$$|t| \leqq \varepsilon(x) \Rightarrow X_t(x) \in \mathscr{V}_{x,X}.$$

On a alors pour $|t| \leqq \varepsilon(x)$,

$$\frac{d}{dt} V^j(t) = \sum_{i=1}^{q} f_i^j(X_t(x))(X_t^*(x))^{-1} \cdot X^i(X_t(x)), \qquad j = 1, \cdots, q.$$

Soit encore

$$\frac{d}{dt} V^j(t) = \sum_{i=1}^{q} f_i^j(X_t(x)) V^i(t), \qquad j = 1, \cdots, q.$$

Les q vecteurs $V^j(t)$ sont solution d'un système de q équations différentielles linéaires. Il existe donc des fonctions

$$\alpha_i^j(t), \qquad i = 1, \cdots, q, \quad j = 1, \cdots, q,$$

telles que

$$V^j(0) = \sum_{i=1}^{q} \alpha_i^j(t) V^i(t).$$

Soit encore

$$X_t^*(x) \cdot X^j(x) = \sum_{i=1}^{q} \alpha_i^j(t) X^i(X_t(x)).$$

Le vecteur $X_t^*(x) \cdot X^j(x)$ appartient donc à $\mathscr{L}(D(X_t(x)))$. Les vecteurs $X^j(x)$ constituant un système de générateurs de $\mathscr{L}(D(x))$; le lemme est démontré.

DÉFINITION 1.2.5. Soit D une F.C.V. sur M. On dit que D est *symétrique* si quel que soit X dans D le champ—X appartient à D.

LEMME 1.2.2. *Soit D une F.C.V. sur M, symétrique, localement de type fini. Soient $X^1, X^2, \cdots, X^p$ p champs définissant une base de $\mathscr{L}(D(x))$. Soit ϕ_x l'application définie par*

$$(t_1, t_2, \cdots, t_p) \in R^p \to \phi_x(t_1, \cdots, t_p) = X_{t_p}^p \circ X_{t_{p-1}}^{p-1} \circ \cdots \circ X_{t_1}^1(x) \in M.$$

Il existe un voisinage $\mathscr{U}_x$ de x tel que ϕ_x restreinte à $\mathscr{U}_x$ ait les propriétés suivantes:

(i) $\phi_x(\mathscr{U}_x) \subset L_x$;

(ii) *ϕ_x est une immersion injective;*

(iii) $\operatorname{Im}(\phi_x^*(t_1, \cdots, t_p)) = \mathscr{L}(D(\phi_x(t_1, \cdots, t_p)))$, *où* $\operatorname{Im}(\phi_x^*(t_1, \cdots, t_p))$ *désigne l'image de l'application*

$$\phi_x^*(t_1, \cdots, t_p)) : R^p \to TM_{\phi_x(t_1, \cdots, t_p)}.$$

En d'autres termes, ϕ_x définit une sous-variété intégrale de D passant par x.

Démonstration. La définition d'un chemin intégral impose que l'on parcoure les courbes intégrales des champs de D dans le sens des t croissants. Comme la famille D est symétrique on aura

$$X_{-t}(x) = -X_t(x),$$

et par suite: le point (i) est démontré. Démontrons le point (ii).

L'application ϕ_x est évidemment de classe C^∞, d'autre part la dérivée à l'origine de ϕ_x est déterminée par les vecteurs:

$$\left(\frac{\partial \phi_x}{\partial t_i}\right)_{(0,\cdots,0)} = X^i(x).$$

ϕ_x est donc de rang p à l'origine ce qui démontre le point (ii). Le point (iii) est une conséquence du lemme 1.2.1. En effet on a vu que

$$\mathscr{L}\left[\left(\frac{\partial \phi_x}{\partial t_i}\right)_{(0,\cdots,0)}; i = 1, \cdots, p\right] = \mathscr{L}(D(x)).$$

On a d'autre part:

$$\left(\frac{\partial \phi_x}{\partial t_i}\right)_{(t_1,0,0,0,)} = X_{t_1}^*(x) \cdot \left(\frac{\partial \phi_x}{\partial t_i}\right)_{(0,\cdots,0)}$$

par conséquence d'après le lemme 1.2.1 :

$$\left(\frac{\partial \phi_x}{\partial t_i}\right)_{(t_1,0,0,0)} \in \mathscr{L}(D(X_{t_1}(x))) = \mathscr{L}(D(\phi_x(t_1, 0, \cdots, 0))).$$

En opérant de proche en proche on démontre ainsi le point (iii).

Démontrons pour terminer un troisième lemme. On considère la situation suivante. Soit D une F.C.V. sur M, symétrique, localement de type fini. Soit

$$\psi : \mathscr{U} \to M$$

une immersion définie sur un ouvert de R^p telle que l'on ait

$$\operatorname{Im}(\psi^*(t)) = \mathscr{L}(D(\psi(t))), \qquad t \in \mathscr{U}.$$

Soit x un point de $\psi(\mathscr{U})$. Soit ϕ_x l'application définie par

$$\phi_x(t_1, \cdots, t_p) = X^p_{t_p} \circ \cdots \circ X^1_{t_1}(x),$$

où les champs $X^1, \cdots, X^p$ définissent une base de $\mathscr{L}(D(x))$.

On peut alors énoncer le lemme suivant.

LEMME 1.2.3. *L'ensemble* $(\phi_x)^{-1}(\psi(\mathscr{U}))$ *est un voisinage de* 0.

Démonstration. Soit $(t_1, \cdots, t_p)$ un point de $\mathscr{V}_x$ tel que

$$\phi_x(t_1, \cdots, t_p) \in \psi(\mathscr{U}).$$

L'application ψ étant une immersion on pourra toujours par application de la proposition 1.1.1 se ramener à la situation suivante :

M est un ouvert de R^n contenant l'origine.

$\psi(\mathscr{U})$ est l'intersection d'un ouvert $\mathscr{W}$ de R^n contenant l'origine avec la variété linéaire :

$$L_p = \{x_1 \cdots x_n ; x_{p+1} = x_{p+2} = x_n = 0\}.$$

La famille D est telle que

$$x \in L_p \cap \mathscr{W} \Rightarrow \mathscr{L}(D(x)) = L_p,$$

$$\phi_x(t_1, t_2, \cdots, t_p) = 0.$$

Notons θ l'application définie par

$$(\theta_1, \cdots, \theta_p) \in R^p \to \theta(\theta_1, \cdots, \theta_p) = \phi_x(t_1 + \theta_1, \cdots, t_p + \theta_p).$$

Cette application est définie sur un ouvert de R^p. On vérifie d'autre part que

$$\theta(\theta_1, \cdots, \theta_p) = Y^p_{\theta_p} \circ \cdots \circ Y^1_{\theta_1}(0),$$

où les champs Y^i sont définis par

$$Y^p(m) = X^p(m),$$

$$Y^{p-1}(m) = X^{p*}_{t_p}(X^p_{-t_p}(m) \cdot X^{p-1}(X^p_{-t_p}(m)),$$

$$Y^i(m) = X^{p*}_{t_p} \circ X^{p-1*}_{t_{p-1}} \circ \cdots \circ X^{i+1*}_{t_{i+1}}(X^{i+1}_{-t_{i+1}} \circ \cdots \circ X^p_{-t_p}(m)) \cdot X^i(X^{i+1}_{-t_{i+1}} \circ \cdots \circ X^p_{-t_p}(m)).$$

D'après le lemme 1.2.1 les champs Y^i sont tels que

$$Y^i(m) \in \mathscr{L}(D(m)).$$

Par conséquence sur $\mathscr{L}_p \cap \mathscr{W}$ on a

$$Y^i(m) \in L_p, \qquad i = 1, \cdots, p, \quad m \in L_p \cap \mathscr{W},$$

et par suite d'après les théorèmes classiques sur l'intégration des équations différentielles on voit que pour $(\theta_1, \cdots, \theta_p)$ suffisamment voisins de 0 le point $\theta(\theta_1, \cdots, \theta_p) = \phi(t_1 + \theta_1, \cdots, t_p + \theta_p)$ appartient à $L_p \cap \mathscr{W}$, donc à $\psi(\mathscr{U})$ ce qui démontre le lemme.

Nous pouvons maintenant démontrer la proposition suivante qui est une petite généralisation de la proposition 2.1 de [8].

PROPOSITION 1.2.3 (Hermann). *Soit D une F.C.V. sur M, symétrique, localement de type fini. Soit x un point de M. La feuille intégrale L_x de D passant par x peut être munie d'une structure de variété différentiable S telle que le couple (L_x, S) soit une sous-variété intégrale de D.*

Démonstration. Soit pour chaque y de L_x l'application :

$$\phi_y : \mathscr{V}_y \to L_x$$

définie au lemme 1.2.2. Montrons que la famille de "cartes" définie par les ϕ_y munit L_x d'une structure de variété différentiable.

(i) Soit $p(y)$ la dimension de l'espace vectoriel sur lequel est défini $\mathscr{V}_y$. On vérifie immédiatement par application du lemme 1.2.1 que $p(y)$ est constante sur tout chemin intégral de D et, par suite, constante sur L_x.

(ii) Les applications $\phi_y : \mathscr{U}_y \to M$ sont des applications injectives de $\mathscr{U}_y$ dans L_x. Montrons que les changements de cartes sont bien C^∞. Soit

$$\phi_z : \mathscr{U}_z \to M$$

une autre carte. Si on applique le lemme 1.2.3 en prenant pour ψ l'application ϕ_z on voit que l'ensemble

$$\phi_y^{-1}(\phi_z(\mathscr{U}_z))$$

est un ouvert de R^p. Le théorème des fonctions implicites montre ensuite que l'application

$$\phi_z \circ \phi_z : \phi_y^{-1}(\phi_z(\mathscr{U}_z)) \to \mathscr{U}_z$$

est de classe C^∞.

(iii) Soit $S = \{\phi_y : \mathscr{U}_y \to L_x \,;\, y \in L_x\}$. L'atlas S définit sur L_x une structure de variété différentiable. Par construction même de S on a

$$T(L_x, S)_y = \phi_y^*(0) \cdot (R^p) = \mathscr{L}(D(y)).$$

Il ne reste donc qu'à montrer de (L_x, S) est une variété connexe ((L_x, S) est trivialement séparée). Pour cela il suffit de montrer qu'un chemin intégral joignant x à y est un arc de (L_x, S), donc vérifier que pour tout X dans D et tout y dans L_x l'application

$$t \to X_t(y) \in L_x$$

est continue ce qui se vérifie aisément en procédant comme au lemme 1.2.3.

Remarque 1. La construction d'une structure de variété différentiable sur L_x qui vient d'être faite diffère sensiblement de celle que propose Hermann dans [8].

Les cartes que nous proposons constituent un système local de "coordonnées curvilignes" de L_x reposant sur la connaissance d'une base de D_x.

Hermann suppose que D est un sous-espace vectoriel et définit une carte au point y en choisissant un sous-espace D_y de D tel que l'application

$$Y \to Y(y)$$

soit un isomorphisme de D_y sur $D(y)$. Ensuite à chaque Y de D_y il associe le point $Y_1(y)$ de L_x.

A condition de choisir Y suffisamment petit on peut ainsi définir une carte au voisinage de y.

D'un point de vue pratique (numérique) dans la construction que nous avons proposé une carte peut être complètement décrite par l'intégration de p équations différentielles.

Remarque 2. L'hypothèse selon laquelle D est localement de type fini est essentielle comme le montre l'exemple suivant. C'est le même exemple que l'exemple 1 mais interprèté de façon différente.

Exemple 2. Soit D la F.C.V. définie sur R^2 dans l'exemple 1. L'allure des courbes intégrales du champ X est donnée par Fig. 2. L'allure des courbes intégrales du champ Y est donnée par Fig. 3. Il est clair que la feuille intégrale de D passant par l'origine est égale à R^2 tout entier alors que la dimension de D à l'origine est seulement égale à 1. Ce contre exemple repose sur le fait qu'il existe des fonctions C^∞ nulles sur un ouvert, non identiquement nulles ce qui n'est pas le cas pour des fonctions analytiques.

On déduit immédiatement de la proposition 1.2.3 et du lemme 1.2.3 la proposition suivante.

PROPOSITION 1.2.4. *Soit D une F.C.V. sur M symétrique, localement de type fini. Soit x un point de M. Toute sous-variété intégrale de D passant par x est une sous-variété de (L_x, S), où S est la structure définie en* 1.2.3.

Supposons maintenant que M soit une variété analytique et $V(M)$ le module des champs de vecteurs analytiques définis sur M. Rappelons les résultats classiques d'algèbre suivants.

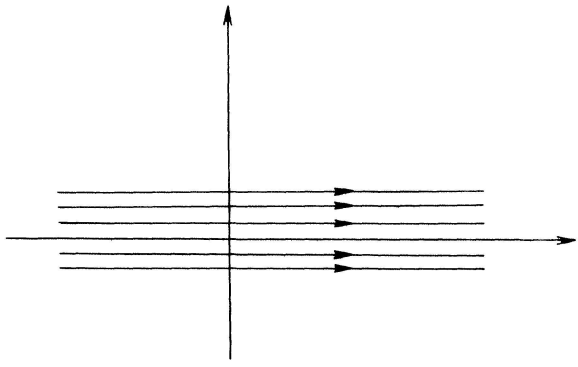


FIG. 2.

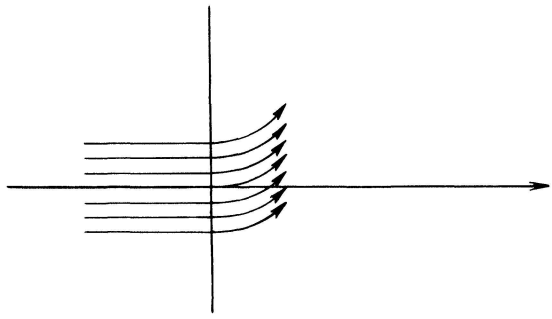


FIG. 3.

DÉFINITION 1.2.6. On dit qu'un *A-module E est Noetherien* si toute famille non vide de sous-modules de E possède un élément maximal. Un *anneau A est Noetherien* si, considéré comme A-module, il est Noetherien.

PROPOSITION 1.2.5. *Si A est un anneau Noetherien, l'A-module* $\prod_{i=1}^{n} A_i (A_i = A)$ *est Noetherien* [20, p. 55].

PROPOSITION 1.2.6. *L'anneau des séries entières convergentes à n indéterminées réelles,* $K[[x_1, \cdots, x_n]]$ *est un anneau Noetherien* [21, p. 149].

PROPOSITION 1.2.7. *Soit D une famille de champs de vecteurs analytiques définis sur la variété M, close pour l'opération de crochet, c'est à dire telle que*

$$(X \in D,\ Y \in D) \Rightarrow [X\,Y] \in D.$$

La famille D est alors "localement de type fini."

Démonstration. La condition étant locale il suffit de démontrer la proposition 1.2.7 pour une famille D d'applications analytiques de R^n dans R^n.

Soit $\mathscr{A}(R^n, R^n)$ l'ensemble des applications analytiques de R^n dans R^n. Soit $\mathscr{GA}(R^n, R^n)$ l'ensemble des germes en 0 d'application analytiques de R^n dans R^n.

On sait que $\mathscr{GA}(R^n, R^n)$ est le quotient de $\mathscr{A}(R^n, R^n)$ par la relation d'équivalence:

"$f \sim g \Leftrightarrow f$ et g coïncident sur un voisinage de 0".

Notons π la projection canonique:

$$\pi : \mathscr{A}(R^n, R^n) \to \mathscr{GA}(R^n, R^n).$$

Le module $\mathscr{GA}(R^n, R^n)$ est isomorphe au module: $(K[[x_1, \cdots, x_n]])^n$ des séries entières convergentes à n indéterminées réelles qui est Noetherien (prop. 1.2.5 et 1.2.6).

Soit $\mathscr{T}_f$ l'ensemble des parties finies de $\pi(D)$. La famille

$$\{M_\alpha ; \alpha \in \mathscr{T}_f\}$$

(M_α = module engendré par α) possède un élément maximal: $M_{(a_1 \cdots a_p)}$; $a_i \in \pi(D)$.

Tout élément a de $\pi(D)$ peut donc s'écrire:

$$a = \sum_{i=1}^{p} \phi_i a_i, \qquad \phi_i \in K[[x_1, \cdots, x_n]].$$

Soient $X^1, \cdots, X^p$ p éléments de D tels que $\pi(X_i) = a_i$. Etant donné un élément quelconque X de D (en particulier de la forme $X = [YX^i]$, $Y \in D$) on a

$$\pi_X = \sum_{i=1}^{p} \phi_i \pi_{Xi}.$$

Par conséquence il existe un voisinage de l'origine sur lequel on a

$$X(x_1, \cdots, x_n) = \sum_{i=1}^{p} \phi_i(x_1, \cdots, x_n) X^i(x_1, \cdots, x_n)$$

ce qui démontre la proposition 1.2.7.

On déduit des propositions 1.2.3 et 1.2.7 la proposition suivante.

PROPOSITION 1.2.8. *Soit D une famille de champs de vecteurs analytiques, symétrique, définie sur la variété analytique M, stable par crochet (i.e., $(X \in D, Y \in D) \Rightarrow [XY] \in D$). Quel que soit x dans M, l'ensemble L_x peut être muni d'une structure de variété S telle que (L_x, S) soit une sous-variété intégrale de D.*

Enfin on obtient la proposition 1.2.1 en appliquant la proposition 1.2.4 à la famille symétrique engendrée par D.

Remarque. La proposition 1.2.3 s'applique à la situation décrite par Kučera dans [14]. En effet ce dernier énonce : Soit $\mathcal{U}$ un sous-espace vectoriel de $\mathcal{L}(R^n, R^n)$ tel que

$$(A \in \mathcal{U}, B \in \mathcal{U}) \Rightarrow AB - BA \in \mathcal{U}.$$

La famille D de champs de vecteurs :

$$D = \{x \to Ax \,;\, A \in \mathcal{U}\}$$

est stable par crochet et localement de type fini ($\mathcal{L}(R^n, R^n)$ est un espace vectoriel de dimension finie).

1.3. Le théorème de Chow. Nous énonçons dans ce paragraphe une forme particulière du théorème de Chow.

Nous donnons sur un exemple une idée de la démonstration. La demonstration générale n'est pas plus difficile mais la multiplication des indices la rend fastidieuse. Elle est rédigée dans [17].

Soit dans R^3 la famille de champs de vecteurs :

$$D = \{\pm X^1 \,;\, \pm X^2\}.$$

PROPOSITION 1.3.1. *Si les vecteurs*

$$X^1(x_0), X^2(x_0), [X^1 X^2](x_0)$$

sont indépendants, alors la feuille intégrale de D passant par x_0 est un voisinage de x_0.

Démonstration. Soit Z_t^λ la famille de groupes à un paramètre définis par

$$Z_t^\lambda(x) = X_{-\lambda}^1 \circ X_t^2 \circ X_\lambda^1(x).$$

Si on note $V_x(\lambda)$ le vecteur

$$V_x(\lambda) = \left(\frac{d}{dt} Z_t^\lambda(x)\right)_{t=0},$$

on a

$$V_x(\lambda) = X^{1*}_{-\lambda}(X^1_\lambda(x)) \cdot X^2(X^1_\lambda(x)) ;$$

donc par définition du crochet de deux champs on a

$$\left(\frac{d}{d\lambda} V_x(\lambda)\right)_{\lambda=0} = [X^1 X^2](x).$$

Soit ψ_λ la famille d'applications de R^3 dans R^3 définie par

$$\psi_\lambda(t_1, t_2, t_3) = Z^\lambda_{t_3} \circ X^2_{t_2} \circ X^1_{t_1}(x_0).$$

Par définition de l'ensemble L_x on a l'inclusion

$$\psi_\lambda(\mathcal{U}) \subset L_x,$$

où $\mathcal{U}$ est un voisinage de 0 dans R^3 sur lequel ψ_λ est bien définie. Montrons que ψ_λ est pour un λ convenable un difféomorphisme local au voisinage de 0 ce qui montrera la proposition. Pour cela calculons $\psi^*_\lambda(0,0,0)$. On a

$$\frac{\partial \psi_\lambda}{\partial t_1}(0,0,0) = X^1(x_0),$$

$$\frac{\partial \psi_\lambda}{t_2}(0,0,0) = X^2(x_0),$$

$$\frac{\partial \psi_\lambda}{t_3}(0,0,0) = V_{x_0}(\lambda).$$

Il est clair que

$$V_{x_0}(0) = X^2(x_0).$$

Soit $\Delta(\lambda)$ l'application de R dans R défini par

$$\Delta(\lambda) = \det(X^1(x_0), X^2(x_0), V_{x_0}(\lambda)).$$

On a

$$\Delta(0) = 0, \qquad \frac{d\Delta}{d\lambda}(\lambda)_{\lambda=0} = \det(X^1(x_0), X^2(x_0), [X^2 X^1](x_0)).$$

Par hypothèse les vecteurs

$$X^1(x_0), X^2(x_0), [X^1 X^2]x_0$$

sont indépendants.

On a donc

$$\Delta(0) = 0, \qquad \Delta'(0) \neq 0.$$

Par conséquence pour λ suffisamment petit on a $\Delta(\lambda) \neq 0$.

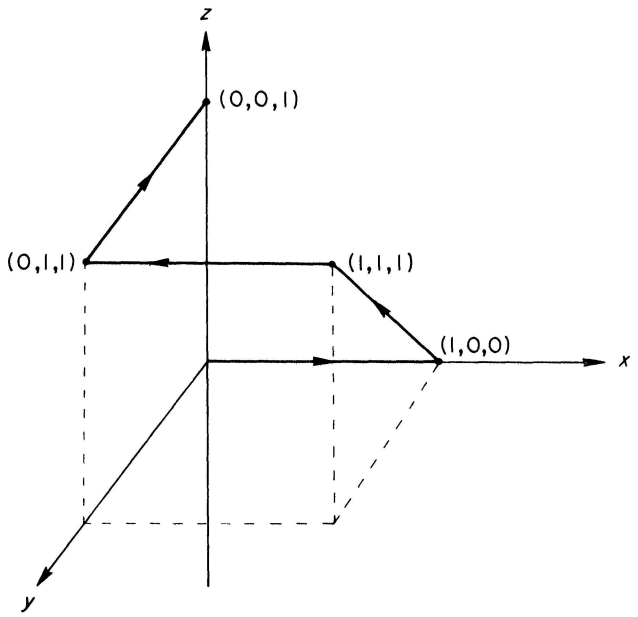


FIG. 4.

Exemple 3. Sur Fig. 4 on voit comment en utilisant les courbes intégrales des champs

$$X = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad Y = \begin{pmatrix} 0 \\ 1 \\ x \end{pmatrix}$$

on peut atteindre le point (0, 0, 1) en partant de l'origine.

Afin de généraliser la proposition 1.3.1 à un système symétrique quelconque de vecteurs de R^n introduisons la notion de saturée d'une famille de champs de vecteurs.

Soit D une famille de champs de vecteurs sur une variété M. On note ΔD la famille:

$$\Delta D = \{[XY]; X; Y : X \in D, Y \in D\}$$

et $\Delta^n(D)$ la famille définie par

$$\Delta^{n+1}(D) = \Delta(\Delta^n(D)).$$

DÉFINITION 1.3.1. Soit D une F.C.V. sur la variété M. On appelle *saturée de* D et on note $\tilde{D}$ la famille définie par

$$\tilde{D} = \bigcup_{n \in N} \Delta^n(D).$$

On remarque que si D est symétrique, $\tilde{D}$ l'est également. D'autre part même si la famille D est finie on peut avoir pour tout n,

$$\Delta^n(D) \not\supseteq \Delta^{n-1}(D).$$

D'autre part, par construction même, $\tilde{D}$ est stable par crochet. On a alors la proposition suivante.

PROPOSITION 1.3.2. THÉORÈME DE CHOW. *Soit D une F.C.V. sur R^n symétrique telle que* $\dim(\mathscr{L}(\tilde{D}(x_0))) = n$. *Alors la feuille intégrale de D passant par x_0 est un voisinage de x_0.*

Démonstration. On procède comme dans la démonstration de la proposition 1.3.1 en utilisant des familles de groupes à un paramètre un peu plus élaborées que la famille Z_λ (cf. [17]).

Du théorème de Chow on déduit immédiatement les propositions:

PROPOSITION 1.3.3. *Soit D une F.C.V. symétrique définie sur une variété M de dimension n. Si la dimension de l'espace vectoriel $\mathscr{L}(\tilde{D}(x_0))$ est égale à n, alors la feuille intégrale L_{x_0} de D passant par x_0 est un ouvert de M.*

Démonstration. C'est une conséquence immédiate de la proposition 1.3.2.

PROPOSITION 1.3.4. *Soit D une F.C.V. symétrique définie sur une variété connexe M de dimension n. Si pour tout x de M on a*

$$\dim \mathscr{L}(\tilde{D}(x)) = n,$$

alors la feuille intégrale passant par x est égale à M.

Démonstration. C'est une conséquence immédiate de la connexité et de la proposition 1.3.1.

PROPOSITION 1.3.5. *Soit D une famille de champs de vecteurs analytique, symétrique sur la variété M. Notons L_x et $\tilde{L}_x$ les feuilles intégrales de D et $\tilde{D}$ passant par x.*

Quel que soit x dans M on a l'égalité

$$L_x = \tilde{L}_x.$$

Démonstration. D'après la proposition 1.2.8 appliquée à la famille $\tilde{D}$ on peut munir $\tilde{L}_x$ d'une structure de variété différentiable. Il suffit ensuite d'appliquer la proposition précédente à la restriction de D à $\tilde{L}_x$.

En l'absence d'hypothèse d'analyticité on obtiendrait une proposition analogue en supposant que $\tilde{D}$ est localement de type fini.

En théorie du contrôle il est souvent important de savoir si l'ensemble des états accessibles est d'intérieur vide. La réponse à cette question sera donnée par la proposition suivante.

PROPOSITION 1.3.6. *Soit D une famille de champs de vecteurs analytiques sur R^n. Une condition nécessaire pour que la feuille intégrale L_x de D passant par x soit d'intérieur non vide est que la dimension de l'espace vectoriel*

$$\mathscr{L}(\tilde{D}(x))$$

soit égale à n. Si de plus la famille D est symétrique la condition est suffisante.

Démonstration. Supposons que

$$\dim(\mathscr{L}(\tilde{D}(x))) = p < n.$$

Soit D' la famille symétrique engendrée par D. On a évidemment

$$\mathscr{L}(\tilde{D}(x)) = \mathscr{L}(\tilde{D}'(x)).$$

La feuille intégrale L'_x de D' passant par x contient L_x. D'autre part on peut munir L'_x d'une structure de variété différentiable S telle que (L'_x, S) soit une sous-variété connexe de R^n de dimension P.

On sait que toute sous-variété connexe et séparée d'une variété σ compacte est σ-compacte [1, p. 97]. La variété R^n est σ-compacte, (L'_x, S) est donc σ-compacte, comme elle est de dimension strictement plus petite que n l'ensemble L'_x est d'intérieur vide. La condition est donc nécessaire. Si D est symétrique la condition est suffisante d'après la proposition 1.3.3.

2. Critères intrinsèques en théorie du contrôle. Nous allons utiliser les résultats précédents pour obtenir des critères intrinsèques en théorie du contrôle. Nous entendons par critère intrinsèque tout critère qui peut se déduire par des opérations *algébriques* sur la fonction $f(x, t, u)$. Par exemple, le critère de contrôlabilité de Kalman [13]: "Le système

$$\frac{dx}{dt} = Ax + Bu, \quad x \in R^n, \qquad u \in R^p,$$

est contrôlable si

$$\operatorname{rang}[A, AB, \cdots, A^{n-1}B] = n,"$$

est un critère intrinsèque. Par contre le critère: "Le système

$$\frac{dx}{dt} = A(t)x + B(t)u, \qquad x \in R^n, \qquad u \in R^p,$$

est contrôlable si

$$\operatorname{rang} \int_0^T \phi^{-1}(t_0 t)B(t)^t B(t)^t \phi^{-1}(t_0 t)\, dt = n,$$

où $\phi(t_0 t)$ est la matrice fondamentale du système," n'est pas un critère intrinsèque car il s'exprime à l'aide de $\phi(t_0 t)$ qui s'obtient à partir des données en *intégrant une équation différentielle*, ce qui n'est pas toujours possible pratiquement.

Le "principe du maximum" n'est pas non plus un critère intrinsèque puisqu'il fait intervenir également la résolvante du système le long d'une trajectoire présumée optimale.

2.1. Théorie du contrôle et familles de champs de vecteurs. Les problèmes de contrôle se formulent classiquement de la manière suivante.

Soit $f: R^n \times R \times R^p \to R^n$ une application suffisamment régulière. Soit $\mathcal{M}$ un sous-ensemble de l'ensemble des applications intégrables d'un intervalle de R dans une partie U de R^p. A tout couple:

$$(x_0, (\mathcal{U} : [t_0 t_1] \to U))$$

constitué d'une condition initiale x_0 de R^n et d'un "contrôle" admissible $(\mathcal{U} : [t_0 t_1] \to U)$ on associe l'état final:

$$(x_0, (\mathcal{U} : [t_0 t_1] \to U)) \to x(x_0, t_0, t_1, \mathcal{U}) \in R^n$$

défini par

$$\frac{d}{dt}x(x_0, t_0, t, \mathcal{U}) = f(x, t, u(t)),$$

$$x(t_0) = x_0.$$

Cette formulation, très générale, rend compte de nombreuses situations pratiques. Nous allons diminuer la généralité de cette situation en supposant que f est une fonction C^∞ ou analytique de tous ses arguments et que $\mathcal{M}$ est l'ensemble des fonctions C^∞ ou analytique par morceaux d'un intervalle de R dans U. La perte de généralité entraînée par ces restrictions n'est pas très grave dans la pratique.

Donnons maintenant quelques définitions précises.

DÉFINITION 2.1.1. L'expression:

"Soit le système

$$\frac{dx}{dt} = f(x, t, u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)."}$$

Sous-entend:

(i) Soit $f: R^n \times R \times R^p \to R^n$ de classe C^∞ (resp. analytique).

(ii) Soit U une partie quelconque de R^p, $\mathcal{M}(U)$ l'ensemble des applications C^∞ par morceaux (resp. analytique par morceaux) d'un intervalle de R dans U.

(iii) Au couple $x_0 \in R^n$, $(\mathcal{U}: [t_0 t_1] \to U)$ de $\mathcal{M}(U)$ on associe l'élément

$$x(x_0, t_0, t_1, \mathcal{U}) \in R^n,$$

défini par

$$(*) \qquad \begin{cases} t \in [t_0 t_1] \dfrac{d}{dt}x(x_0, t_0, t, \mathcal{U}) = f(x(x_0, t_0, t, \mathcal{U}), t, \mathcal{U}(t)), \\ x(x_0, t_0, t_0, \mathcal{U}) = x_0. \end{cases}$$

L'application f étant supposée telle que les solutions de l'équation différentielle (*) soient toujours définies pour $t = t_1$.

DÉFINITION 2.1.2. Soit le système

$$\frac{dx}{dt} = f(x, t, u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)}.$$

On appelle ensemble des états (x_0, t_0)-*accessibles à l'instant* t_1 $(t_1 > t_0)$, et on note $A(x_0, t_0, t_1, U)$ l'ensemble:

$$A(x_0, t_0, t_1, U) = \{x(x_0, t_0, t_1, \mathcal{U}) : \mathcal{U} \in C^\infty_{\mathrm{M}}([t_0 t_1] \to U)\}$$

$$\text{(resp. } \mathrm{anal}_{\mathrm{M}}([t_0 t_1] \to U)),$$

où $C^\infty_{\mathrm{M}}([t_0 t_1] \to U)$ (resp. $\mathrm{anal}_{\mathrm{M}}([t_0 t_1] \to U)$) désigne l'ensemble des applications C^∞ (resp. analytiques) par morceaux de $[t_0 t_1]$ dans U.

On appelle ensemble des états (x_0, t_0)-accessibles et on note $A(x_0, t_0, U)$ l'ensemble:

$$A(x_0, t_0, U) = \bigcup_{t_1 > t_0} A(x_0, t_0, t_1, U).$$

Pour finir définissons la contrôlabilité "bang-bang" de la manière suivante.

DÉFINITION 2.1.3. Soient les systèmes:

$$\text{(i)}\quad \frac{dx}{dt} = f(x,t,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)},$$

$$\text{(ii)}\quad \frac{dx}{dt} = f(x,t,u), \qquad x \in R^n, \qquad u \in \overline{U} \subset R^p, \qquad C^\infty \text{ (resp. anal.)}.$$

Supposons que $\overline{U}$ soit inclus dans U. On dit que le système (i) est *strictement* "$\overline{U}$-*bang-bang*" *contrôlable* si quel que soient x_0, t_0, t_1 on a l'égalité

$$A(x_0, t_0, t_1, \overline{U}) = A(x_0, t_0, t_1, U).$$

On dit que le système (i) est "$\overline{U}$-*bang-bang*" *contrôlable* (*au sens large*) si

$$A(x_0, t_0, \overline{U}) = A(x_0, t_0, U).$$

Il est clair que lorsqu'on s'intéresse à des problèmes de contrôle en temps minimum seule la notion de stricte "bang-bang" contrôlabilité est utilisable. Par contre si dans un problème d'optimisation on étudie un critère sur l'état final on pourra utiliser la "bang-bang" contrôlabilité large. Dans la suite nous n'examinerons que la contrôlabilité "bang-bang" au sens large.

Nous relions maintenant la notion de système à celle de famille de champs de vecteurs.

DÉFINITION 2.1.4. Soit le système:

$$\frac{dx}{dt} = f(x,t,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal)}.$$

Notons I l'ensemble des applications C^∞ (resp. analytiques) de R dans U. On appelle *famille de champs de vecteurs associée à* (f, U) et on note:

$$(X^i)_{i \in I}; \qquad I = C^\infty(R, U) \text{ (resp. anal. } (R \to U)),$$

la famille de champs de vecteurs sur R^{n+1} définie par

$$X^i(x,t) = \begin{pmatrix} f(x,t,i(t)) \\ 1 \end{pmatrix}, \qquad i \in I.$$

On utilise également la notation:

$$(X^i)_{i \in I} = D(f, U).$$

La proposition suivante est une conséquence immédiate des définitions 1.2.2, 2.1.2, 2.1.4.

PROPOSITION 2.1.1. *Soit le systéme*:

$$\frac{dx}{dt} = f(x,t,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)}$$

et $D(f, U)$ *sa famille de champs de vecteurs associée.*

Soit x_0 un point de R^n, t_0 un réel. On a les égalités suivantes:

$$A(x_0, t_0, t_1, U) = \pi(L_{(x_0,t_0)} \cap \{(x,t); t = t_1\}),$$
$$A(x_0, t_0, U) = \pi(L_{(x_0,t_0)}),$$

où π est la projection de R^{n+1} sur R^n:

$$(x,t) \to x$$

et $L_{(x_0,t_0)}$ est la feuille intégrale (déf. 1.2.3) *de $D(f, U)$ passant par (x_0, t_0).*

Les systèmes autonomes sont susceptibles d'un traitement différent que nous définissons maintenant.

DÉFINITION 2.1.5. On appelle système *autonome* un système de la forme:

$$\frac{dx}{dt} = f(x,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)}.$$

On dit de plus que le système est dénombrable si l'ensemble U est dénombrable.

DÉFINITION 2.1.6. Soit

$$\frac{dx}{dt} = f(x,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)}$$

un système autonome *dénombrable*. On appelle famille de champs de vecteurs associée à un tel système et on note:

$$(X^u)u \in U$$

la famille de champs de vecteurs sur R^n définie par

$$X^u(x) = f(x,u).$$

On a alors la proposition suivante.

PROPOSITION 2.1.2. *Soit*

$$(*) \qquad \frac{dx}{dt} = f(x,u), \qquad x \in R^n, \qquad u \in U \subset R^p, \qquad C^\infty \text{ (resp. anal.)}$$

un système autonome dénombrable. On a l'égalité:

$$A(x_0, 0, U) = L_{x_0},$$

où L_{x_0} est la feuille intégrale de la famille de champs de vecteur $(X^u)u \in U$ associée à (*).

Démonstration. Il suffit de remarquer que si U est dénombrable les applications C^∞ (resp. analytiques) par morceaux de R dans U sont constantes par morceaux.

2.2. Ensemble des états accessibles étude locale. Comme exemples d'application des résultats de §§ 1.2 et 1.3 nous donnons les propositions suivantes. Elles ne sont pas très fines et il est certain qu'un travail du type de celui de Kučera [14], [15] permettrait d'obtenir des résultats plus intéressants.

On considère le système:

$$(1) \qquad \frac{dx}{dt} = f(x, t, u), \qquad x \in R^n, \qquad u \in U \subset R^p \text{ (anal.)},$$

et sa famille de champs de vecteurs associée $D(f, U)$ (cf. déf. 2.1.4).

On note enfin $\tilde{D}(f, U)$ la saturée (déf. 1.3.1) de $D(f, U)$.

PROPOSITION 2.2.1. *Soit dans* R^{n+1} *un couple*:

$$(\bar{x}, t_1); \quad \bar{x} \in A(x_0, t_0, t_1, U).$$

L'ensemble des points de R^{n+1} *de la forme*

$$(x, t); \qquad x \in A(x_0, t_0, t, U)$$

est inclus dans une sous-variété de R^{n+1} *dont l'espace tangent au point*

$$(\bar{x}, t_1)$$

est défini par

$$\mathscr{L}(\tilde{D}(f, U)(\bar{x}, t_1)).$$

Démonstration. D'après la proposition 2.1.1 le point $(\bar{x}, t_1)$ appartient, ainsi que tous les points (x, t) tels que $x \in A(x_0, t_0, t, U)$, à la feuille intégrale $L_{(x_0 t_0)}$ de $D(f, U)$ qui d'après (prop. 1.2.1) est incluse dans une sous-variété de R^{n+1} dont l'espace tangent en $(\bar{x}, t_1)$ est $\mathscr{L}(\tilde{D}(f, U)(\bar{x}, t_1))$.

Remarque 1. Ce résultat concernant le couple

$$(\bar{x} = x(x_0, t_0, t_1, \bar{u}), t_1)$$

dans l'espace des (x, t) ne fait intervenir que le couple $(\bar{x}, t_1)$ et pas le contrôle $\bar{u}:[t_0 t_1] \to U$ conduisant à $\bar{x}$. En principe l'espace $\mathscr{L}(D(f, U)(\bar{x}, t_1))$ peut être déterminé par des opérations de crochet (pas nécessairement en nombre fini!).

Remarque 2. Le résultat est très faible et ne renseigne que très peu sur le comportement de l'ensemble des états accessibles au voisinage de $\bar{x}$.

Remarque 3. On ne peut rien obtenir par ce procédé concernant les états accessibles $A(x_0, t_0, U)$ car la projection selon π de la sous-variété $L_{(x_0, t_0)}$ n'est pas une sous-variété de R^n. Par contre concernant l'ensemble des états $A(x_0 t_0, t_1, U)$ on peut obtenir la proposition suivante.

PROPOSITION 2.2.2. *Soit dans* R^n *le point* $\bar{x}$ *appartenant à* $A(x_0, t_0, t_1, U)$. *L'ensemble* $A(x_0, t_0, t_1, U)$ *est inclus dans une sous-variété dont l'espace tangent au point* $\bar{x}$ *est l'espace vectoriel*:

$$\mathscr{L}(D(f, U)(\bar{x}, t_1)) \cap \{(x, t); t = 0\}.$$

Démonstration. L 'ensemble de R^{n+1}

$$\{(x, t_1); x \in A(x_0, t_0, t_1, U)\}$$

est inclus dans la sous-variété de R^n:

$$L_{(x_0, t_0)} \cap \{(x, t); t = t_1\}.$$

Remarque. Ce résultat est un peu plus précis que le précédent. Il reste cependant très insuffisant comme le montre l'exemple classique suivant, du à Filippov.

Exemple 4.

$$\frac{dx}{dt} = -y^2 + u^2, \qquad x \in R,$$

$$u \in R, \ |u| = 1.$$

$$\frac{dy}{dt} = u, \qquad y \in R,$$

Dans R^3 la famille de champs de vecteurs associée est :

$$X^1(x, y, t) = \begin{pmatrix} -y^2 + 1 \\ 1 \\ 1 \end{pmatrix},$$

$$X^2(x, y, t) = \begin{pmatrix} -y^2 + 1 \\ -1 \\ 1 \end{pmatrix}.$$

Calculons les crochets. On a

$$[X^1X^2](xy) = \begin{pmatrix} 0 & -2y & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} -y^2 + 1 \\ -1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 & -2y & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} -y^2 + 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4y \\ 0 \\ 0 \end{pmatrix},$$

$$[[X^1X^2]X^1] = \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}.$$

Les vecteurs X^1, X^2, $[[X^1X^2]X^1]$ sont toujours indépendants. L'ensemble $A(x_0, 0, 1, U)$ est inclus dans un ouvert de R^2, ce qui n'apporte évidemment aucun renseignement supplémentaire!

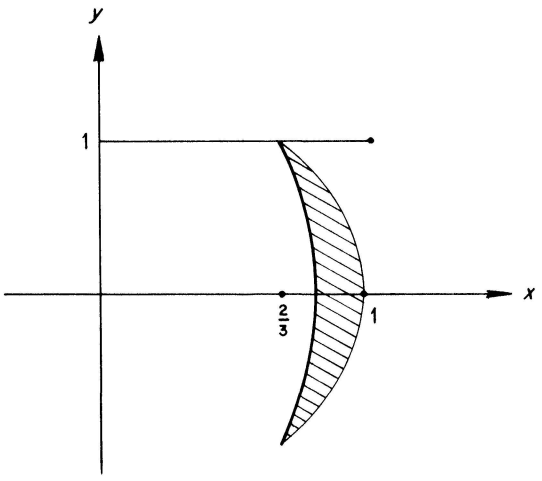


FIG. 5.

Quelques calculs élémentaires montrent que l'allure de $A(x_0, 0, 1, U)$ est donnée par Fig. 5. L'ensemble des états accessibles à l'instant 1 est un ensemble limité par deux arcs, l'arc passant par le point (1, 0) n'appartenant pas à l'ensemble. Filippov a montré par cet exemple qu'en l'absence d'hypothèse de convexité sur l'ensemble :

$$\{f(x, u); u \in U\},$$

il ne peut pas exister de contrôle en temps minimum. Ceci est très clair lorsqu'on représente dans R^3 la feuille intégrale du système défini par les deux champs X^1 et X^2. La troisième composante des champs étant constamment égale à 1 la troisième coordonnée s'interprète comme le temps écoulé depuis le départ. On voit facilement que cette feuille intégrale est une partie de R^3 limitée par deux surfaces, la surface inférieure n'appartenant pas à la partie, ce qui entraine la non-existence d'un contrôle en temps minimum.

On peut enfin obtenir une condition nécessaire pour que l'ensemble $A(x_0, t_0, t_1, U)$ soit d'intérieur non vide.

PROPOSITION 2.2.3. *Une condition nécessaire pour que l'ensemble $A(x_0, t_0, t_1, U)$ des états (x_0, t_0)-accessibles à l'instant t_1 du système* (1) *(analytique) soit d'intérieur non vide est que*

$$\mathscr{L}(\tilde{D}(f, U) \cdot (x_0, t_0))$$

soit de dimension $n + 1$.

Démonstration. Supposons que $\dim(\mathscr{L}(\tilde{D}(f, U) \cdot (x_0, t_0))) \leqq n$; alors pour tout $\bar{x} \in A(x_0, t_0, t_1, U)$ on a

$$\dim\{\mathscr{L}(\tilde{D}(f, U) \cdot (\bar{x}, t_1)) \cap \{(x, t); t = 0\}\} < n$$

car l'espace $\mathscr{L}(\tilde{D}(f, U) \cdot (\bar{x}t_1))$ est toujours transverse à l'hyperplan $\{(x, t); t = 0\}$.

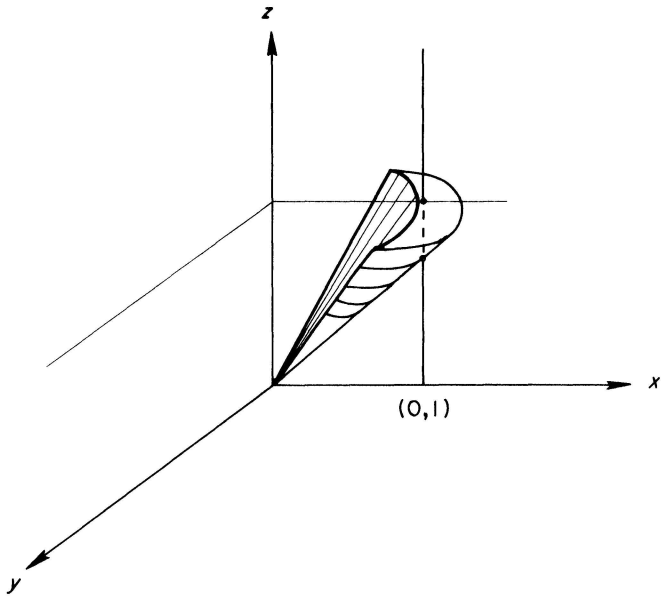


FIG. 6.

La proposition découle alors de la proposition 1.3.5.

Exemple 5. Appliquons ce résultat au système:

$$\frac{dx}{dt} = Ax + Bu, \qquad x \in R^n, \qquad u \in R^p, \qquad U = \{(u_1, \cdots, u_p); |u_i| = 1\},$$

$$A \in \mathscr{L}(R^n, R^n), \qquad B \in \mathscr{L}(R^p, R^n),$$

$$x(0) = 0.$$

On sait qu'une condition nécessaire et suffisante (Kalman et Halkin) pour que l'ensemble des états (0, 0)-accessibles du système ci-dessus soit d'intérieur non vide est que

$$\text{rang}\,[B, AB, \cdots, A^{n-1}B] = n;$$

en appliquant la proposition 2.2.2 on obtient la condition *nécessaire*. En effet, la famille de champs de vecteurs associée est:

$$D(f, U) = \{Ax \pm V_i; i = 1, \cdots, p\},$$

où V_1, V_p représentent les colonnes de B.

On a alors

$$[Ax + V_i, Ax + V_j] = AV_j - AV_i = A[V_j - V_i].$$

On obtient donc tous les champs:

$$AV_1, AV_2, \cdots, AV_p$$

(à un facteur multiplicatif près). Si on recommence on a

$$[Ax + V_i, AV_j] = A \cdot AV_j = A^2V_j.$$

On voit donc que la saturée de la famille $D\,(f, U)$ contient les champs:

$$Ax \pm V_i, \pm AV_i, \pm A^2V_i, \cdots, \pm A^{n-1}V_i, \qquad i = 1, \cdots, p.$$

Le théorème de Hamilton-Caeley montre qu'il est inutile de continuer (A^n s'exprime à partir des A^i, $i \leqq n - 1$) et que par conséquence on a obtenu la saturée de $\tilde{D}\,(f, U)$. D'après la proposition 2.2.3 une condition nécessaire pour que $A\,(0, 0, t_1, U)$ soit d'intérieur non vide est que l'espace vectoriel engendré par les vecteurs

$$\pm V_i, \pm AV_i, \cdots, \pm A^{n-1}V_i, \qquad i = 1, \cdots, p,$$

soit de dimension n, ce qui peut également s'écrire

$$\text{rang}\,[B, AB, \cdots, A^{n-1}B] = n.$$

2.3. Etats accessibles d'un système autonome symétrique. Nous étudions ici les systèmes autonomes particuliers du type suivant.

DÉFINITION 2.3.1. On appelle système *symétrique* un système de la forme

$$\frac{dx}{dt} = H(x) \cdot u, \qquad x \in R^n, \qquad u \in U \subset R^p \text{ (anal.)},$$

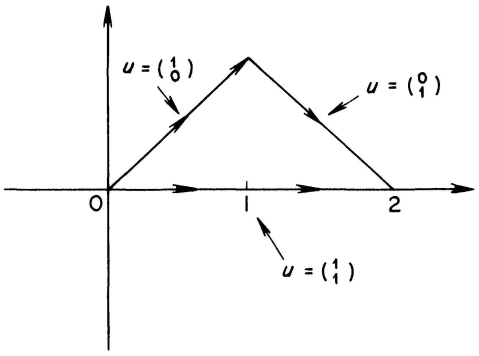


FIG. 7.

En effet soit dans R^2 le système défini par:

$$\frac{dx}{dt} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} u, \qquad x \in R^2, \qquad u \in U \subset R^2,$$

$$U = \{(u_1, u_2); |u_i| \leqq 1\}.$$

Il est clair que le point (2, 0) peut être atteint au temps $t = 1$ à partir de l'origine si on utilise des contrôles appartenant à U alors qu'avec des contrôles appartenant à $\overline{U}$ il faut attendre le temps $t = 2$ (cf. Fig. 7).

Ceci réside dans le fait qu'en choisissant les contrôles "bang-bang" dans l'ensemble $\overline{U}$ on n'utilise pas toute "l'énergie" du système. Pour obtenir des propositions de contrôlabilité "bang-bang" stricte il faut choisir un U permettant d'utiliser toute l'énergie tel que, par exemple,

$$\{(u_1, \cdots, u_p); |u_i| = 1\}.$$

La proposition 2.3.1 est la conséquence des résultats qui vont suivre.

Introduisons la famille D des champs de vecteurs analytiques tels que

$$X \in D \Leftrightarrow \forall x \in R^n : X(x) \in H(x)(U).$$

Cette famille est évidemment symétrique; notons $\tilde{D}$ la saturée de D pour l'opération de crochet. Soit $\tilde{L}_{x_0}$ la feuille integrale de $\tilde{D}$ passant par x_0, on sait que $\tilde{L}_{x_0}$ est une "sous-variété" intégrale (Prop. 1.3.5) de $\tilde{D}$. Nous allons utiliser cette propriété de $\tilde{L}_{x_0}$ pour démontrer la proposition suivante.

PROPOSITION 2.3.2. *Quel que soit x_0 on a l'inclusion:*

$$A(x_0, 0, U) \subset \tilde{L}_{x_0},$$

où $A(x_0, 0, U)$ est l'ensemble des états (x_0, t_0)-accessibles (déf. 2.1.2).

Remarquons pour commencer que la proposition suivante est fausse. "*Soit $t \to \sigma(t)$ un arc différentiable tel que*

$$\frac{d\sigma}{dt} \in \mathscr{L}(\tilde{D}(\sigma(t)))$$

quel que soit t le point $\sigma(t)$ appartient à $\tilde{L}_{\sigma(0)}$."

Exemple 7. En effet considèrons dans R^2 les deux champs:

$$X^1 = \begin{pmatrix} 1 \\ y \end{pmatrix}, \quad X^2 = \begin{pmatrix} 1 \\ -y \end{pmatrix}.$$

Soit $D = \{X^1, X^2\}$, il est clair que les feuilles intégrales de $\tilde{D}$ sont les 3 ensembles:

$$E^+ = \{(x, y)\,; y > 0\}, \qquad E^0 = \{(x, y)\,; y = 0\}, \qquad E^- = \{(x, y)\,; y < 0\}.$$

Soit maintenant l'arc

$$t \to \begin{pmatrix} t^3 \\ t^3 \end{pmatrix} = \sigma(t), \qquad t \in R.$$

On a

$$\frac{d\sigma}{dt} = \begin{pmatrix} 3t^2 \\ 3t^2 \end{pmatrix};$$

par conséquence,

$$\frac{d\sigma}{dt} \in \mathscr{L}(\tilde{D}(\sigma(t))),$$

et pourtant l'arc $\sigma(t)$ traverse les trois feuilles intégrales de D (cf. Fig. 8).

Le fait que $A\,(x_0, 0, U)$ est inclus dans L_{x_0} ne sera donc pas une conséquence du fait que si

$$x(x_0, 0, t_1, \mathscr{U})$$

est un élément de $A(x_0, 0, U)$ l'arc

$$t \to x(x_0, 0, t, \mathscr{U}) = \sigma(t) \qquad \text{(déf. 2.1.2)}$$

est tel que

$$\frac{d}{dt}\sigma(t) = H(\sigma(t)) \cdot \mathscr{U}(t) \in \mathscr{L}(D(\sigma(t))).$$

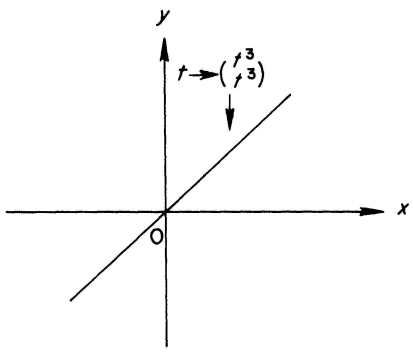


FIG. 8.

Par contre supposons maintenant que l'application :

$$f : R^n \times R \to R^n,$$

$$(x, t) \to f(x, t)$$

soit une application continue, différentiable en x, telle que

$$\forall x \in R^n, \qquad \forall t \in R, \qquad f(x, t) \in \mathscr{L}(\tilde{D}(x)).$$

Soit $x\,(x_0, t)$ la solution maximale de l'équation différentielle :

$$\frac{dx}{dt} = f(x, t), \qquad x(0) = x_0.$$

L'application $x\,(x_0, t)$ est une application d'un intervalle ouvert de J de R dans R^n.

Si nous montrons que pour tout t_0 appartenant à J il existe un voisinage $\mathscr{U}_{t_0}$ de t_0 tel que

$$t \in \mathscr{U}_{t_0} \Rightarrow x(x_0, t) \in \tilde{L}_{x(x_0,t_0)},$$

on aura montré que la solution maximale $x\,(x_0, t)$ appartient entièrement à L_{x_0}. Soit donc t_0 un élément de J. Soit l'application :

$$\tilde{f} : \tilde{L}_{x(x_0,t_0)} \times R \to T\tilde{L}_{x(x_0,t_0)},$$

où $T\,L_{x(x_0,t_0)}$ est le fibré tangent à $L_{x(x_0,t_0)}$ définie par

$$\tilde{f}(x, t) = f(x, t),$$

où $f\,(x, t)$ est considéré comme un élément de l'espace tangent en x à la variété $\tilde{L}_{x(x_0,t_0)}$, ce qui est possible puisque $f(x, t)$ appartient à $\mathscr{L}(\tilde{D}(x))$. On a défini ainsi une "équation différentielle" sur la variété $\tilde{L}_{x(x_0,t_0)}$. Soit $\tilde{x}\,(x(x_0, t_0), t)$ une solution dans $\tilde{L}_{x(x_0t_0)}$ de

$$\begin{cases} \dfrac{dx}{dt} = \tilde{f}(x, t), \\ x(t_0) = x(x_0, t_0). \end{cases}$$

Cette solution peut évidemment être considérée comme une solution dans R^n de

$$\begin{cases} \dfrac{dx}{dt} = f(x, t), \\ x(t_0) = x(x_0, t_0). \end{cases}$$

Par conséquence sur un voisinage de t_0 la solution $x(x_0, t)$ ne quittera pas la variété $\tilde{L}_{x_0}$.

Pour démontrer la proposition 2.3.2 il suffit d'appliquer ce qui vient d'être dit pour $f(x, t)$ à la fonction $H(x)\mathscr{U}(t)$ les discontinuités de $\mathscr{U}(t)$ n'introduisant pas de difficultés supplémentaires.

Introduisons maintenant la famille $\bar{D}$ de champs de vecteurs définie par

$$\bar{D} = \{\pm V^1, \pm V^2, \cdots, \pm V^p\}.$$

La proposition suivante est une conséquence immédiate des définitions.

PROPOSITION 2.3.3. *Soient $\bar{L}_{x_0}$ la feuille intégrale de $\bar{D}$ passant par x_0* (déf. 1.2.3), *$A(x_0, 0, \bar{U})$ l'ensemble des états (x_0, σ)-accessibles* (déf. 2.1.2) *du système*

$$\frac{dx}{dt} = H(x)u, \qquad u \in \bar{U}.$$

On a l'égalité

$$\bar{L}_{x_0} = A(x_0, 0, U).$$

La proposition 2.3.1 est une conséquence immédiate de la proposition suivante.

PROPOSITION 2.3.4. *Soient $\bar{L}_{x_0}$ et $\tilde{L}_{x_0}$ les feuilles intégrales des familles $\bar{D}$ et $\tilde{D}$ passant par x_0. On a les égalités:*

$$\bar{L}_{x_0} = A(x_0, 0, \bar{U}) = A(x_0, 0, U) = \tilde{L}_{x_0}.$$

Démonstration. On vient de voir que

$$\bar{L}_{x_0} = A(x_0, 0, \bar{U}) \subset A(x_0, 0, U) \subset \tilde{L}_{x_0}.$$

Il suffit donc de montrer que $\bar{L}_{x_0} = \tilde{L}_{x_0}$. Il est clair que pour tout x de R^n on a

$$\mathscr{L}(\tilde{\bar{D}}(x_0)) = \mathscr{L}(\tilde{D}(x_0)).$$

Par conséquence l'égalité désirée est une conséquence de la proposition 1.3.3 (théorème de Chow).

On peut enfin énoncer, lorsque le système n'est pas analytique, le théorème de contrôlabilité suivant.

PROPOSITION 2.3.5. *Soit le système autonome symétrique:*

$$\frac{dx}{dt} = H(x)u, \qquad x \in R^n, \qquad u \in \bar{U} \subset R^p, \qquad C^\infty,$$

$$\bar{U} = \left\{ (u_1, u_p) ; |u_i| = 0 \text{ ou } 1, \sum_{i=1}^{p} |u_i| = 1 \right\}.$$

Soit $\bar{D}$ la famille de champs

$$\bar{D} = \{\pm V^1, \cdots, \pm V^p\},$$

et $\bar{D}$ la saturée $\bar{D}$ par crochet.

Si $\mathscr{L}(\bar{D}(x_0))$ est de dimension n l'ensemble $A(x_0, 0, \bar{U})$ est un ouvert contenant x_0, si quel que soit x dans R^n $\mathscr{L}(\bar{D}(x))$ est de dimension n l'ensemble $A(x_0, 0, \bar{U})$ est égal à R^n.

Démonstration. C'est une conséquence immédiate des propositions 1.3.3 et 1.3.4.

Exemple 8. Pour terminer remarquons que la proposition de contrôlabilité "bang-bang" est certainement valable pour des systèmes non symétriques.

Considérons par exemple dans R^2 le système:

$$\frac{dx}{dt} = u(t)X^1(x) + (1 - u(t))X^2(x), \qquad x \in R^n, \qquad u \in U \subset R,$$

$$U = \{u; 0 \leqq u \leqq 1\},$$

$$\overline{U} = \{0\} \cup \{1\}.$$

On peut montrer assez facilement que les ensembles

$$\mathscr{U} \cap A(x_0, 0, \overline{U}) \quad \text{et} \quad \mathscr{U} \cap A(x_0, 0, U)$$

sont égaux lorsque $\mathscr{U}$ est un voisinage de x_0. En effet les courbes intégrales de X^1 et X^2 définissent au voisinage de x_0 une "figure" dont l'allure est donnée par Fig. 9.

Il est clair que si on intègre

$$\frac{dx}{dt} = u(t)X^1(x) + (1 - u(t))X^2(x), \qquad D \leqq 1 \leqq 1,$$

$$x(0) = x_0,$$

on obtient un arc qui appartient à la région hachurée (pour t assez petit). Cette région se définit simplement à l'aide des deux courbes intégrales de X^1 et X^2 passant par x_0. Il n'en est plus de même dans le cas où l'on considère le même système dans R^3.

Si les vecteurs $X^1(x_0)$, $X^2(x_0)$, $[X^1X^2]$ (x_0) sont indépendants, on a alors au voisinage de x_0 une figure du type de Fig. 10. Sur cette figure on a fait apparaître les deux arcs:

$$X_t^1(x_0), \quad t \geqq 0,$$

$$X_t^2(x_0), \quad t \geqq 0.$$

Ces deux arcs étant tracés on obtient deux "surfaces" en traçant d'une part:

$$\text{les arcs } X_t^1(x), \qquad t \geqq 0, \qquad x \in X_t^2(x_0), \qquad t \geqq 0;$$

d'autre part:

$$\text{les arcs } X_t^2(x), t \geqq 0, \qquad x \in X_t^1(x_0), \qquad t \geqq 0.$$

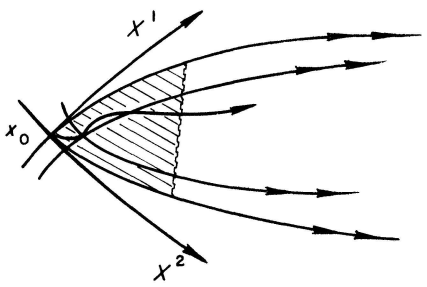


FIG. 9.

Ces deux surfaces S_1 et S_2 sont donc décrites par

$$S_1 = \{X^2_{t_2} \circ X^1_{t_1}(x_0), t_1 \geqq 0, t_2 \geqq 0\},$$
$$S_2 = \{X^1_{t_1} \circ X^2_{t_2}(x_0), t_1 \geqq 0, t_2 \geqq 0\}.$$

Grâce à ces deux surfaces on peut définir une région dans R^3. On montrera ensuite qu'une courbe intégrale de

$$\frac{dx}{dt} = u(t)X^1(x) + (1 - u(t))X^2(x), \qquad 0 \leqq u \leqq 1,$$

$$x_{(0)} = x_0,$$

reste dans cette région. Dans le cas général, pour p champs dans R^n, on devra définir un certain nombre d'hypersurfaces qui permettront de définir une région "bang-bang" accessible. La définition de ces hypersurfaces fera intervenir le comportement mutuel des champs donc les crochets au point x_0.

L'exemple de Filippov (exemple 4) prouve qu'il n'existe pas de théorème général de "bang-bang" contrôlabilité (même au sens large). Cet exemple fait apparaitre le phénomène suivant.

On a vu que

$$[X^1X^2](xyz) = \begin{pmatrix} 4y \\ 0 \\ 0 \end{pmatrix}.$$

Ce crochet est donc nul sur l'hyperplan $y = 0$.

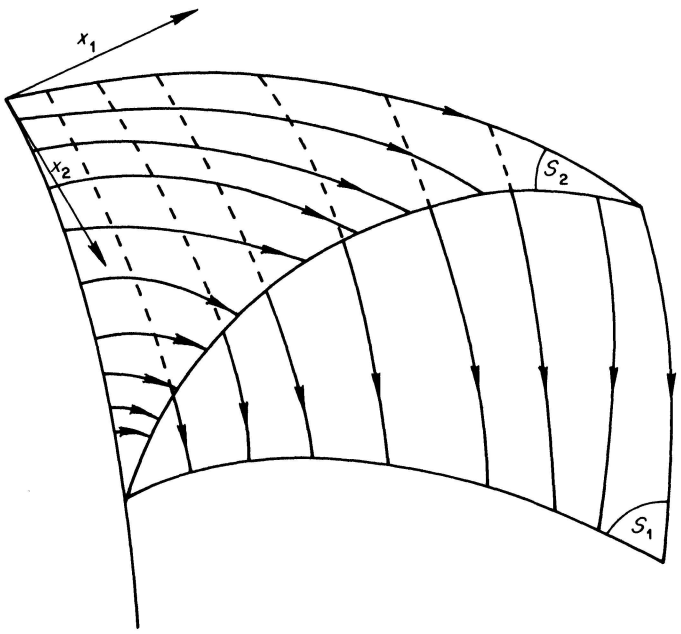


FIG. 10.

Si d'autre part on applique la construction décrite dans l'exemple précédent à cette situation on constate que les deux surfaces S_1 et S_2 sont imbriquées, ce qui ne permet évidemment plus de limiter une région de l'espace. Le calcul des crochets suivants tels que

$$[[X^1X^2], X^1](xyz) = \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}$$

permet simplement d'affirmer que l'ensemble des états accessibles est situé au "dessus" d'une partie de S_1 ou S_2.

Le fait que les deux surfaces soient imbriquées est lié à la "singularité" du crochet $[X^1X^2]$.

Ces deux exemples suggèrent donc de donner une définition adéquate des "singularités" qui peuvent se produire et d'énoncer alors, pour des F.C.V. qui ne présentent pas de "singularités," un théorème de "bang-bang" contrôlabilité du type suivant:

"*Soit D une F.C.V. ne présentant pas de "singularités." Soit*

$$\frac{dx}{dt} = f(x,t,u), \qquad x \in R^n, \qquad u \in R^p, \qquad u \in U,$$

$$x(t_0) = x_0,$$

un système tel que pour tout x et tout t on ait l'inclusion

$$f(x,t,u) \subset \text{co}\,(D(x)).$$

Alors l'ensemble des états accessibles du système est inclus dans la feuille intégrale de D passant par le point x_0."

Remerciement. Pour terminer je desire remercier J. Martinet qui pourrait revendiquer bien des idées exploitées ici.

REFERENCES

[1] CLAUDE CHEVALLEY, *Theory of Lie Groups*, Princeton University Press, Princeton, New Jersey, 1946.

[2] W. L. CHOW, *Uber Systeme von linearen partiellen Differentialglechungen erster Ordnung*, Math. Ann., 117 (1939), pp. 98–105.

[3] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.

[4] H. HALKIN, *A generalization of LaSalle's "bang-bang" principle*, this Journal, 2 (1964), pp. 199–202.

[5] ———, *On a generalization of a theorem of Liapunov*, J. Math. Anal. Appl., 10 (1965), pp. 325–329.

[6] ———, *Mathematical foundations of systems optimization*, Topics in Optimization, Academic Press, New York, 1967, pp. 197–262.

[7] R. HERMANN, *On the accessibility problem in control theory*, Internat. Sympos. Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325–332.

[8] ———, *The differential geometry of foliation. II*, J. Math. Mech., 11 (1962), pp. 305–315.

[9] ———, *Differential Geometry and the Calculus of Variations*, Academic Press, New York, 1968.

[10] H. HERMES, *Controllability and the singular problem*, this Journal, 2 (1964), pp. 241–260.

[11] H. HERMES AND G. W. HAYNES, *Nonlinear controllability via Lie theory*, this Journal, 8 (1970), to appear.
[12] H. HERMES, *Attainable sets and generalized geodesic spheres*, J. Differential Equations, 3 (1967), pp. 256–270.
[13] R. E. KALMAN, Y. C. HO AND K. S. NARENDA, *Controllability of linear dynamical systems*, Contribution to Differential Equations, 1 (1963), pp. 189–213.
[14] J. KUČERA, *Solution in large of control problem*: $\dot{x} = (A(1 - u) + Bu)x$, Czechoslovak Math. J., 16 (91) (1966), pp. 600–623.
[15] ———, *Solution in large of control problem*: $\dot{x} = (Au + Bv)x$, Ibid., 17 (92) (1967), pp. 91–96.
[16] J. P. LASALLE, *The time optimal control problem*, Contribution to the Theory of Nonlinear Oscillations, vol. 5, Princeton University Press, Princeton, New Jersey, 1960, pp. 1–24.
[17] C. LOBRY, *Séminaire d'analyse numérique*, Grenoble, 1968.
[18] YOZÔ MATSUSHIMA, *Groupes de Lie*, Polycopié, Université de Grenoble, Service de Mathématiques pures, 1966.
[19] G. DE RHAM, *Variétés différentiables*, Hermann, Paris, 1960.
[20] P. SAMUEL, *Théorie algébrique des nombres*, Hermann, Paris, 1967.
[21] S. ZARISKI, *Commutative Algebra*. II, D. Van Nostrand, Princeton, N. J., 1960.
[22] JEAN MARTINET, *Sur les singularités des formes différentielles*, Thèse, Grenoble, 1969.
[23] D. LUNA, *Stratifications*, Séminaire, Grenoble, 1969.
[24] R. L. BISHOP AND R. J. CRITTENDEN, *Geometry of Manifolds*, Academic Press, New York, 1964.